SPEECH ACOUSTICS AND PHONETICS

Text, Speech and Language Technology

VOLUME 24

The titles published in this series are listed at the end of this volume.

Speech Acoustics and Phonetics

by

GUNNAR FANT
Department of Speech, Music and Hearing,
Royal Institute of Technology,
Stockholm, Sweden

KLUWER ACADEMIC PUBLISHERS
DORDRECHT / BOSTON / LONDON

A C.I.P Catalogue record for this book is available from the Library of Congress.

ISBN 1-4020-2789-3 (PB)
ISBN 1-4020-2373-1 (HB)
ISBN 1-4020-2790-7 (e-book)

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Sold and distributed in North, Central and South America
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

CONTENTS

FOREWORD

by Louis C.W. Pols
University of Amsterdam

Since almost 60 years Gunnar Fant has actively contributed to the field of Speech Acoustics and Phonetics. Almost every speech scientist in the world knows him because he is still a frequent visitor of conferences and workshops all over the world. They may also know about some of his work, but only few have a proper knowledge about the details and the breadth of his pioneering and still ongoing research. This is partly related to the fact that in the early days many of his contributions were only published in the Quarterly Progress and Status Reports (QPSR) of the Speech Transmission Laboratory (STL) of the Speech Communication and Music Acoustics department, later called the department of Speech, Music and Hearing (TMH) of the Stockholm Royal Institute of Technology (KTH). Several other important publications only appeared in various conference proceedings or in not easily accessible journals. It is most fortunate that Gunnar Fant has taken up the challenge to produce this book of Selected Writings. It is his own unique selection and it only contains publications from his own hand with or without colleague co-authors. Via the Introductions per section he guides us himself through the multitude of topics and explains the historical developments and the various connections. It makes especially those older publications accessible that otherwise would have been very hard to find. I suggested to him to extend as much as possible the introductions to each of the six main chapters (Speech research overview; Speech production and synthesis; The voice source; Speech analysis and features; Speech perception; and Prosody). This, in my opinion has substantially contributed to the readability of the 19 individual contributions. It was difficult for him to limit himself to these 19 papers only, as one can imagine from his full list of publications with over 260 old and new titles, that he presents both chronologically as well as topic-wise. Nevertheless you have this goldmine now in front of you and I hope and expect that this will be joyful and informative reading.

Gunnar Fant has been awarded many times, most recently in June 2004 with the new IEEE James L. Flanagan Speech and Audio Processing Award, together with another speech giant Ken Stevens, for their "fundamental contributions to the theory and practice of acoustic phonetics and speech perception". Still I believe that the most valuable contribution of a scientist to the scientific and world community are his products in writing, especially when they are made accessible in such a splendid form as in this book.

His pioneering book on the "Acoustic theory of speech production" published in 1960 by Mouton, is of course a classic that is for instance still invaluable in articulatory synthesis. However, many other topics, like the formant banana, the Jakobsen-Fant-Halle distinctive features, the LF source model, the OVE synthesiser,

the invariance-variability dispute, syllable prominence and the speech code, get much attention and are presented in the proper perspective in the present book. The final section is about Prosody, the topic that keeps him most busy these days. He works on it together with Anita Kruckenberg and Johan Liljencrants, and it concerns not just the prosody of spoken Swedish but also that of poetry.

PREFACE

This is a collection of articles spanning half a century of speech research. It started at the Ericsson Telephone Company in Stockholm, 1946–1949. The following two years were spent at MIT. In 1951 a small research group was established at the KTH in Stockholm. This unit, the Speech Transmission Laboratory, became the foundation for our present department of Speech, Music and Hearing. An early expansion was promoted by US grants in the 1960's. Research in speech acoustics, phonetics, hearing and handicap aids dominated the activities up to 1990, after which more applied projects in speech technology gained dominance at the department.

Much of our work was published in our Speech Transmission Laboratory Quarterly Progress and Status Reports (STL-QPSR). It had the advantage of reaching a large international forum with a minimum of delay, but gave less time for publications in established journals. The purpose of the present book is to make available a collection of articles from various reports and publications, which contribute to the knowledge foundation. It is not a structured textbook, but it provides a reference material for quite a wide range of topics in the field.

It is with great gratitude that I acknowledge the contributions from all those who have been involved in developing our department and its research and have served as co-workers. They are too many to be listed here, but there are two persons from the early days that I want to mention. Marianne Richter, my first employee, participated in language statistics and became our first finance officer. Si Felicetti started our STL-QPSR in 1960 and promoted our international contacts. My present research is carried out together with Anita Kruckenberg and Johan Liljencrants. They and my many friends in the scientific community have contributed to the delight of cooperation and scientific discoveries. Closest in research profile is Ken Stevens. His book, Acoustic Phonetics, is of monumental value and provides deeper insights in matters of speech production.

Valuable suggestions for the planning of the volume were given by Louis Pols.

INTRODUCTION

As time goes by, a look ahead in science and technology gains by a spotlight on earlier periods. Speech technology has provided important tools for applications in man-machine communication systems and is growing rapidly. But there is a risk that expansion will be limited by insufficient attention to the potentialities of speech and language research. The symbiosis between technology and basic research that has made possible the advance, now shows a tendency to turn into polarization. Speech technology is highly dependent on statistical tools and large data bases, whilst phonetics tends to become fractionalised by narrowly defined problems or by abstract issues with small or no relevance for the overall code of spoken language.

There is a great need for integrated basic knowledge of speech production, acoustics, perception and cognitive processes and of the encoding of linguistically defined units in the speech wave and other parts of the overall speech chain. In several articles I have attempted to coin the concept of these relations as ***the speech code***. The relative success of speech synthesis has created an illusion that we have a profound insight in the speech code. This illusion becomes especially apparent when operating in the reverse direction, that is, given a record of the speech wave we attempt to decipher what was said. An example is spectrogram reading, a difficult but rewarding exercise, which is mediated by knowledge about speech production.

In quest of the speech code, we are faced with issues concerning invariance and variability. However, the invariance issue ceases to present a problem if we systematically develop rules for structuring variability of all kinds, not only language, dialectal and contextual variations but also variations specific to speaker, speaking style and emotions. Much more effort is needed to develop the code and make it available for practical applications, as well as for the advance of general phonetics and linguistics. It is only with a profound knowledge of the speech code and human behaviour that we can realize ultimate goals of advanced and reliable systems.

The 19 selected articles span a period of almost 50 years. Some of the older ones still maintain a pedagogical value, or they document unique studies of some significance. I have arranged them in six chapters according to categories, which have some inevitable overlap of contents.

They have been selected from my complete publication list, which has been structured in annual order with a number tag for each item of a particular year. In a separate table they have been sorted in categories conforming to those of the selected articles and additional activities.

Seven of the articles originate from our STL-QPSR laboratory reports, now the TMH- QPSR of the department of Speech, Music and Hearing of KTH. They have been distributed in 800 copies to 50 different countries. From 2001 they are available on the Internet only.

The first chapter contains a single article devoted to the development of phonetics and speech technology in a historical perspective, viewed from my personal experiences during more than half a century.

The topic of the second chapter is speech production and synthesis. It contains five articles related to my Acoustic Theory of Speech Production (Fant, 1960), earlier work at the Ericsson Telephone Company, and follow-up studies of vocal tract configurations and associated circuit theory. Our early formant based synthesis system is described, and also a novel synthesizer intended for articulatory coding.

Chapter 3 is devoted to the voice source. Three major articles are included. They deal with properties of the LF-model, glottal source-vocal tract interaction, inverse filtering and applications in speech synthesis. Frequency domain aspects are given special attention.

Chapter 4 contains four articles devoted to acoustic phonetic descriptive analysis of speech within a spectrographic frame, as well as interpretations with reference to distinctive feature analysis and the concept of the speech code.

Chapter 5 contains two articles, one on auditory modelling applied to vowel perception, including theory and experiments on two-formant approximations. The second article addresses a problem in audiology, to predict intelligibility at a specific hearing loss from a pure tone audiogram in relation to the frequency-intensity distribution of speech formants.

During the last 15 years a substantial part of my work has been devoted to prosody together with Anita Kruckenberg. The first article in chapter 6 reviews our findings of quantal patterns in segmental timing, and the second is devoted to the acoustic realization of meter and rhythm in poetry reading, The third article reports on the great individual spread in pausing within a sentence, and differences between prose reading and news reading. The forth article provides a comprehensive and detailed overview of our work on Swedish prosody, covering the acoustic phonetic basis of syllabic stress and perceived prominence, intonation patterns and models of speech synthesis. The descriptive frame has a foundation in voice source properties and aerodynamics.

More detailed comments supplementing the brief presentation above are to be found in the introductions to the separate chapters, with lists of articles recommended for additional reading.

LIST OF SELECTED ARTICLES

5. SPEECH PERCEPTION

6. PROSODY

- *STL-QPSR* and *TMH-QPSR* are the quarterly reports from KTH department of Speech, Music and Hearing.

CHAPTER 1

SPEECH RESEARCH OVERVIEW

Speech technology of today derives from a pioneering era around 1945–1965, as an outgrowth of interdisciplinary contacts and joint activities in a large number of fields directed to various aspects of speech communication and technical applications.

At the technical end it has involved acoustics, electronics, circuit theory and the advent of computer science, and at the humanities end linguistics, phonetics, psychology, physiology, information theory.

My first stay at M.I.T was in 1949–1951, the start of the pioneering era, when I got involved in activities together with Ken Stevens, Roman Jakobson and Morris Halle, and later with James Flanagan at Bell Laboratories, relations maintained over a lifetime. My article is a personal account for how the field has developed and of my experiences, how it started in Sweden 1945–1949, the M.I.T period and then back in Sweden 1951, when I formed the Speech Transmission laboratory at the Royal Institute of Technology, KTH, now the department of Speech, Music and Hearing. Much has happened during these years, including events that have deserved some anecdotal remarks.

In perspective, I find that basic research had a relative prominence during the first part of the half century (Fant, 1968), whilst applications took over in the second half as a result of the expanding involvement of computer science. I have described this as a polarization of research efforts. Computer power can not substitute a profound insight in the speech code. This is a recurrent theme in many of my articles.

A more detailed account of perspectives and personal contacts can be found in an unpublished manuscript structured as interview questions (Fant, 1996).

CHAPTER 1 ARTICLE

[1.1] Fant, G. (2004). More than half a century in phonetics and speech research.*Gunnar Fant Selected Writings*. (Revised version of a presentation at the Swedish phonetics meeting in Skövde, May 24–26, 2000).

ADDITIONAL READING

Fant, G. (1968). Analysis and synthesis of speech processes. In *Manual of Phonetics*, Chapt. 8, 173–276 (B. Malmberg, ed.). Amsterdam, North-Holland Publ. Co.

Fant, G. (1996). Historical notes. Response to interview questions posed by Louis-Jean Boe and Pierre Badin. *KTH-TMH manuscript,* 15 pages.

Fant, G. (2004). Speech research in a historical perspective. In J. Slifka, S. Manuel and M. Matthies (Eds.), From Sound to Sense: 50+ Years of Discoveries in Speech Communication, Research Laboratory of Electronics MIT, June 11–13, 2004, pp. 20–40.

CHAPTER 1.1

MORE THAN HALF A CENTURY IN PHONETICS AND SPEECH RESEARCH*

ABSTRACT

This is a brief outlook of my experiences during more than 50 years in phonetics and speech research. I will have something to say about my scientific career, the growth of our department at KTH, and I will end up with an overview of research objectives in phonetics and a summary of my present activities.

1. INTRODUCTION

As you are all aware of, phonetics and speech research are highly interrelated and integrated in many branches of humanities and technology. In Sweden by tradition, phonetics and linguistics have had a strong position and speech technology is well developed and internationally respected. This is indeed an exciting field of growing importance that still keeps me busy. What have we been up to during half a century? Where do we stand today and how do we look ahead? I am not attempting a deep, thorough study, my presentation will in part be anecdotal, but I hope that it will add to the perspective, supplementing the brief account presented in Fant (1998).

2. THE EARLY PERIOD 1945–1966

2.1. *KTH and Ericsson 1945–1949*

I graduated from the department of Telegraphy and Telephony of the KTH in May 1945. My supervisor, professor Torbern Laurent, a specialist in transmission line theory and electrical filters had an open mind for interdisciplinary studies. My thesis was concerned with theoretical matters of relations between speech intelligibility and reduction of overall system bandwidth, incorporating the effects of different types of hearing loss.

This work paved the way for my employment 1945–1949 at the acoustics laboratory of the Ericsson Telephone Company. They needed basic knowledge of the formant structure of Swedish speech sounds as a support to intelligibility tests and for predictions of effects of selected frequency band elimination in telephony. I was given free hands to construct equipment for speech analysis and go ahead with studies of the time-frequency-intensity distributions of Swedish speech sounds and sentences. My tools were primitive but effective. I constructed a wave analyser for manual continuous variation of the center frequency and a choice of a narrow or a broad bandwidth setting. The subject was seated in a proper anechoic chamber. A special feature, usually not retained in present day's routines, was absolute intensity calibrations in all recordings. Vowels and sonorants were analysed with the subject sustaining the sound for 3 seconds, (Fant 1948). Spectrograms were manually compiled from intensity versus time oscillograms in a substantial number of frequency

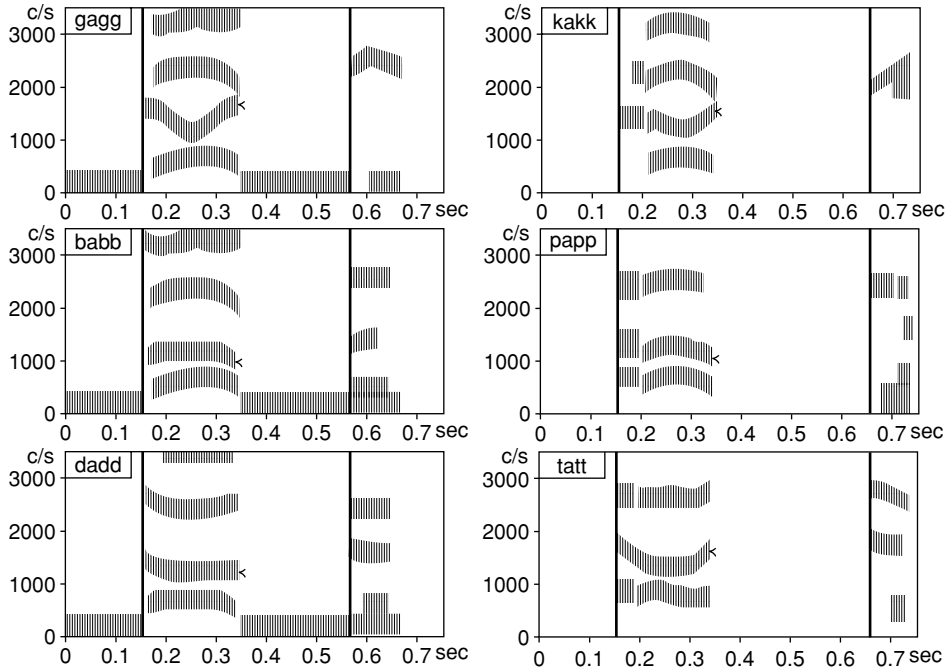


Figure 1. Stylized spectrograms of "gagg", "babb", "dadd", "kakk, "papp", "tatt" compiled from band-pass oscillograms. From Fant (1949, 1959).

bands (Fant, 1949). These spectral patterns closely resembled the spectrograms I had just learned about from the Bell Labs Visible Speech, see Figure 1.

Except for these early Ericsson internal reports, results were not published until 1959, in an Ericsson Technics monograph, Fant (1959), that also included basic speech synthesis theory. It was later added to my two other thesis publications, (Fant, 1960) and Fant (1958). An outcome of the Ericsson work (Fant, 1949) published in Fant (1959) was a scatter diagram of the frequency intensity distribution of Swedish vowel and consonant formants, referred to a specified talking distance. A summary view of major formant areas in the frame of an audiogram was constructed, see Figure 2. This has become a standard reference in audiology referred to as the "formant banana" (Fant 1995B). The acoustic-phonetic data acquired at that time was included in a compendium on Swedish Phonetics (Fant 1957) which was adopted for phonetic courses at the Stockholm University.

2.2. *MIT 1949–1951*

My Ericsson work included detailed spectral sections of the burst of unvoiced stop consonants (Fant, 1949, 1959), see Figure 3. When I came to America in late 1949 these data were presented at a seminar at Harvard University. In the audience was Roman Jacobson who found my data to fill a missing link in his feature theory, the

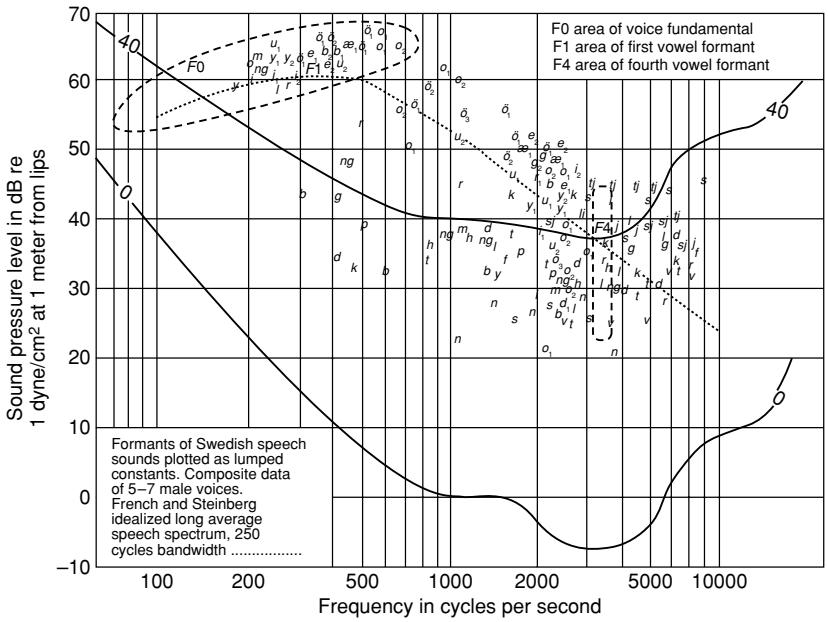


Fig. 45. Sound pressure level versus frequency plot of the vowel and consonant formant data.

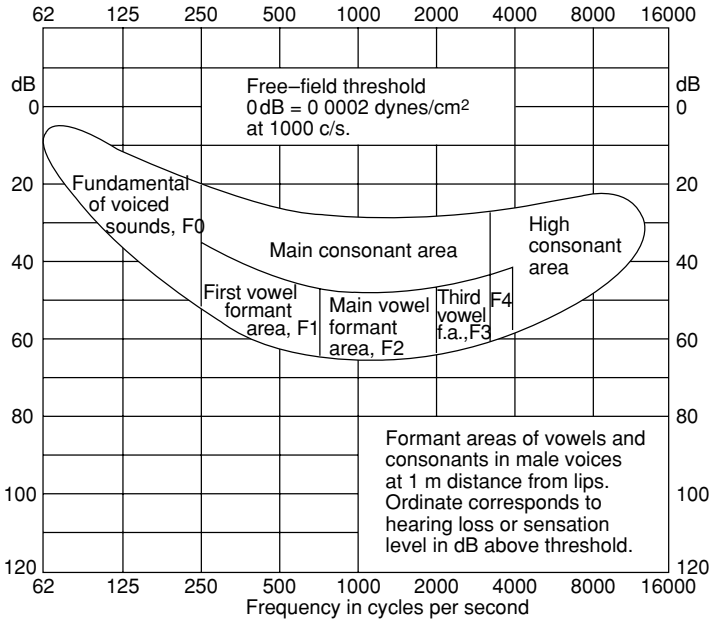


Fig. 46. Schematized presentation of the average speech spectrum within an audiogram in terms of main formant areas. Sensation levels (down) are relative to the standardized free-field threshold.

Figure 2. Above, intensity-frequency scatterplot of Swedish vowels and consonants. Below formant regions within the frame of an audiogram, referred to as the speech banana. From Fant (1949, 1959).

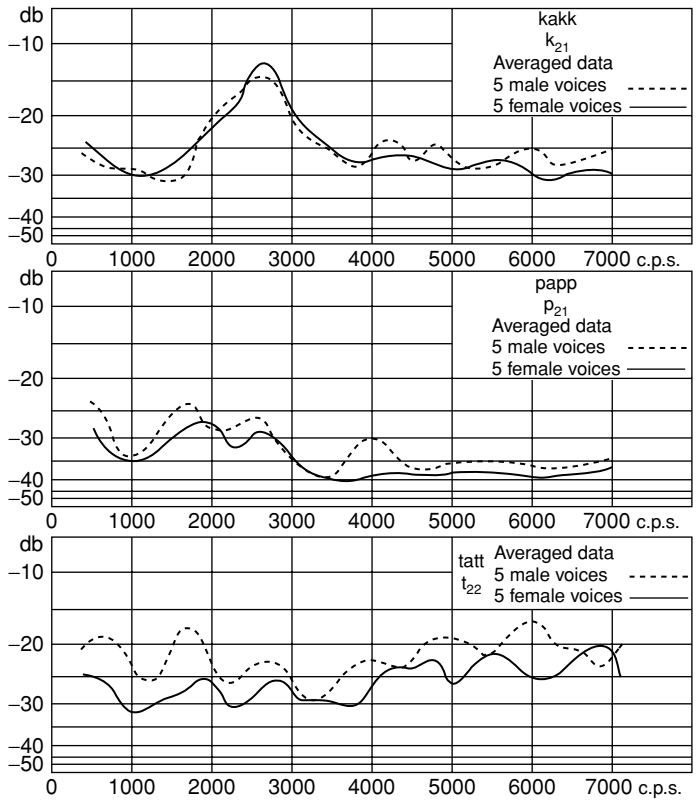


Figure 3. Spectral sections of the release phase of syllable final unvoiced Swedish stop. From Fant (1949, 1959).

realisation of compactness and gravity in consonants. The graph shows the single peak realisation of [k], the low-frequency emphasis of [p] and the high frequency emphasis of [t]. This is how our co-operation started. It was followed up by the teamwork of Jakobson, Fant, Halle (1952). My later views on distinctive features were published in Fant (1973).

The stay at MIT 1949–1950 was extremely rewarding. This was a truly pioneering era in speech research as an outgrowth from linguistics, electrical circuit theory, psycho-acoustics and information theory. A fascinating experience was the Pattern Play back at Haskins laboratories (Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1952). It had a pedagogical value unsurpassed by recent year's digital technology. With my internalised knowledge of spectral representations in Swedish I could paint a sentence and have it replayed. It sounded monotone but was understandable. My work on speech production theory had a good start at that time. With the support of Roman Jakobson and Morris Halle X-ray studies of sustained articulations of a Russian subject were pursued. These became the foundation for my book, Acoustic Theory of Speech Production (Fant, 1960).

At MIT I gained much from ongoing work on acoustics, circuit theory, transform theory and information theory. I had close contacts with Ken Stevens then a graduate student, a relation which has remained throughout our lives in virtue of common interests and scientific approach. A common interest was in vocal tract electrical line analogs (Stevens, Kasowsky and Fant, 1953; Fant, 1960).

2.3. *The Speech Transmission Laboratory, 1951–1966*

Back in Sweden in 1951 the nucleus of the Speech Transmission Laboratory was formed. A project already started in 1948, when I was part time employed at KTH, was to carry out statistics of the relative occurrence of letters, phonemes and words in written material and in telephone conversations. This data was now processed by my first employee, Marianne Richter, who also became the secretary and finance officer of the group. Results from our survey were published in the proceedings of the Linguistics conference in Oslo 1958 (Fant and Richter, 1958). The main part is retained in my speech communication course compendium (Fant, 1969).

It is of some historical interest that in 1947 when we acquired the telephone speech material from the Swedish Telecom lab, it was stored on an old-fashioned wire recorder. After we had transcribed the text the wire had to be cut into pieces to preserve the privacy.

Our first synthesiser OVE 1, (Fant, 1953) was capable of producing quite natural vowels and simple sentences made up of transitional voiced gestures. Its name was given in a program over the Swedish radio in 1953. The interviewer asked if the machine had a name. Well, I could make it produce the name OVE and so it was christened. Later it was interpreted as "Orator Verbalis Electricus". Our present OVE 1 designed by Johan Liljencrants in the 1970 still maintains a substantial pedagogical value.

OVE II, capable of producing unrestricted connected speech, appeared in 1961, (Fant and Martony, 1962). It was programmed from function lines drawn with conducting ink on a plastic sheet and employed vacuum tube electronics. A transistorised version was later developed by Johan Liljencrants. During the 1962 international speech communication seminar in Stockholm it created a substantial interest. Janos Martony and our guest John Holmes had demonstrated high quality copy synthesis with OVE 2.

An incident is worth while mentioning. One night during the conference, without our consent, a representative of the Melpar Company in USA, Joseph Campanella, took photographs and measurements of the function generator, which enabled them to produce a rather exact copy. A year later their synthesiser EVA was on sale. Arne Risberg who had been involved in our original design and I now made a claim of compensation, which was granted by a small royalty. However, not many Eva systems were sold and our revenue was small.

At this time an important object of speech technology was vocoder systems for bandwidth reduction in telephony to be combined with secrecy coding. We were consulted by the Ericsson Company in the design of their channel vocoder. At that time an alternative approach was formant coded analysis-synthesis systems, which

potentially is closer to production theory. But formant vocoders were not a success. A well known system, that of the Melpar Company sounded natural but was quite unintelligible, an instance of a failure of both the model and the formant tracking. More sophisticated systems are now considered in bandwidth reduction schemes.

2.4. *The Later Period, 1967–Present Days*

A major break-through in speech technology was text-to-speech synthesis. Our system was developed by Rolf Carlson and Björn Granström, who worked out a language and phonetics program interfacing the OVE II (Carlson and Granström, 1975, 1991). Applications in reading aids for the blind and in speaking aids for vocally handicapped have been of great importance. and the prototype has paved the way for applications in general information services.

A financial support was received in 1959–1973 through American grants from the US Air Force and US army and from the National Institute of Health. These made possible a substantial increase of our basic research, which continued with increasing support from Swedish funding agencies.

Research in speech production, speech synthesis and speech perception including auditory modelling was promoted, and a wide range of handicap applications were considered. Music acoustics developed strongly and gradually attained its important role in general and clinically oriented research in the human voice.

Our Speech Transmission Laboratory Quarterly Progress and Status Report, STL-QPSR, now TMH-QPSR, was started in 1960 and has been distributed free of charge, at present to over 800 individuals and institutions in 50 countries. Over the years we have had a large number of contributors. Phonetically oriented research was initiated by Björn Lindblom and Sven Öhman already in 1960. The most productive authors in terms of number of contributions, see the cumulative index in STL-QPSR 1/1995, have been: Eva Agelfors, T.V-Ananthapadmanabha, Mats Blomberg, Rolf Carlson, Kjell Elenius, Gunnar Fant, Frans Franson, Karoly Galyas, Jan Gauffin-Lindqvist, Björn Granström, Brita Hammarberg, Sharon Hunnicutt, Inger Karlsson, Johan Liljencrants, Qi-guang Lin, Björn Lindblom, Janos Martony, Lennart Nord, Arne Risberg. Karl-Erik Spens, Sven Öhman: In music acoustics the main contributors were Anders Askenfelt, Frans Fransson, Erik Janson, Johan Sundberg and Sten Ternström, the latter two also contributing to studies of voice and speech production. The largest number of contributions 1960–1995 were those of Sundberg (97) and Fant (77).

Working contacts were established with Russian and French research groups through symposia and frequent scientific visits, from Leningrad (Ludmilla Chistovich) and from Grenoble (Pierre Badin). The editor of our STL-QPSR was Si Felicetti, who also had an important role in the organisation of scientific meetings in Sweden and abroad.

The STL-QPSR made possible publication with a minimum of delay to a large number of recipients, but a disadvantage still remains. It often becomes a substitute for refereed publications in established journals. There are several major contributions in the STL-QPSR on speech production theory, speech analysis and speech

perception which have not been published elsewhere. I have in mind articles related to properties of the voice source (Fant, Liljencrants, Lin, 1985, which is the original publication on the LF model); (Fant and Lin 1988; Fant, 1995A). A comprehensive collection of articles on the voice source appears in Gobl (2003).

Furthermore, work on vocal tract transfer functions (Fant, 1972); Badin and Fant (1984); Båvegård, Fant, Gauffin and Liljencrants (1993) and nonlinear formant normalization, Fant, 1975) were initiated.

An activity well represented in the period 1960–1980 was auditory modelling (Carlson, Granström and Fant, 1970, Carlson and Granström, 1982, Carlson, Fant and Granström, 1975, Fant, 1978). Apart from studies in speech prosody we nowadays lack projects on auditory functions and perception. The symposium volume edited by Carlson and Granström (1982) contains several articles that deserve a renewed attention.

Two major publication covering broad areas of speech research and contributing to the historical perspective are a chapter in Malmberg, Manual of Phonetics, (Fant 1968) and the book Speech Sounds and Features (Fant, 1973).

A peripheral problem which engaged me during this early period was to explain the distortion encountered in the speech of divers when breathing different gas mixtures adjusted for operation at a certain depth. While the gas mixture accounts for a simple Donald Duck linear frequency transposition, the pressure component at high depths produces a non-linear frequency scale compression perceived as a nasal quality (Fant and Lindqvist-Gauffin, 1968).

A minor, handicap oriented, project I happened to get involved in was inspired by Arne Risberg. It was a phonetically structured hand-finger alphabet for non-hearing subjects (Fant, 1972). It has been tested in a Japanese school for the deaf.

There has always been a tension between forensic science and speech research. There exists a seeming but deceptive analogy between finger prints and voice prints which would motivate spectrographic analysis as legal evidence. However, such applications have been refuted by a majority of speech scientists as being premature and insecure. In 1970 Stockholm was visited by the head of the FBI, who gave an interview in Dagens Nyheter on the use of voiceprints for identifying terrorists. The newspaper turned to me for comments, which were quite negative. On the first page of Dagens Nyheter of January 16 our controversy was reviewed. There were two photographs, one of Edgar J. Hoover head of the FBI, and one of Gunnar Fant, suggested as a possible FBI enemy number one.

The close ties between linguistics, phonetics, speech technology, auditory research and music acoustics and their importance for a number of applied areas such as handicap aids, auditory and speech rehabilitation have been especially well developed in Sweden. A recognition was the funding through grants of a project named "Tal, ljud och hörsel" (Speech, sound and hearing) which sponsored some research and annual meetings in the years 1983–1985. When the support terminated our annual Swedish phonetics conferences took over as a forum for these interdisciplinary meetings.

In recent years the major recipient of funding within our department has been the Center for Speech Communication (CTT). It has had a productive start, to a

substantial part sponsored by Swedish industry. The main activity is development of interactive man-computer information systems. It involves work in speech synthesis with an added talking head display, speech recognition and language engineering.

But what about the basic research? It still prevails but not with the same breadth as in earlier periods. On the other hand, we have extended contacts with phonetics departments in Sweden, e.g. in prosody research and dialect studies.

3. RECENT ENGAGEMENTS

During the last 10–20 years a substantial part of my time has been spent in studies of prosody (Fant, Nord and Kruckenberg, 1986; Fant and Kruckenberg, 1989, 1994, 1996, 1999, 2000; Fant, Kruckenberg and Liljencrants, 2000A, 2000B; Fant, Kruckenberg, Gustafson and Liljencrants 2002. A detailed account of the role of sub-glottal pressure in production and perception appears in Fant, Kruckenberg, Liljencrants and Hertegård (2000). Related problems in singing have been treated in our Music Acoustics group (Sundberg, Andersson and Hultqvist, 1999).

Studies of poetry reading have also been carried out (Fant, Kruckenberg and Nord, 1991; Kruckenberg and Fant, 1993).

Studies of the voice source have been continued (Fant, 1993, 1995A, 1997, Fant and Lin, 1988). The LF model originating from Fant, Liljencrants and Lin (1985) has now been widely accepted in synthesis systems.

A novelty, increasing the insight in the speech chain, is to supplement a traditional speech analysis display of oscillogram, spectrogram, F0 and intensity with data from perceptual scaling of syllable prominence labelled RS (Fant and Kruckenberg, 1989, 1994, 1999, Fant, Kruckenberg and Liljencrants, 2000A, 2000B).and if available also records of subglottal pressure. The RS parameter is of special importance for relating degrees of stress and emphasis to all relevant physical parameters, including voice source properties. We have adopted two intensity curves, one is plain SPL and the other has a high-frequency pre-emphasis, SPLH, which in combination with SPL allows an estimate of spectral tilt.

From detailed acoustic phonetic studies of text reading during the last 15 years, we have now been able to develop quite efficient prosody rules for text-to-speech synthesis (Fant, Kruckenberg, Gustafson and Liljencrants, 2002) A novelty lies in the frequency and time-domain normalization of intonation which makes possible the averaging of data from male and female subjects. In synthesis, local accent modulations are superimposed on base curves for successive prosodic groups within a sentence. The shape and peak height of the local modulations as well as phoneme durations are varied according to predicted prominence RS and the position within a sentence. A high degree of naturalness has been achieved by imposing a syntactical frame for prosodic grouping and pausing. Our rules can to some extent be applied to other languages than Swedish.

An ambition for future work is to test our accumulated knowledge of prosodic realisations not only in concatenated synthesis, which nowadays has gained dominance, but also in formant coded synthesis, revised to include higher level rules for articulatory continuities and prosodic modifications of segmental structures and

voice source characteristics. The prominence parameter RS will have an important role for scaling both prosodic and segmental parameters.

It would carry too far to perform a detailed review of developments and the status of Swedish phonetics. Over the years there has been a substantial progress and a close co-operation with speech technology. The most prominent area is prosody which by tradition has a firm foundation in Sweden and is of considerable importance for text-to-speech synthesis and in language teaching. Co-operation networks involving KTH, and the Stockholm, Lund and Umeå universities are well established.

4. RESEARCH OBJECTIVES AND TRENDS

In the last 50 years, research in phonetics has expanded over a wide range of areas reflecting diverse interests, backgrounds and methods of approach. I shall take the liberty of executing a personal view of structures and trends.

One is the apparent dispersion of phonetic research into many sub-areas of narrowly defined problems. On the other hand, there is a search for universal principles of language development and unifying principles in human behaviour, e.g. the relations of speech and language to other human activities.

However, a central object of general phonetics is the speech communication chain, the many stages of message encoding within the speaker, the transmission medium and the receiving partner. It can be studied in functional details, for instance with respect to vocal tract acoustics, articulatory gestures and properties of the auditory system, including transformational relations between links in direct or reverse order, e.g. articulatory interpretation of spectrograms. The latter technique is of great importance as a supplement to direct studies of articulatory dynamics. Our insight in brain functions of the speaker and the listener is limited and restricted to indirect observations and functional analogies.

All these aspects of the function of the speech chain can be referred to by a major category labelled MECHANISMS which is largely a matter of physiology, physics and sensory psychology. A second major category, introduced by adding the linguistic function and linguistic competence is the SPEECH CODE, the relation between message units and their realisation in the speech chain, involving both discrete linguistic units as in phonology and social codes of expressing attitudes, denotations and emotions (Fant, 1985, 1989).

A challenge for future research, and the ultimate aim of general phonetics, is to force the speech code, i.e. to predict the articulatory, acoustic and perceptual manifestation of any utterance given the message transcript and the particular language, dialect, speaker and situational context. One can conceive of this task as a very advanced project of deriving rules for text-to-speech synthesis. Our present knowledge of the speech code is incomplete and in part hidden in text-to-speech programs. We need an extension of code oriented analysis such as in the comprehensive early studies of Dennis Klatt for American English and of Carlson and Granström for Swedish.

In recent years, large data banks of spoken material have been collected, e.g. for studies of Swedish dialects and dialect independent speech recognition. They have been used mainly in automatic computer training, but have not been much exploited in acoustic-phonetic studies. However, in quest of the speech code, we

need comprehensive analysis of all acoustics correlates, segmental and prosodic, which puts a limit to the size of a corpus. Up till now most attempts of this kind have been fragmentary, i.e. directed to some detail aspect only.

Irrespective of the aim of a project it can be pursued with more or less emphasis on the knowledge to be gained or to instrumental tools, either those available or new tools developed within the project. This dichotomy has been referred to as "knowledge driven" versus "instrumentally driven" research. An example of instrumental approach is to derive formant locus equations from established plotting techniques, which in my view is of rather limited interest for studies in speech perception. Knowledge driven projects e.g. for multi-parameter investigations of the speech code, are of greater importance but more demanding.

We need to expand our knowledge base, but there is also a lag in transfer of knowledge to consider. Present systems, especially for speech recognition, are not designed to make use of available knowledge on prosody and segmental constraints. The question is now whether we shall be able to effectively handle all new developing information in explicit rules, or if we have to continue to rely on computers to learn the code for speech recognition? Furthermore, will concatenated synthesis remain a prerequisite for high quality text-to-speech systems? At present "Unit selection" concatenation is gaining ground in virtue of the human quality but it is a poor medium for introducing prosody rules and lacks flexibility with respect to voice type. Diphone systems are better suited as we have shown. Articulatory oriented parametric systems will eventually take over. Some kind of hybrid systems may emerge.

The two main lines of approach, the statistical and the knowledge based, appear in all branches of speech technology. Computing power can not substitute crucial knowledge. My forecast for the future is that a more solid and integrated view of speech and language structure will develop and find its way also into speech recognition and synthesis work.

5. IN QUEST OF THE SPEECH CODE

A main point in my look ahead has been the challenge to force the speech code. This is not a matter of a single intellectual undertaking to find the key. It requires coordinated, multilevel investigations and integrated modelling over long periods of time. The seeming lack of invariance which has discouraged so many investigators ceases to be a problem if we are able to structure the variability as a part of the code (Fant, 1985).

The complexity is great but, as time goes by, research proceeds towards increasing insights. Mankind is making much progress in mapping the genetic code. We need some of the same patience and persistence in mapping the speech code. The reward will be a more solid theoretical basis of phonetics as well as new methods in speech technology design and improved quality of performance.

REFERENCES

Badin, P. and Fant, G. (1984). Notes on vocal tract computation. *STL-QPSR* 2–3/1984, 53–108.

Båvegård, M., Fant, G., Gauffin, J. and Liljencrants, J. (1993). Vocal tract sweep-tone data and model simulations of vowels, laterals and nasals. *STL-QPSR* 4/1993 43–76.

Carlson, R. and Granström, B. (1975). A text-to-speech system based on a phonetically oriented programming language. *STL-QPSR 1/1975*, 1–4.
Carlson, R. and Granström, B. (1982). Towards an auditory spectrograph. In R. Carlson and Granström, B. (eds), *The representation of speech in the auditory system*. Elsevier Biomedical Press, Amsterdam, 109–115.
Carlson, R. and Granström, B. (1991). Performance rules in a text-to-speech system. In J. Sundberg, L. Nord and Carlson, R. (eds) *Music Language, Speech and Brain,* MacMillan Press, 121–131.
Carlson, R., Fant, G. and Granström, B. (1975). Two-formant models, pitch, and vowel perception. In G. Fant and. Tatham M. (eds.). *Auditory Analysis and Perception of Speech*. London: Academic Press Inc, 55–82.
Carlson, R., Granström, B. and Fant, G. (1970). Some studies concerning perception of isolated vowels. *STL-QPSR 2–3/1970*: 19–35.
Fant, G. (1948). *Analys av de svenska vokalljuden*, LM Ericsson report H/P-1035.
Fant, G. (1949). *Analys av de svenska konsonantljuden.* LM Ericsson report H/P-1064.
Fant, G. (1953). Speech communication research*, IVA, Royal Swedish Academy of Engineering Sciences,* Stockholm 24, 331–337.
Fant, G. (1957). *Den akustiska fonetikens grunder.* KTH, Inst. för Telegrafi-Telefoni, Rapport nr 7, Taltransmissionslaboratoriet (61 pages).
Fant, G. (1958). Modern instruments and methods for acoustic studies of speech, *Acta Polytechnica Scandinavica,* No 1, 1–81.
Fant, G. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. *Ericsson Technics* 1 1959, 1–106.
Fant, G. (1960). *Acoustic theory of speech production*. Mouton, the Hague. 2nd ed., 1970.
Fant, G. (1962). Descriptive analysis of the acoustic aspects of speech. *LOGOS*, Vol **5**, No. 1, 3–17.
Fant, G. (1964). Auditory patterns of speech. *Models for the Perception of Speech and Visual Form,* Boston, Mass., Nov. 11–14.
Fant, G. (1968). Analysis and synthesis of speech processes". In B. Malmberg (ed) *Manual of Phonetics,* Amsterdam, North-Holland Publ. Co. Chapt. 8, 173–276.
Fant, G. (1969). *Kompendium i talöverföring, del 1*. KTH, inst för talöverföring 1967.
Fant, G. (1972A). Vocal tract wall effects, losses, and resonance bandwidths. *STL-QPSR* 2-3/1972, 28–52.
Fant, G. (1972B). Q-codes. In *Intl. Symp. on Speech Communication Ability and Profound Deafness, Stockholm 197*0, A.G. Bell Ass. for the Deaf, Washington DC, eds., 261–268.
Fant, G. (1973). *Speech Sounds and Features*. Cambridge, MA, USA The MIT Press.
Fant, G. (1975). Non-uniform vowel normalization. *STL-QPSR* 2-3/1975, 1–19.
Fant, G. (1978). Vowel perception and specification. *Rivista Italiana de Acustica* II, 69–87.
Fant, G. (1980). The relation between area functions and the acoustical signal. *Phonetica* **37**, 55–86.
Fant, G. (1985). Features, fiction and fact. In J. Perkell et al. (eds). *Invariance and Variability of Speech Processes, Brain and Reading.* Lawrence Erlbaum Ass. Publ., 480–492.
Fant, G. (1986). Glottal flow, models and interaction. *Journal of Phonetics,* **4** (3/4) Theme issue, Voice Acoustics and Dysphonia. Gotland, Sweden, August 1985, 393–399.
Fant, G. (1989). The speech code. In C. von Euler, I. Lundberg and G. Lennerstrand G. (eds.) *Brain and Reading* MacMillan, London, 171–182.
Fant, G. (1991). What can basic research contribute to speech synthesis? *Journal of Phonetics*, **19**, 75–90.
Fant, G. (1992). Vocal tract area functions of Swedish vowels and a new three-parameter model. *Proc. ICSLP-92,* Vol. 1, 807–810.
Fant, G. (1993). Some problems in voice source analysis, *Speech Communication* 13, 7–22.
Fant, G. (1995A). The LF-model revisited. Transformations and frequency-domain analysis. *STL-QPSR* 2-3/1995, 119–156.
Fant, G. (1995B). Speech related to pure tone audiograms. In (eds.) G. Plant and K.E. Spens. *Profound deafness and speech communication.* Whurr Publ. Ltd, London, 299–305.

Fant, G. (1997). The voice source in connected speech. *Speech Communication* 22: 125–139.
Fant, G. (2000). Half a century in phonetics and speech research. *Swedish phonetics meeting in Skövde, May 24–26, 2000.* (Revised version).
Fant, G., Hertegård, S. and Kruckenberg, A. 1996. Focal accent and subglottal pressure. *TMH-QPSR* 2/1996, 29–32.
Fant, G. and Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style*, STL-QPSR* 2/1989, pp. 1–83.
Fant, G. and Kruckenberg, A. (1994). Notes on stress and word accent in Swedish. *Proceedings of the International Symposium on Prosody, Sept 18 1994, Yokohama.* Also published in *STL-QPSR* 2-3/1994, 125–144.
Fant, G. and Kruckenberg, A. (1996). On the quantal nature of speech timing. *Proceedings of the International Conference on Spoken Language Processing. ICSLP* 1996, 2044–2047.
Fant G. and Kruckenberg, A. (1999). Prominence correlates in Swedish prosody. *Proc. of International Conference of Phonetic Sciences, 1999*, San Francisco 3, 1749–1752.
Fant G. and Kruckenberg, A. (2001). F0 analysis and prediction in Swedish prose reading. In N. Grønnum and J. Rischel (eds.), *To honour Eli Fischer-Jørgensen.* Travaux du Circle Linguistique de Copenhague. Copenhagen, Reitzel, 124–147.
Fant, G., Kruckenberg, A., Gustafson, K. and Liljencrants, J. (2002). A new approach to intonation analysis and synthesis of Swedish. *Speech Prosody 2002, Aix en Provence.* Also in *Fonetik 2002, TMH-QPSR* 2002.
Fant, G., Kruckenberg, A and Liljencrants, J. (2000A). Acoustic-phonetic analysis of prominence in Swedish. Antonis Botinis (ed*) Intonation. Analysis, Modelling and Technology*, Kluwer, Academic Publishers, pp. 55–86.
Fant, G. Kruckenberg, A and Liljencrants, J (2000B). The Source-Filter Frame of Prominence. *Phonetica* 57, 113–127.
Fant, G., Kruckenberg, A., Liljencrants, J. and Hertegård, S. (2000). Acoustic phonetic studies of prominence in Swedish. *TMH-QPSR* 2/3 2000, pp. 1–52.
Fant, G., Kruckenberg, A. and Nord, L. (1991A). Durational correlates of stress in Swedish, French and English. *Journal of Phonetics* 19, 1991, 351–365.
Fant, G., Kruckenberg, A. and Nord, L. (1991B). Stress patterns and rhythm in the reading of prose and poetry with analogies to music performance. In J. Sundberg, L. Nord, R. Carlson (eds) *Music, Language, Speech, and Brain*, Wenner-Gren International Symposium Series, Vol. 59, 1991, 380–407.
Fant, G., Kruckenberg, A. and Nord, L. (1992). Prediction of syllable duration, speech rate and tempo. *Proceedings of the International Conference on Spoken Language Processing. ICSLP* 92, Banff, Vol 1, 667–670.
Fant, G., Liljencrants, J. and Lin, Q. (1985). A four-parameter model of glottal flow. *STL-QPSR* 4/1985, 1–13.
Fant, G. and Lin, Q. (1988). Frequency domain interpretation and derivation of glottal flow parameters, *STL-QPSR* 2-3/1988, 1–21.
Fant, G. and Lindqvist-Gauffin, J. (1968). Pressure and gas mixture effects on divers speech. *STL-QPSR* 1/1968, 7–17.
Fant, G., Nord, L. and Kruckenberg, A. (1986). Individual Variations in Text Reading. A Data-Bank Pilot Study*, STL-QPSR* 4/1986 1-17 and in *RUUL 17, 1987*, 104–114.
Fant, G. and Mártony, J. (1962). Instrumentation for parametric synthesis, OVE II synthesis strategy, and quantization of synthesis parameters. *STL-QPSR* 2/1962 18–24.
Fant, G.and Richter, M. (1958). Some notes on the relative occurrence of letters, phonemes, and words in Swedish. *Proc. of the VIIIth Intl. Congress of Linguists, Oslo 1958,* 815–816.
Gobl, C. (2003) *The voice source in speech communication*. Thesis at the Department of Speech, Music and Hearing, KTH.
Jakobson, R., Fant, G. and Halle, M. (1952). Preliminaries to speech analysis. The distinctive features and their correlates. *Acoustics Laboratory, Massachusetts Inst. of Technology, Technical Report* No. 13. *MIT press*, seventh edition, 1967.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. Journal of the Acoustic Society of America, 67, 971–995.

Kruckenberg, A. and Fant, G. (1993). Iambic versus trochaic patterns in poetry reading, *Nordic Prosody* VI, Stockholm, 1993, 123–135.

Kruckenberg, A. and Fant, G. (1995). Notes on syllable duration in French and Swedish. *Proc. ICPhS 95,* Vol II, 158–161.

Liberman, A. M., Delattre, P.C. and Cooper, F. S. (1952). The role of selected stimulus variables in the perception of the unvoiced consonants. *American Journal of Psychology,* 65, 497–516.

Lin, Q. and Fant, G. (1992). An articulatory speech synthesizer based on a frequency domain simulation of the vocal tract, *IEEE-ICASSP San Francisco,* 1992, paper 173, 1992.

Stevens, K.N., Kasowski, S. and Fant, G. (1953). An electrical analog of the vocal tract. *J. Acoust. Soc. Amer*. 25, 734–742.

Sundberg J, Andersson M and Hultqvist C (1999). Effects of subglottal pressure variations on professional baritone singers' voice sources. *J Acoust Soc Am* 105/3: 1999.

STL-QPSR = Speech Transmission Laboratory Quarterly Progress and Status Report, KTH, since 1998 the TMH-QPSR.

CHAPTER 2

SPEECH PRODUCTION AND SYNTHESIS

Much of my early work on production and synthesis appeared in an Ericsson publication (Fant, 1959), which also contains data from speech analysis and speech processing during the period 1946–1949. Excerpts from this publication is the first item of chapter 2. It serves as a supplement to my Acoustic Theory of Speech Production (Fant, 1960), providing a detailed account for the derivation of the higher pole correction in formant synthesis.

The experimental part, although performed with primitive means, is unique in the respect that it included measurements of formant amplitudes with absolute calibration of sound pressure levels. These data made possible an experimental verification of the predictability of changes in formant amplitudes within a sequence of vowels from their formant frequencies, see also Fant (1956).

The second article of chapter 2, (Fant, 1980), contains essentials from Fant (1960) such as the three parameter vocal tract model with associated nomograms. A particular illustration that deserves a highlight is the close correspondence between the measured and the calculated spectra of [k] and [p] from the X-ray study. The [k] spectrum is dominated by a single formant, a pole, whilst the [p] spectrum shows a dominant spectral minimum, in mathematical terminology a zero.

Other items of general interest in Fant (1980) are the treatment of time-varying vocal tract shapes, and a generalized perturbation theory based on vocal tract energy functions, which enables a calculation of the sensitivity of formants to local perturbations of cavity dimensions, see also the more detailed treatment in Fant (1975A).

Differences between male and female vocal tracts and systematic differences in vowel formant patterns are reviewed. The original data derives from a study of six languages (Fant, 1975B), which also provided mean value graphs of F2 versus F1 and F3 versus F1 of all vowels with formant scale factors inserted. These are reproduced in Appendix 2. An interesting finding was that tenor/base scale factor relations mirrored those of females/males.

Fant (1980) also provides data on vowel formant bandwidths. These derive from vocal tract computations and sweeptone data in Fant (1972). The influence of the subglottal system on formant patterns, notably in consonant burst spectra, has been treated by Fant, Ishizaka and Lindqvist-Gauffin (1972). Subglottal formants in vowels are exemplified in Fant and Lin (1988) in connection with voice source studies, see chapter 3.

The third item (Fant, 2001) is a less well known but now updated study of Swedish vowels, originating from X-ray recordings from the early 1960's with supplementary tomographic registrations. Some of the tomograms, included here in Appendix 1, show the presence of substantial air passages on both sides of the tongue for [u:] and [ɑ:] Calculated formant frequencies show a good correspondence to measured data. The Fant (2001) study also contains a revised version of the three-parameter model, based on overall constraints of cavity dimensions.

A further development of the model with consonant production included was adopted by Fant and Båvegård (1997), with attempts of inverse transformation, to predicted vocal tract configurations from formant patterns.

Pole-zero matching of vocal tract sweeptone data provides a detailed insight in vocal tract system functions, see Båvegård, Fant, Gauffin and Liljencrants (1994).

In 1960 there was a break-through in synthesis techniques. We were able to demonstrate a high degree of naturalness in parametric synthesis with control data copied from spectrograms. The forth item (Fant and Martony, 1962) is a basic report on the synthesis configuration and on synthesis strategies. Our present formant synthesis has retained essentials of the configuration.

Articulatory coding will in the long run make possible more flexible and natural synthesis. The synthesis architecture outlined in the fifth item by Lin and Fant (1990) deserves attention in future development work.

SELECTED ARTICLES

[2.1] Fant, G. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. *Ericsson Technics,* **1**–1959, 1–106. (Excerpts from pages 43–45, 54–60, 70–73).

[2.2] Fant, G. (1980). The relation between area functions and the acoustical signal. *Phonetica* **37**, 55–86.

[2.3] Fant, G. (2001). Swedish vowels and a new three-parameter model. *TMH-QPSR* 1/2001.

[2.4] Fant, G. and Mártony, J. (1962). Instrumentation for parametric synthesis (OVE II). Synthesis strategy, and quantization of synthesis parameters. *STL-QPSR* 2/1962, 18–24.

[2.5] Lin, Q. and Fant, G. (1990). A new algorithm for speech synthesis based on vocal tract modelling. *STL-QPSR* 2–3/1990, 45–52.

APPENDIX

A1. Tomographic data

A2. Female/male formant data

A3. Diver speech

ADDITIONAL READING

Fant, G. (1956). On the predictability of formant levels and spectrum envelopes from formant frequencies. *For Roman Jakobson,* Mouton and Co., 's-Gravenhage, 109–120.

Fant, G. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. *Ericsson Technics* **1**–1959, 1–106.

Fant, G. (1960). *Acoustic theory of speech production.* The Hague, Netherlands: Mouton, 2nd edition. 1970 (Translated into Russian, *Nauka, Moskva,* 1964).

Fant, G. (1964). Formants and cavities. In E. Zwirner and W. Bethge, (eds.) *Proc. of the Fifth Intl. Congr. of Phonetic Sciences*, Munster. Basel: S Karger, 120–141.

Fant, G. (1975 A). Vocal-tract area and length perturbations. *STL-QPSR* 4/1975, 1–14.

Fant, G. (1975 B). Non-uniform vowel normalization. *STL-QPSR* 2–3/1975, 1–19.

Båvegård, M., Fant, G., Gauffin, J. and Liljencrants, J. (1994). Vocal tract sweeptone data and model simulations of vowels, laterals and nasals. *STL-QPSR* 4/1993, 43–76.

Fant, G. (1972). Vocal tract wall effects, losses, and resonance bandwidths. *STL-QPSR* 2–3/1972, 28–52.

Fant, G., Ishizaka, K. and Lindqvist-Gauffin, J. (1972). Subglottal formants. *STL-QPSR* 1/1972, 1–12.

Fant, G. and Båvegård, M. (1997). Parametric model of the vocal tract area function. Vowels and consonants. *TMH-QPSR* 1/1997, 1–20. Also published in *ESPRIT/BR SPEECHMAPS (6975).* Delivery 28, WP2.2, 1995, 1–30.

CHAPTER 2.1

ACOUSTIC ANALYSIS AND SYNTHESIS OF SPEECH WITH APPLICATIONS TO SWEDISH

Excerpts from Ericsson Technics No 1, 1959 (pp. 43–45, 54–58, 70–73)

2.4. THE SINGLE-TUBE RESONATOR. REST TERM TREATMENT OF HIGHER POLES

Particular attention will be devoted to the single-tube resonator, approximately closed in the driving end and open at the radiating end:

$$U_o/U_q = H_p(s) = \frac{1}{\cosh(\alpha + s/c)l_e} \tag{2.4-1}$$

The pole frequencies are apparently

$$s_n = \sigma_n + j\omega_n = -\alpha c \pm j2\pi c(2n-1)/4l_e \tag{2.4-2}$$

The theoretical justification of this solution is discussed by Zinn (1952), Pipes (1946). The formant frequencies

$$F_n = F_1(2n-1) \tag{2.4-3}$$

are odd integers to the first formant frequency

$$F_1 = c/4l_e \tag{2.4-4}$$

where l_e is the effective length of the tube, end corrections included and c is the velocity of sound.
The bandwidths are

$$B_n = \alpha c/\pi \tag{2.4-5}$$

An F-pattern of $F_1 = 500$ c/s, $F_2 = 1,500$ c/s, $F_2 = 2,500$ c/s, etc., defines the neutral synthetic vowel [∃] which is a convenient reference for discussion of formant patterns as well as for laboratory experiments on speech synthesis. It can be produced from a tube of effective length c/2000 cm.

When vowel synthesis is attempted by means of a set of cascaded resonant circuits, each of these represents a conjugate pair of poles $s = \sigma_n \pm j\omega_n$ of the expansion of Eq. 2.2-6:

$$H_p(\omega) = \frac{1}{\prod_1^\infty [1 - j\omega/(\sigma_n + j\omega_n)][1 - j\omega/(\sigma_n - j\omega_n)]} = \frac{1}{\prod_1^\infty [1 - x_n^2 + j\delta_n x_n]} \tag{2.4-6}$$

where

$$\begin{aligned} x_n &= \omega/\omega_{0n} = f/F_{0n} \simeq f/F_n \\ \delta_n &= -2\sigma_n/\omega_{0n} \simeq B_n/F_n \\ \omega_{0n} &= (\omega_n^2 + \sigma_n^2)^{1/2} \simeq \omega_n \end{aligned} \tag{2.4-7}$$

The first attempts to simulate the vocal transfer function by means of the three or four first poles, neglecting all higher poles, were not very successful. There resulted an appreciable loss of level at the higher formants, as could for instance be seen by comparing the amplitude versus frequency response curve of the neutral vowel approximation employing a number of r poles:

$$|H_p(\omega)| = \frac{1}{\prod_1^r \left[\left(1 - x_n^2\right)^2 + \delta_n^2 x_n^2\right]^{1/2}} \tag{2.4-8}$$

with the small loss approximation of Eq. 2.4-1:

$$|H_p(\omega)| = \frac{1}{\left[\cos^2 \omega l_e/c + \alpha^2 l_e^2 \sin^2 \omega l_e/c\right]^{1/2}} \tag{2.4-9}$$

The failing correction to be applied to Eq. 2.4-8 will be denoted $|K_{rr}(\omega)|$. Thus

$$|H_p(\omega)| = \frac{K_{rr}(\omega)}{\prod_{n=1}^r \left[1 - x_n^2 + j\delta_n x_n\right]} \tag{2.4-10}$$

$$|K_{rr}(\omega)| \simeq \frac{1}{\prod_{n=r+1}^{\infty} \left[1 - x_n^2\right]} \tag{2.4-11}$$

The terms $\delta_n x_n$ of the higher poles have been neglected since they are of importance mainly for the shape of the corresponding higher formant peaks. The correction-factor K_{rr} is to be made use of only up to and including the frequency range of formant F_r. The determination of K_{rr} for the single-tube resonator characterized by

$$x_n = f/F_1(2n - 1) = x_1/(2n - 1) \tag{2.4-12}$$

involves the expansion of the logarithm of Eq. 2.4-11 into a series comprising the first and second order terms:

$$20 \log_{10} k_{rr} = 20 \log_{10} e \sum_{r+1}^{\infty} \left[\frac{1}{(2n-1)^2} x_1^2 + \frac{1}{2} \frac{1}{(2n-1)^4} x_1^4 \quad \cdots\right] \text{(dB)} \tag{2.4-13}$$

From the identity

$$-\text{logcos}(\omega l_e/c) = -\text{logcos} \frac{\pi x_1}{2} = \log \prod_1^{\infty} \left[1 - x_1^2/(2n-1)^2\right] \tag{2.4-14}$$

and a series expansion of these two alternatives there may be derived the relation

$$\left.\begin{aligned}\sum_{1}^{\infty}\frac{1}{(2n-1)^2}&=\frac{\pi^2}{8}\\ \sum_{1}^{\infty}\frac{1}{(2n-1)^4}&=\frac{\pi^4}{96}\end{aligned}\right\} \tag{2.4-15}$$

and it is thus possible to substitute for Eq. 2.4-13 the following more practical form:

$$\begin{aligned}20\log_{10}k_{rr}=20\log_{10}e\Bigg\{&\left[\frac{\pi^2}{8}-\sum_{1}^{r}\frac{1}{(2n-1)^2}\right]x_1^2\\&+\frac{1}{2}\left[\frac{\pi^4}{96}-\sum_{1}^{r}\frac{1}{(2n-1)^4}\right]x_1^4\Bigg\}\ \text{(dB)}\end{aligned} \tag{2.4-16}$$

where $x_1 = f/F_1 = 4\,f l_e/c$. This formula gives a very good fit up to a frequency halfway between F_r and F_{r+1}. It may also be adopted for any specific vocal tract configuration specified by the length measure l_e, but with less precision. The generalized usage is motivated by the significance of the average spacing of the higher formants, which is mainly influenced by l_e. The particular configuration of the higher poles is not crucial as long as their average distance to any particular frequency where the correction shall apply—i.e., their residue at a low frequency—is constant. The error is naturally maximal if the first few formants above F_r deviate appreciably from their neutral positions. The total length of the vocal tract is by no means a constant since it varies from one sound to another. The K_{rr}-function derived above, however, has been found experimentally to be useful as an average correction for cascaded resonance analog synthesizers. The practical incorporation of $K_{rr}(\omega)$ into the circuitry is described in the Appendix.

For calculations of vowel spectra from a system function analysis or for the design of frequency correction networks simulating the $K_{rr}(\omega)$-function, the following numerical expressions derived from Eq. 2.4-16 are given:

$$\begin{aligned}20\log_{10}K_{r2}&=1.06x_1^2+0.0102x_1^4\quad\text{(dB)}\\ 20\log_{10}K_{r3}&=0.72x_1^2+0.0033x_1^4\quad\text{(dB)}\\ 20\log_{10}K_{r4}&=0.54x_1^2+0,00137x_1^4\quad\text{(dB)}\end{aligned} \tag{2.4-17}$$

Contrary to the statements of DUNN (1950), the spectrum of a voiced sound does not fall 12 dB/octave more steeply for each resonance that is passed. The $K_{rr}(\omega)$-factor provides the compensation in terms of a steeply rising high frequency emphasis.

A similar correction for the lack of higher zeroes in an expansion of a zero function may also be derived. In most cases, however, the zeroes of a vocal tract system function can be paired with poles that are mainly influenced by the shunting network responsible for the zeroes. The alternation of poles and zeroes along the frequency scale will cause a reduction of the residue at lower frequencies from higher poles and zeroes. A pole and a zero that coincide will cancel each other and may accordingly be removed from the spectral description. If the distance between

a pole $\hat{F}_n$ and zero $\bar{F}_n$ is large enough to motivate a calculation of their compound effect, the following formula derived from Eq.2.4-6 may be utilized:

$$H_{pn}(s)\, H_{zn}(s) = \frac{1 - \bar{x}_n^2 + j\bar{\delta}_n \bar{x}_n}{1 - \hat{x}_n^2 + j\hat{\delta}_n \hat{x}_n} \tag{2.4-18}$$

where $\hat{x}_n = f/\hat{F}_n$ and $\hat{x}_n = f/\hat{F}_n$. The degree of coupling between the shunt and the main transmission system will influence the degree of pole-zero proximity. The smaller the area of the opening to the shunt and the higher the frequency, the closer will a zero of the shunt come to the corresponding pole and the less will the shunting effect be as specified by Eq. 2.4-18. Applied to speech production this implies that the effect of the cavities behind the source on the spectrum of a fricative or stop decreases with a decrease of the constriction area. This is why F2 and in some instances F3 are found to be very weak in fricatives but more apparent in aspirated sounds.

The effect of nasalization may be described as a distortion of the ideal vowel spectrum by a set of anti-resonances and extra nasal resonances. The main shape of the vowel spectrum will not be seriously influenced by the overlaid nasalization pattern. If formants of the subglottal system appear, they must occur paired with anti-resonances. The greater the coupling to the subglottal system the more apparent will this spectral distortion be. This theory thus provides a criterion for deciding what formants belong to the main or "oral" part of the vocal tract and which formants are to be described to shunting systems. A change in the frequency of an oral formant, i.e., of a frequency included in the F-pattern, will cause a pronounced change in the spectral shape contrary to the case of a non-oral formant, the latter being partially compensated by its zero and displaying less frequency variations.

2.7. SPECTRUM ENVELOPE SYNTHESIS FROM FORMANT FREQUENCIES

The predictability of spectrum envelopes and formant levels from formant frequencies dealt with in an earlier publication (FANT, 1956) is based on the unique relation between the residues along the jω-axis and at the formant poles to the poles and zeroes of the complex frequency plane. The phase information, although useful for the measurement of formant frequencies, is completely redundant. Amplitude frequency spectrum envelopes have proved to be a very useful visual correlate to the phonetic quality of most speech sounds. In addition to the frequency-intensity specification of formants within the spectrum, it is fruitful to discuss the general shape of the spectrum with regard to the gross distribution of sound energy. The spectral building blocks for frequency domain synthesis of voiced sounds may be derived from Eq. 2.5-4 by substituting s for jω and factorizing the absolute values. The logical ordering of the data can conveniently start by bringing together all frequency functions that are to be regarded as constant characteristics. The $K_T(\omega)$-factor is generally disregarded and so is the excitation periodicity function. The voice source poles and the zero at the origin due to radiation are lumped together with the higher pole factor $K_r r(\omega)$.

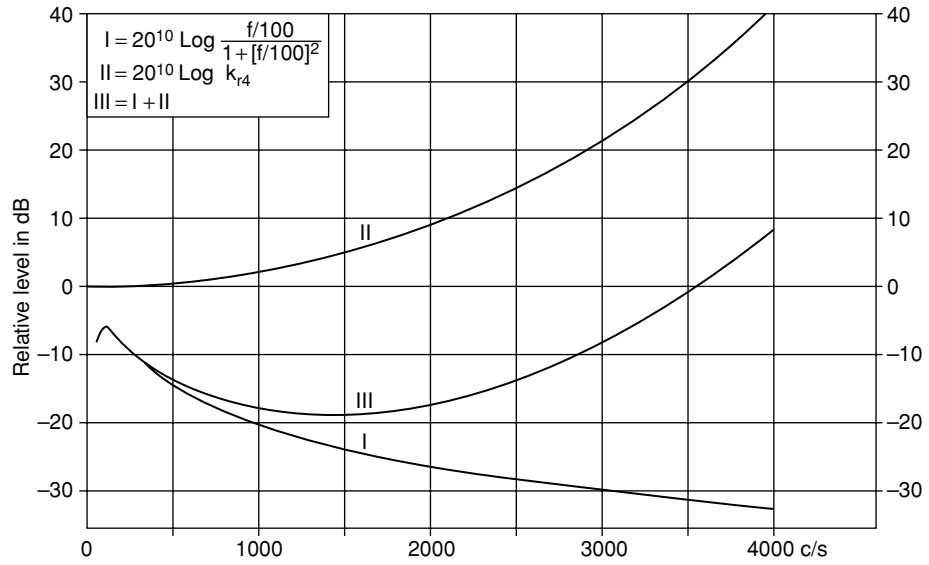


Figure 19. The constant frequency characteristics representing voice source, radiation transfer, and higher poles to be added to those of a four-formant resonance circuit analog.

The summation of these constant factors on a decibel scale is performed in *Fig. 19* pertaining to a voice source of $s_1 = s_2 = -2\pi \cdot 100\,\mathrm{sec}^{-1}$ and $K_{rr} = K_{r4}$. The magnitude and sharply rising character of the $K_{r4}(\omega)$-factor should be observed. It amounts to as much as 30 dB at 3,500 c/s.

The elementary resonance curve constituting the spectral contribution from a pair of conjugate poles has the form

$$|H_n(f)| = \frac{F_n^2 + B_n^2/4}{\left[(f - F_n)^2 + B_n^2/4\right]^{1/2}\left[(f + F_n)^2 + B_n^2/4\right]^{1/2}} \tag{2.7-1}$$

which may be conveniently memorized from the vectorial construction in the complex frequency plane. Under small loss assumptions the form given by Eq. 2.4-8 is useful:

$$|H_n(f)| = \frac{1}{\left[\left(1 - x_n^2\right)^2 + \delta_n^2 x_n^2\right]^{1/2}} \qquad x_n = f/F_n \tag{2.7-2}$$

The spectrum envelope of a voiced sound produced according to the assumptions above thus takes the form

$$L(f) = 20\,\log_{10}\left[\frac{K f}{1 + f^2/100^2} \cdot |K_{r4}(f)| \cdot \prod_{n=1}^{4} |H_n(f)|\right] \text{ (dB)} \tag{2.7-3}$$

where K determines the reference level. A decomposition of the neutral standard

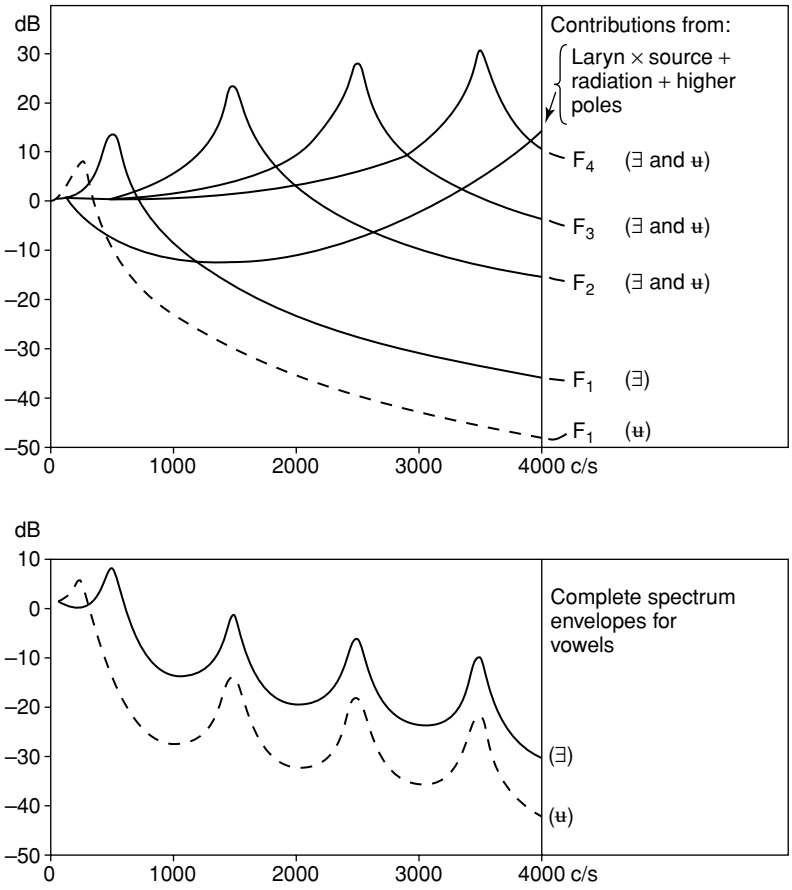


Figure 20. Summation of the separate resonance curves and the constant spectral characteristics of two vowels. One of these, [ʉ], differs from the other, [ə], by an octave lower frequency of the first formant resulting in a 12-dB spectrum-level loss at higher frequencies.

vowel [ə], $F_n = (2n - 1)$ 500 c/s, into the four resonance curves and the constant factor is demonstrated in Fig. 20. A shift down in F_1 from 500 c/s to 250 c/s causes a shift of the sound quality from [ə] to the [ʉ] as in Norwegian "hus". It may be seen that the spectrum level in [ʉ] is constantly 12 dB below that of [ə] at frequencies above 700 c/s. This shift, apparent from Eq. 2.7-2, may be referred to as the low-pass filtering effect of the first formant resonance curve on the rest of the spectrum. At frequencies well above F_1, $H_n(f)$ may be approximated by x_n^{-2} which implies an attenuation of the higher frequency region by 12 dB for every octave decrease in F_1.

A second rule relating formant frequencies and formant levels is that two formants that approach in frequency both gain 6 dB per halving of their frequency distance. The valley between the formants gains in level 12 dB as a result of this shift. This assumes that the two formants are appreciably closer to each other than to any other

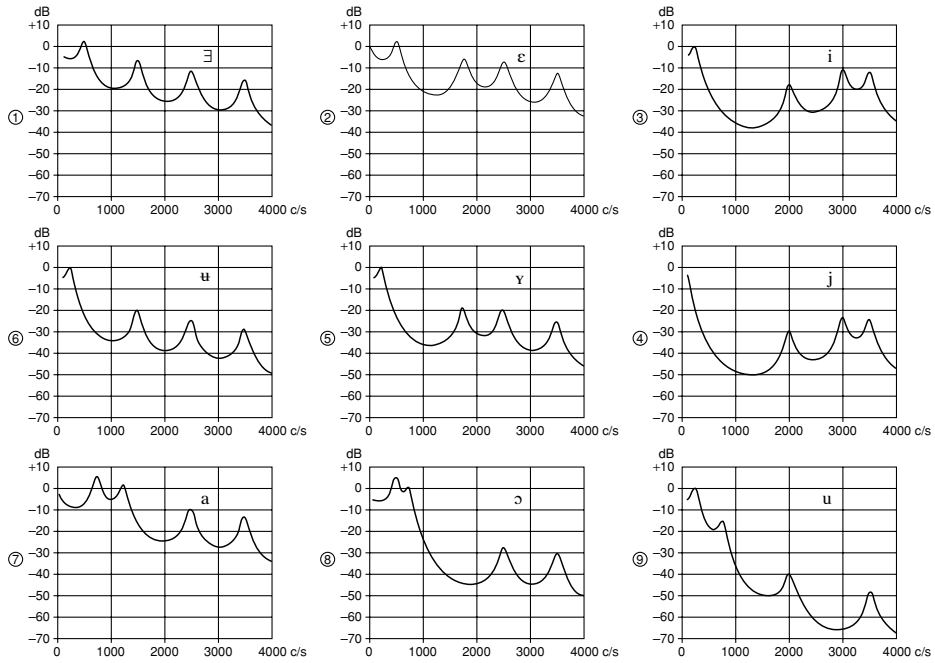


Figure 21. Spectrum envelopes synthesized numerically by the procedure of Fig. 20. Formant frequencies are varied in steps of 125 c/s in F_1, 250 c/s in F_2, and 500 c/s in F_3.

formant. If a single formant is shifted in frequency it will be necessary to study its vectorial relations to all points of the complex frequency plane.

These rules may be studied from the systematic synthesis performed in Fig. 21. Forman frequencies were shifted in quantal steps of minimal 125 c/s change in F_1, 250 c/s change in F_2, and 500 c/s change in F_2. All formant bandwidths were 100 c/s as in Fig. 20. It should be observed that the levels of the third and fourth formants of [u], L_2 and L_4, are very weak because both F_1 and F_2 have low frequency positions. In the [a]-spectrum on the other hand, L_3 and L_4 are appreciably higher due to the higher geometrical mean position of F_1 and F_2. The rise in L_1 and L_2 in the intermediate valley following the shift of F_2 and F_1 closer to each other is seen by comparing [ɜ] with [a]. Comparing [i] to [u] it may be seen how a shift up in frequency of F_3 and of F_2 at constant F_1 shifts the balance of spectral energy to the favor of the higher frequencies. These shape characteristics have been discussed in more detail in other publications (FANT, 1956, 1958a).

As an experimental check on these techniques an attempt has been made to match a measured vowel spectrum with a calculated spectrum envelope. Formant frequencies of $F_1 = 700$ c/s, $F_2 = 1,100$ c/s, $F_3 = 2,700$ c/s, and $F_4 = 3,300$ c/s were utilized. Bandwidths were chosen according to the empirical formula $B_n = 50(1 + f/2,000)$ c/s. The total length of the vocal tract was considered to be 16.7 cm, and the standard voice source spectum envelope of —12 dB/octave was adopted as in the previous

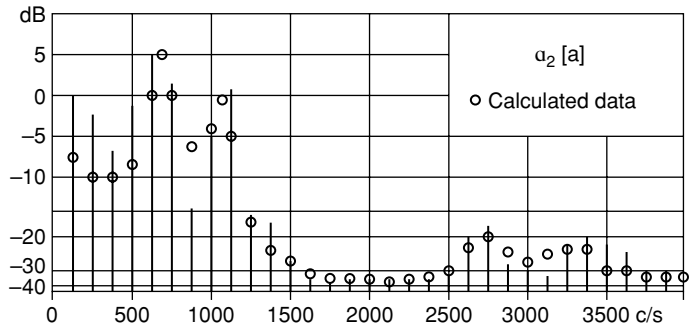


Figure 22. Spectrum of a sustained vowel a_2. The calculated points have been derived from a set of preselected formant frequencies.

examples. It may be seen from Fig. 22 that the general fit is good. There is some lack of base level in the calculated curve.

3.22. ACOUSTIC VOWEL DIAGRAMS BASED ON FORMANT PARAMETERS

The most extensive vowel measurements ever reported on are those performed at the Bell Telephone Laboratories by POTTER and STEINBERG (1950) and by PETERSON and BARNEY (1952). An attempt has been made to map Potter and Steinberg's data on to the Swedish vowel data, represented by the subject Gj-n. The main purpose of this matching was not to make phonetic comparisons on an acoustic basis but to check the general similarity of the formant patterns. As seen from *Fig. 27*, the fit is good enough in phonetically similar pairs as [o] $å_1$, [ɔ] a_1, [æ] $ä_3$, [i] i_1, that it may be concluded that F_1, F_2, and F_3 refer to the same formants and that there accordingly exists a basis for future detailed comparisons.

The investigation of PETERSON and BARNEY (1952) is the only survey providing data on both formant frequencies and formant levels besides the earlier Swedish measurements (FANT, 1948), discussed in this chapter. The sequence diagram of *Fig. 28* (earlier discussed by FANT, 1956) is intended to show that providing two vowels, one Swedish and one American English vowel, have approximately the same formant frequencies, they will also have approximately the same formant levels. This agreement exists in spite of the systematic differences involved; the Swedish vowels being sustained, the American English vowels sampled from mono-syllabic test words. Both sets of data pertain to the average figures for a group of males; several general observations of interest may be made from *Fig. 28*, e.g., the rapid increase of L_2 and L_3 within the series o_1 to a_2. L_2 and L_3 of front vowels are found to be of the same order of magnitude. Exceptions are $ö_3$ and u_2 which behave more like back vowels.

The relative formant levels within a spectrum determine the main shape of the spectrum envelope. Because of the analytical relations between formant frequencies and formant levels discussed in Section 2.7, there exists the theoretical possibility of

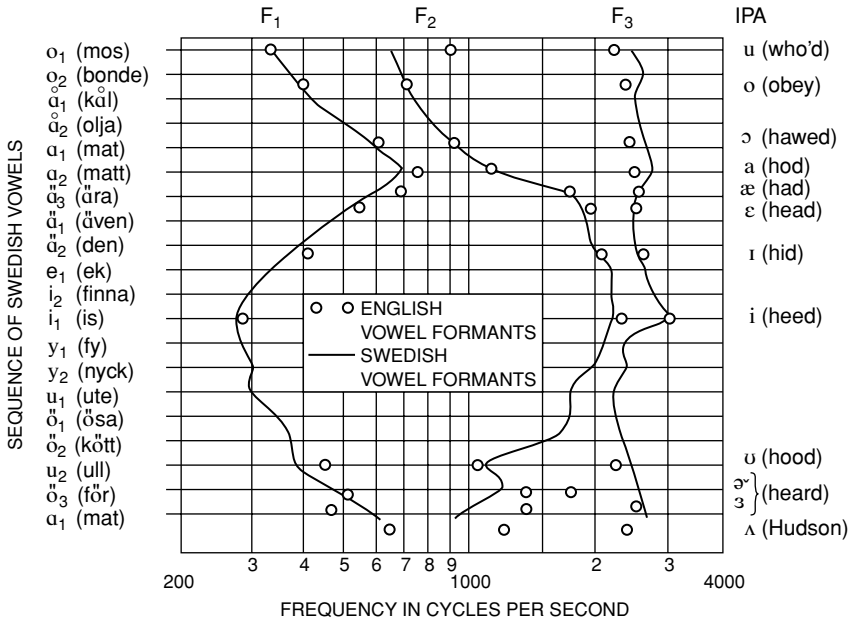


Figure 27. Sequential diagram of Swedish and American English vowels.

phonetic classification of vowels according to relative formant levels. However, the formant level information is related not only to articulation, but also to phonation, i.e., to the voice source characteristics, whereas the formant frequencies are related to articulation alone.

The spread of formant levels comparing different speakers is of the order of 4 dB in L_2–L_1 and 5 dB in L_3–L_1. From a perceptional point of view these numbers are not very large. According to FLANAGAN (1957 a) the minimum perceptual difference, DL, in second and third formant levels is $\pm$ 3 dB and $\pm$ 5 dB respectively. The DL of the first formant level is 1 dB, which also is valid for the overall loudness sensation of the vowel. It can thus be concluded that the range of L_2-variation, 25 dB, covers approximately 5 of the $\pm$ 3 dB DL, whereas the range of F_2 frequency variation, 1,500 c/s, contains approximately 15 of the $\pm$ 50 c/s frequency DL; see further FLANAGAN (1957 a).

There is one additional conformity comparing the American and Swedish data that should be mentioned. As shown by FANT (1953), the female versus male average differences in formant frequencies are similar even when a specific vowel is studied. The following tabulation summarizes the results of this study in terms of the frequency percentage relation of each of F_1, F_2, and F_3 of the female average data to corresponding male data. The 9 selected vowels are paired on an F-pattern matching basis. Numerical data on the male formant frequencies are also given below.

The agreement is appreciable in some instances, e.g., comparing the front vowels i_1, e_1, and $ä_3$ with the corresponding American vowels. The overall correlation is

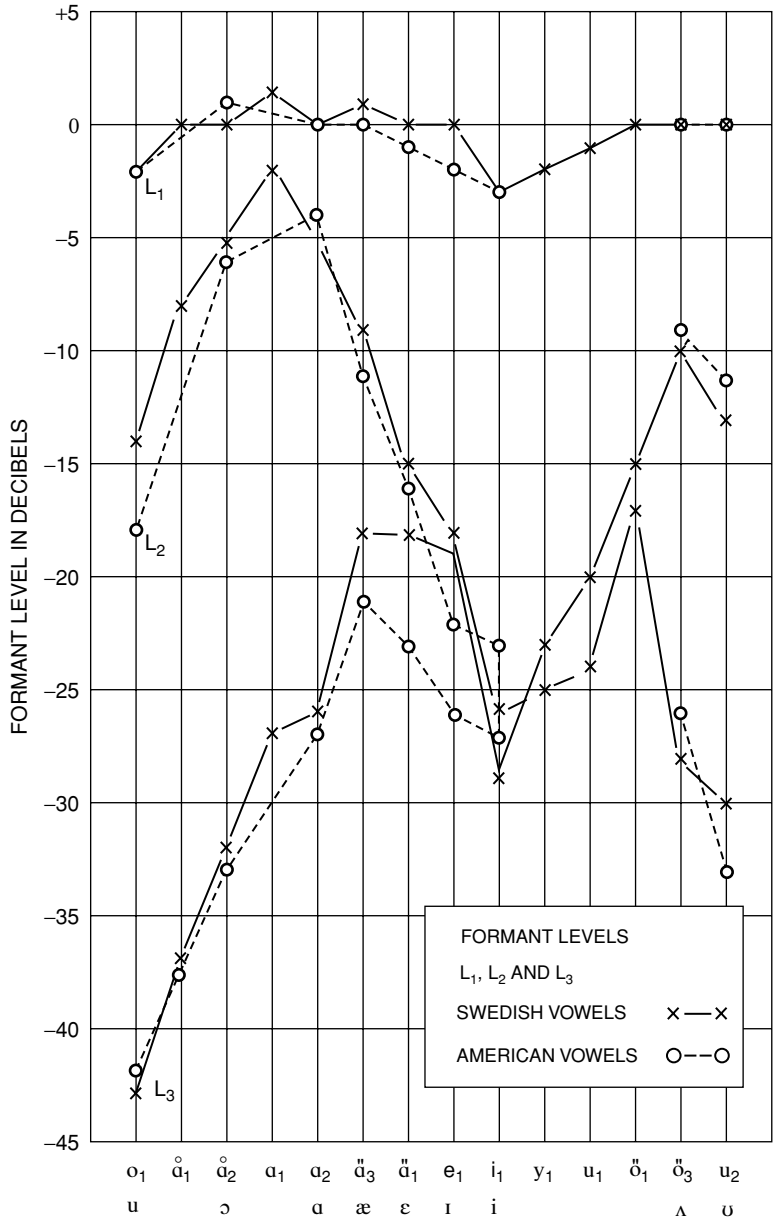


Figure 28. First, second, and third formant amplitudes of a sequence of Swedish and American English vowels.

significant. It suggests that the female-male differences are greater for formants of a standing wave origin, e.g., F_2 and F_3 of front vowels, and F_1 of the very open $ä_3$. The differences are smaller for F_1 and F_2 of back vowels, e.g., $å_1$, and for F_1 of close front vowels, e.g., i_1. This could be conceived of in the light of the partial applicability

TABLE 3.2-2
Average male F-patterns and the female/male frequency percentage

$$k = 100\left(\frac{F_{female}}{F_{male}} - 1\right) \quad \%$$

American English data (A) 33 subjects
Swedish data (S) 7 subjects

IPA	STA		F_1 c/s	k_1 %	F_2 c/s	k_2 %	F_3 c/s	k_3 %
[i]		(A)	270	10.0	2,290	22.0	3,010	10.0
	i_1	(S)	260	8.5	2,070	22.0	2,960	16.5
[I]		(A)	390	10.0	1,990	24.5	2,550	20.0
	e_1	(S)	330	9.5	2,050	24.0	2,510	17.5
[ɛ]		(A)	530	15.0	1,840	26.5	2,480	20.0
	ä$_1$	(S)	440	24.5	1,800	19.0	2,390	20.0
[æ]		(A)	660	30.0	1,720	19.0	2,410	18.5
	ä$_3$	(S)	610	29.5	1,550	17.5	2,450	20.0
[a]		(A)	730	16.5	1,090	12.0	2,440	15.0
	a_2	(S)	680	26.5	1,070	12.0	2,520	12.5
[ɔ]		(A)	570	3.5	840	9.5	2,410	12.5
	å$_2$	(S)	490	6.0	820	2.0	2,560	10.5
[U]		(A)	440	7.0	1,020	14.0	2,240	19.5
	u_2	(S)	420	−1.0	1,070	10.0	2,320	16.5
[u]		(A)	300	23.0	870	9.0	2,240	19.0
	o_1	(S)	310	10.5	710	−3.0	2,230	30.0
[Λ]		(A)	640	19.0	1,190	17.5	2,390	16.5
	ö$_3$	(S)	520	8.0	1,100	17.0	2,430	12.5

of the double or single Helmholtz resonator formula in these instances. A reduction of cavity volume can be compensated for by a narrowing of the associated orifice. In the case of standing wave resonances on the other hand, the length dimensions alone are crucial.

CHAPTER 2.2

THE RELATIONS BETWEEN AREA FUNCTIONS AND THE ACOUSTIC SIGNAL

GUNNAR FANT
Department of Speech Communication, Royal Institute of Technology, Stockholm

ABSTRACT

To derive speech wave data from area function specifications and the reverse, to predict the area function from the speech wave, are fundamental problems of acoustic theory of speech production. Deviations from ideal resonator theory in terms of vocal tract boundary conditions and source filter interactions are discussed. Perturbation theory is related to special problems of male-female vocal tract scaling. Shortcomings of the inverse transforms are discussed. Merits of lossy transmission line theory over standard linear prediction procedures are emphasized. The use of bandwidths for removing ambiguities is illustrated in simple models. A limited amount of bandwidth data supplementing formant frequency data and model related vocal tract constraints appears to be optimal.

INTRODUCTION

The topic of this paper is to discuss how configurations, shapes, and detailed outlines of the vocal tract cavity system influence the acoustic signal and the reverse, how to predict vocal tract resonator dimensions from speech wave data. As far as the direct transform is concerned, this is a revisit to my old field of acoustic theory of speech production.

What progress have we had in vocal tract modeling and associated acoustic theory of speech production during the last 20 years? My impression is that the large activity emanating from groups engaged in speech production theory and in signal processing has not been paralleled by a corresponding effort at the articulatory phonetics end. Very little original data on area functions have accumulated. The FANT [1960] Russian vowels have almost been overexploited. Our consonant models are still rather primitive and we lack reliable data on details of the vocal tract as well as of essential differences between males and females and of the development of the vocal tract with age.

The slow pace in articulatory studies is of course related to the hesitance in exposing subjects to X-ray radiation. Much hope was directed to the transformational mathematics for deriving area functions from speech wave data. These techniques have as yet failed to provide us with a new reference material. The so-called inverse transform generates 'pseudo-area functions' that can be translated back to high quality synthetic speech but which remain fictional in the sense that they do not

necessarily resemble natural area functions. Their validity is restricted to nonnasal, nonconstricted articulations and even so, they at the best retain some major aspects of the area function shape rather than its exact dimensions. However, some improvements could be made if more representative acoustic models than linear prediction (LPG) analysis are considered.

Once a vocal tract model has been set up it can be used, not only for studying articulation-to-speech wave transformations, but also for a reverse mapping of articulations and area functions to fit specific speech wave data. These analysis-by-synthesis remapping techniques, as well as perturbation theory for the study of the consequences of incremental changes in area functions or of the inverse process, are useful for gaining insight in the functional aspect of a model. However, without access to fresh articulatory data the investigator easily gets preoccupied with his basic model and the constraints he has chosen.

The slow advance we have had in developing high quality synthesis from articulatory models is in part related to our lack of reliable physiological data, especially with respect to consonants, in part to the difficulty involved in modeling all relevant facors in the acoustic production process. The most successful attempt to construct a complete system is that of FLANAGAN *et al*. [1975] at Bell Laboratories. A variety of studies at KTH in Stockholm and at other places have contributed to our insight in special aspects of the production process such as the influence of cavity wall impedance, glottal and subglottal impedance, nasal cavity system, source filter interaction, and formant damping.

FROM AREA FUNCTION TO THE ACOUSTIC SIGNAL

The acoustic signal or, in other words, the speech wave is the product of a source and a filtering process. The most common approach is to disregard the source and relate a vocal tract area function to a corresponding formant pattern only, i.e. a set of formant frequencies F_1 F_2 F_3 F_4, etc. This correspondence is illustrated by figure 1. I shall not go into the mathematics of the wave equations and the equivalent circuit theory. Instead I will attempt to develop a perspective around some basic models and current problems.

To derive an area function from X-ray data on vocal tract dimensions is by no means a straightforward procedure, see FANT [1960; 1965] and LINDBLOM and SUNDBERG [1969]. The estimation of cross-sectional shapes and dimensions in planes perpendicular to the central pathway of propagation through the vocal tract has to rely on crude conventions and involves uncertainties, e.g. with respect to variations with articulation and for different types of subjects. The lack of basic data is especially apparent for female and child speech and for consonants, e.g. laterals and nasals. In spite of the accessibility of the speech wave to quantitative analysis there is a similar lack of reference data concerning the acoustic correlates. Most studies have been concerned with male speech and vowels.

A specification of an area function as a more or less continuous graph of cross-sectional area from the glottis to the lips allows detailed calculations of the acoustic response but is not practical for systematic descriptions. A data reduction in terms of

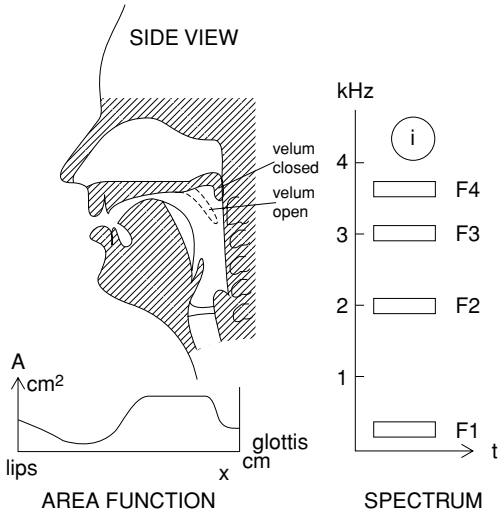


Figure 1. Principle illustration of vocal tract sagittal view with area function and corresponding resonance frequency pattern.

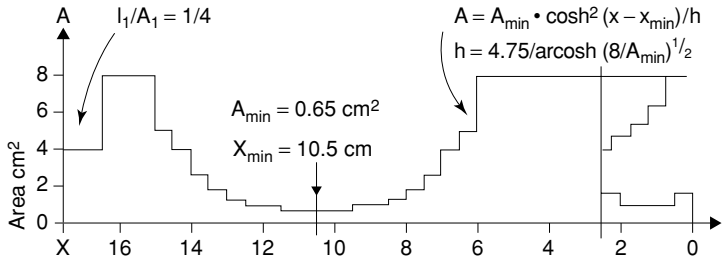


Figure 2. Three-parameter vocal tract model [FANT, 1960]. X = Constriction coordinate in centimeter from glottis.

parametric models brings out the acoustically relevant aspects. The three-parameter models of STEVENS and HOUSE [1955] and FANT [1960] differ somewhat in the details but have the same set of parameters, the place of minimum cross-sectional area of the tongue section, the area at this coordinate, and the length over area ratio $1_0/A_c$ of the lip section.

My model is shown in figure 2. The shunting sinus piriformis cavity around the outlet of the larynx tube was a constant feature in my model. A weakness is that it is not reduced in volume for back vowels which does not allow F_1 to reach a sufficiently high value for [a]. Figure 3 shows the variation of the F-pattern with the place of tongue constriction. This is a well-established graph which retains basic patterns such as the rise of F_2 with advance of the tongue constriction from back to front up to an optimal place at a midpalatal location after which F_2 drops again. A limitation of the parameter range to a region bounded by [a], [u], and [i] as proposed

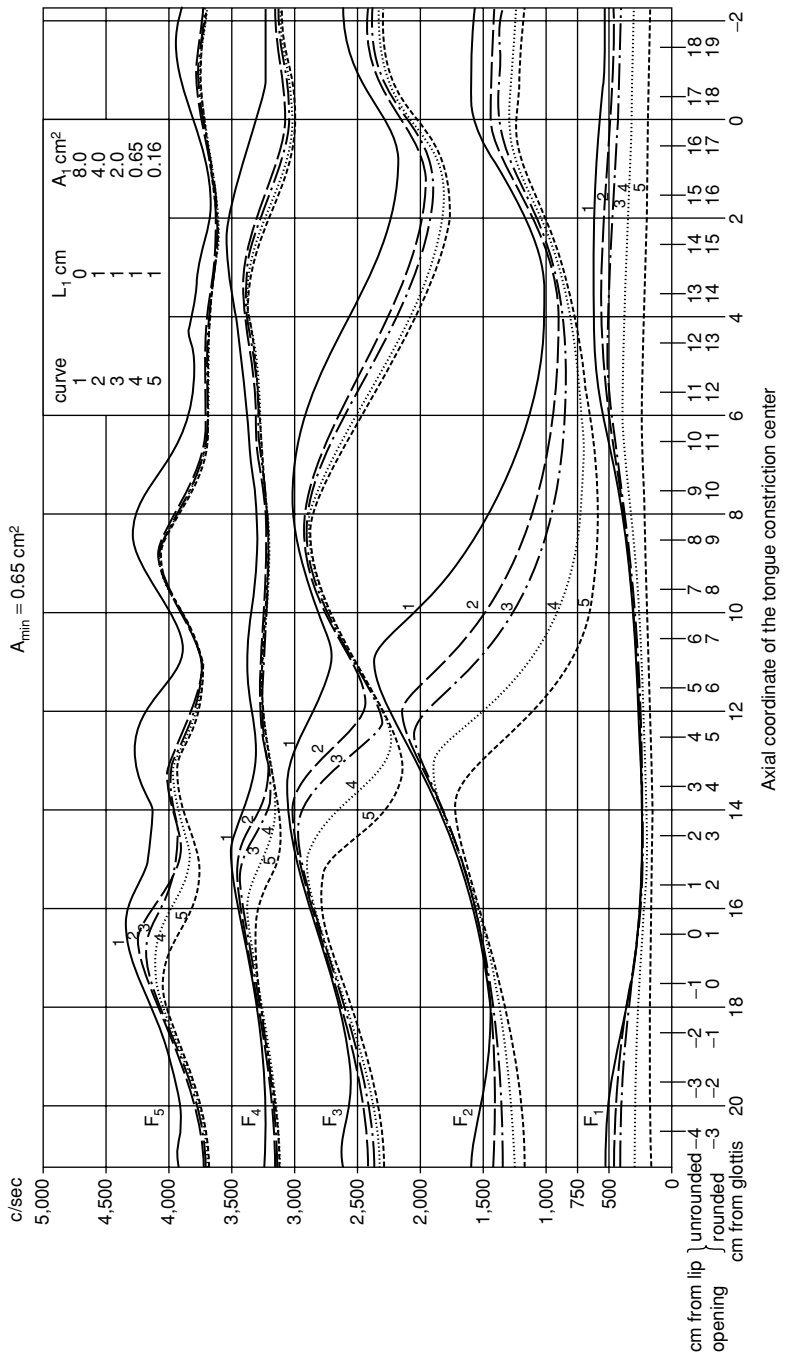


Figure 3. F-pattern variation with constriction coordinate x_c at different sets of lip parameter l_1/A_1 at constant constriction area A_{min}. The constriction coordinate is zero at the glottis.

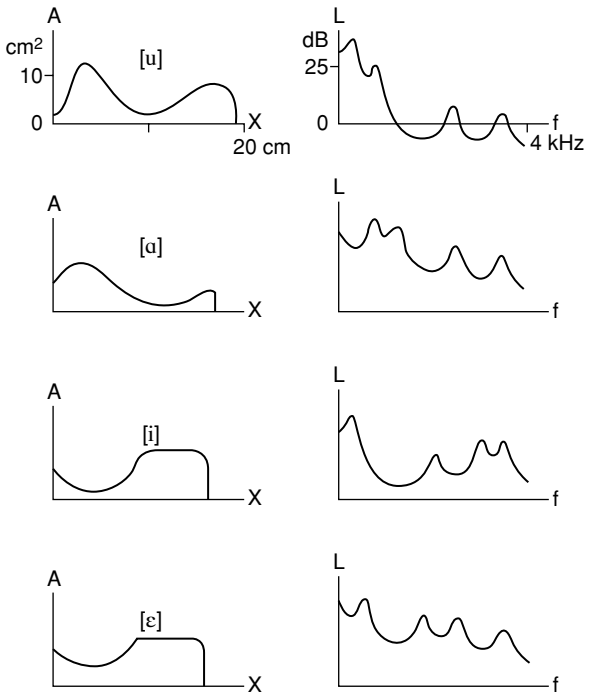


Figure 4. Stylized area functions and corresponding spectrum envelopes of [u], [*a*], [i] and [ɛ]. The constriction coordinate is zero at the lips.

in several articulatory models, e.g. LINDBLOM and SUNBERG [1969], would exclude the standard Swedish pronunciation of the vowel [ʉ] which, contrary to traditional classifications, has a constriction somewhat anterior to that of [i] [FANT, 1973].

The constriction coordinate is an acoustically more relevant classifier than the 'highest point of the tongue' of classical phonetics. Most stressed vowels have a definite 'place of articulation' as evidenced by a region of minimum cross-sectional area which we may exemplify by [i], [u], [ɔ], [*a*] ending with a variant of [æ] with-major narrowing just above the glottis [FANT, 1960]. On the other hand, it may be argued that the traditional classification in terms of tongue locations and related parameters belongs to a production stage one step higher up than area functions and could be directly related to formant patterns.

The [a] and [i] vowels are polar opposites, the [i] vowel requiring a wide pharynx and narrowed mouth, whilst the opposite is true of [a] type vowels. A production of a vowel [u] requires a double resonator configuration with a narrow lip opening to ensure a low F_1 and a narrow constriction between the two major cavities as a correlate of a low F_2. These shape aspects are brought out in the stylized area functions of figure 4. A basic issue in acoustic phonetics is that it is not possible to produce these vowels without retaining the major shape aspects of the area functions. To this extent area functions are predictable from the acoustic signal as will be

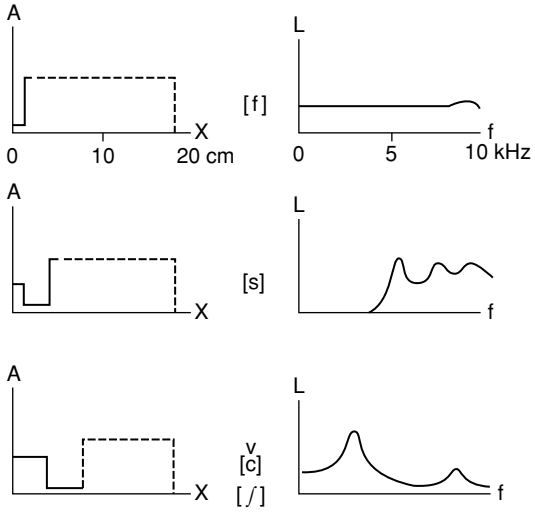


Figure 5. Stylized area functions and corresponding spectra of three basic consonant categories. The constriction coordinate is zero at the lips.

discussed in greater detail in a later section. PETER LADEFOGED would back me up here with his competence of transforming phonetic qualities to equivalent resonator configurations.

Another basic issue is that the vocal tract filtering is determined by the location of formants only and that the spectrum envelope between peaks cannot contain any other irregularities than those originating from the source function. Minor irregularities in the outline of the area function may have some influence on formant locations but will not give rise to irregularities in the spectrum envelope. This is not evident without an insight in the mathematical constraints imposed by acoustic theory. It is related to the one-dimensional wave propagation, wavelengths generally being short compared to vocal tract cross dimensions. Systematic perturbations of vocal tract area functions will be discussed in a later section.

Highly simplified area functions of fricatives (or corresponding stops) and their filtering functions are shown in figure 5. As discussed by FANT [1960], the 'compact' sibilant [$\int$] or the stop [k] has a definite cavity in front of the major constrictions which accounts for a central dominance of the spectrum, usually a single formant, if the cavity is abruptly terminated by the constriction. The [s] or [t] has a narrow channel of a few centimeters length behind the source which may combine with a small front cavity to produce resonances above 4000 Hz which build up a high-pass filtering. The [f] or [p] has no significant resonance in its closed state.

In general, the cavities behind the source do not influence the spectrum much, provided that the consonantal constriction is effective. Resonances of the back cavities may appear if the constriction tapers off gradually as in palatals or if a palatal tongue articulation builds up a supporting constriction behind the lips. Back cavity

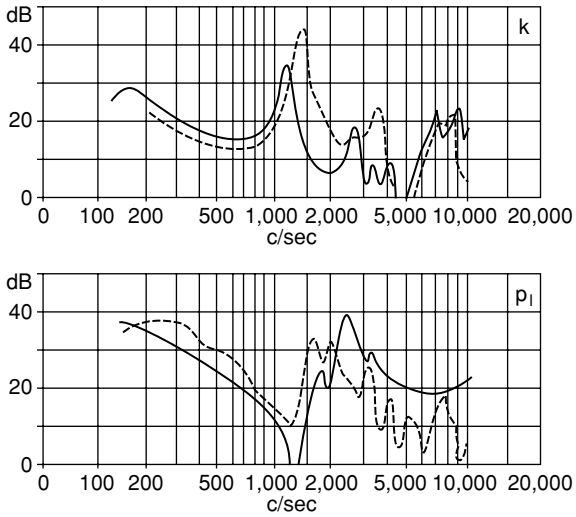


Figure 6. Calculated (—) and measured (- -) stop release spectra of a velar [k] and a palatalized [p′]. The minimum in [p′] at 1,400 Hz is a free zero in the sublip impedance whilst the main formant of [k] is a mouth cavity formant. After FANT [1960].

resonances combine with and are cancelled by spectral zeroes at complete closure but move away from their zero mates during release and are then more or less free to appear. In figure 6 we can study measured and calculated spectra of [k] and a palatalized [p′] [FANT, 1960]. The labial burst spectrum contains peaks at around 2–3 kHz but has a free spectral minimum at 1400 Hz. In contrast, the [k] spectrum has a single formant peak around 1400 Hz. It is interesting to note that the calculations from the area function data back up the measured spectra. We need more studies of this type.

VOCAL TRACT BOUNDARY CONSTRAINTS AND DYNAMICS

The simplified static models relating a single area function without parallel branches to a set of formant frequencies have obvious limitations. On a higher level of ambition we must include proper boundary conditions such as radiation load and a finite coupling to the subglottal and nasal systems. In order to predict formant bandwidths we must consider the energy loss during an oscillatory cycle of a formant associated with 'loss elements' on the surface of the vocal tract resonator system and other dissipative elements [FANT and PAULI, 1975]. Source functions must be defined with respect to place of insertion in the vocal tract, their spectrum or waveform, and the degree of coupling to other parts of the system [STEVENS, 1971]. In addition, these properties are highly time variable within a voice fundamental period [FANT, 1979] and within intervals of transition from various states of the glottis or of other terminations of the vocal tract. Rapid opening and closing gestures pose specific

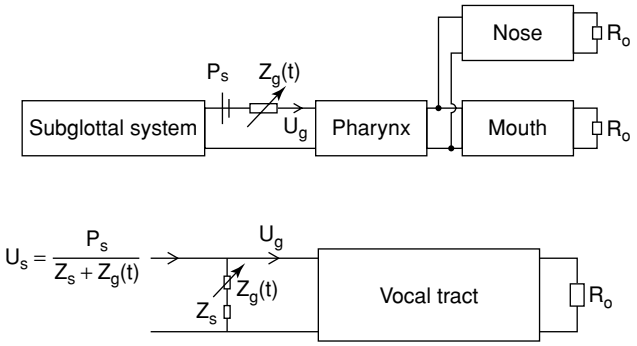


Figure 7. Block diagram of the production of voiced sounds illustrating the principle difference between an ideal source U_s and a glottal flow U_g.

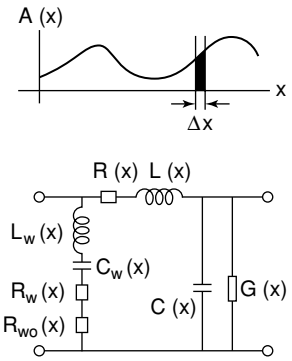


Figure 8. Lumped constant approximation of a small slice of the area function.

problems in relating area functions to acoustic data. In a proper analysis of connected speech we need two sets of acoustic variables: the continuous variations of the F-pattern as a correlate of the continuous movements of the articulators and the often abruptly varying patterns of spectral energy distributions associated with discrete events of production.

The acoustic production model of figure 7 may serve as a starting point for a brief discussion of these problems. First of all, we should note an important element in converting area functions to a filter function. The walls of the vocal tract are not rigid. They may expand during a voiced occlusion as represented by the element C_w in the equivalent circuit of a small slice of the area function (fig. 8), and they have a finite mass L_w which adds to the tuning of vocal resonances and which dominates the impedance of the shunting branch at frequencies above 40 Hz. A small fraction of sound is radiated externally from the outside of the head through R_{wo}. It is negligible except as a constituent of the voice bar of a voiced occlusion.

Disregarding the cavity wall mass element L_w, calculations would provide $F_1 = 0$ for an area function starting and ending with complete closure. The finite F_1 of around

150–250 Hz found in the spectrogram of the voiced occlusion is determined by the resonance of the entire air volume compliance in the tract with the total lumped cavity wall mass shunt. This resonance can easily be measured acoustically [FANT *et al.*, 1976] and amounts to $F_{1w} = 190$ Hz with a bandwidth of $B_{1w} = 75$ Hz, typically for a male voice, and around 20% higher for females. The wall mass element L_w is thus an important constituent in calculating F_1 from the area function. The procedure is to start out with a derivation of an ideal F_{1i} without mass shunt and add a correction factor

$$F_1 = F_{1i}(1 + F_{1w}^2/F_{1i^2})^{1/2} \quad (1)$$

The distribution of the wall impedance along the vocal tract and its dependence on particular articulations are not known. The experiments of FANT *et al*. [1976] suggest that regions around the larynx and the lips are especially important. Experiments by ISHIZAKA *et al*. [1975] provide data of the same order of magnitude but have not revealed conclusive distribution patterns.

The resistive component R_w in the cavity wall branch determines a major part of the bandwidth B_1 of low F_1 formants. The resistive part of the radiation load which is proportional to frequency squared is the essential bandwith determinant of resonances above 1000 Hz originating from an open front resonator. Internal surface losses from friction and heat conduction enter through the elements R and G in figure 8. They are proportional to the half power of frequency and to the inverse of the cross-sectional area. A detailed analysis of formant bandwidths and their origin appears in FANT [1972], FANT and PAULI [1975], and WAKITA and FANT [1978].

The time variable glottal impedance accounts for variations of formant frequencies and bandwidths within a voice fundamental period [FLANAGAN, 1965]. A more detailed analysis of glottal damping requires a reconsideration of the process of voice generation [FANT 1979] and adoption of perceptual criteria for deriving equivalent mean values [FANT and LILJENCRANTS, 1979]. The main excitation of the vocal tract occurs at the instant of interruption of glottal flow by glottal closure. At this instance, damped oscillations are evoked and subjected to the damping from supraglottal loss elements.

When the glottis opens for the next flow pulse the vocal tract becomes loaded by the time variable glottal plus subglottal impedance. Providing a resonance mode is much dependent on the part of the area function immediately above the glottis, the glottal damping becomes severe. This is especially apparent if the lower pharynx is narrowed thus facilitating an impedance match between the cavity system and the glottal resistance. A complete extinction of the formant oscillation in the glottal open interval may result. This is typical of F_1 of the vowel [a] produced at low or moderate voice effort by a male subject.

In general most of the energy excited during a voice fundamental period is lost during the timespan of the following period. Since glottal resistance decreases with lowered transglottal pressure the damping effect is especially apparent at weak voice levels. The mean glottal bandwidth in normal voice production is of the order of 0–100 Hz with 20 Hz as a typical value for male medium intensity phonation.

It is apparent that any model of voice production which adopts the actual flow through the glottis as the primary source will create problems. With this convention, which happens to apply to inverse filtering techniques, the source attains components of formant oscillations and becomes dependent of the vocal tract area function [MRAYATI and GUÉRIN, 1976]. Their approach is intended to define a proper source for a formant synthesizer.

A different approach more suited for production models is to incorporate the combined glottal and subglottal impedance as a termination paralleling the input end of the tract and to define the source as the flow through the glottis which would have occurred with the input to the vocal tract short circuited. This representation adopted by FANT [1960] preserves a realistic definition of the vocal tract transfer function but fails to take into account source modifications due to aerodynamic losses in supraglottal constrictions. In the transition from a vowel to a voiced consonant there is generally some loss of transglottal pressure which reduces the excitation strength of the voice source.

The interplay of glottal and supraglottal sources associated with articulatory narrowing and release becomes an important part of a dynamic oriented theory of predicting acoustic signals from area functions [STEVENS, 1971].

What about the subglottal system? How does it influence speech? In normal voice production the influence appears to be small. As long as the glottal opening is small and the flow velocity high, the glottis impedance becomes high compared to the subglottal impedance. Unless there is a constant leakage bypassing the vibrating part of the glottis, the subglottal system should have a minor influence only.

This reasoning is concerned with the modification of the supraglottal formants only. At the instance of flow interruption when the glottis closes there is a simultaneous excitation of resonances in the trachea and other parts of the subglottal system. Potential frequencies are 600, 1250, and 2150 Hz for a male voice [FANT *et al.*, 1972]. The transmission losses associated with the penetration of these components through the walls of the trachea and the chest to externally radiated sound appear to be sufficiently high to rule out any significance, but this remains to be proved.

As shown by FANT *et al.* [1972], subglottal formants may occasionally be seen in spectra from aspirated sound segments, e.g. in the release phase of unvoiced stops. 'F_1-cutback' in the first part of the voiced interval after release, which appears as a relative delay in onset of F_1 compared to F_2 and higher formants, may be explained as an instance of excessive F_1 damping through an incompletely closing glottis. The upper formants are less dependent on the glottal termination and thus less affected. This relative weakening of F_1 is a filtering effect, whilst the relative weakness of F_1 in a preceding unvoiced, aspirated segment is also a matter of low source energy in the F_1 region. The F_1 intensity reduction is also seen in the terminating periods of a vowel before the occlusion of an unvoiced stop (preocclusion aspiration).

Nasalization and aspiration have similar effects on F_1. In nasalized sounds the F_1 intensity is typically reduced by a spectral zero [FANT, 1960; FUJIMURA and LINDQVIST, 1971]. The nasal model of FANT [1960] produces too high values of the lowest nasal pole. The possible occurrence of several low frequency pole-zero

pairs is made plausible by the study of LINDQVIST and SUNDBERG [1972]. More anatomical and acoustic data are needed.

In connection with the voice source studies of FANT [1979] it has been noted that the spectral maximum often seen below F_1 in vowels is a voice source characteristic, which becomes especially enhanced in contrast to a weak F_1 in nasalized or aspirated, voiced segments. This is especially apparent in a time domain study. Another way of expressing this finding is to say that nasal sounds retain more source characteristics than nonnasal sounds.

If an area function is subjected to a substantial change in a very short time, one may expect some deviations from the linear stationary behavior. Point-by-point calculations of resonance frequencies are still valid but additional bandwidth terms enter which may be positive or negative. A rapid opening of a constriction is accordingly associated with a negative bandwidth component and a rapid closure with a positive bandwidth component. The analysis is simple. Consider a flow U(t) through an acoustic inductance $L(t) = \varrho 1/A(t)$. The pressure drop is:

$$P(t) = \frac{d}{dt}[L(t)U(t)] = L'U + LU' \tag{2}$$

$L' = dL/dt$ apparently has the dimension of a resistance R_d

$$R_d = \frac{dL}{dt} = \frac{-A'(t)\varrho 1}{A^2(t)} \tag{3}$$

In a single resonator system the bandwidth component associated with a resistance R in series with an inductance L is simply $R/2\pi L$.

Accordingly, the bandwidth associated with R_d is

$$B_d = \frac{-A'(t)}{2\pi A(t)} \tag{4}$$

which implies a bandwidth component of opposite sign to that of the rate of change of the area. Figure 9 illustrates the temporal course of the bandwidth when a resonator of volume 100 cm^3 is coupled to a neck of 4 cm in length and a cross-sectional area A(t) varying exponentially from closure to complete opening of 2 cm^2 with a time constant of 10 msec.

The time varying negative bandwidth overrides the frictional bandwidth components up to 8 msec after release which could tend to increase the amplitude of the oscillation during that period. However, in the speech case there enter additional positive bandwidth components related to flow-dependent resistance and to cavity wall losses and possibly also glottal losses which tend to reduce the importance of the negative terms. In a detailed analysis of the glottis resistance the dynamics calls for some decrease of glottal resistance in the rising branch of the glottal pulse and an increase in the falling branch, as noted by GUÉRIN *et al.* [1975]. Except for the analysis above, a proper evaluation of the practical significance has to my knowledge not been performed. The most detailed thesis on the theoretical aspects is that of JOSPA [1975]. I feel that dynamic effects are of academic rather than practical

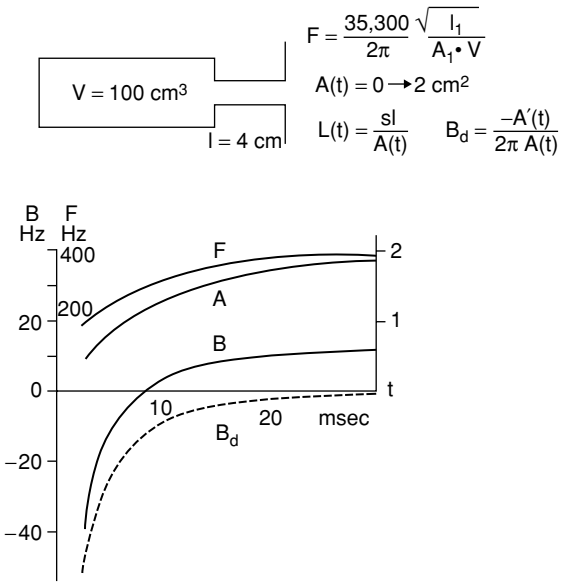


Figure 9. Resonator outlet area A, resonance frequency F, and total bandwidth B as a function of time during an exponential release with a time constant of 10 msec. B_d is the negative dynamic component of the bandwidth.

significance. Of greater importance is probably the mere fact that a rapid transition of a formant creates a special perceptual 'chirp' effect.

PERTURBATION THEORY AND VOCAL TRACT SCALING

Perturbation theory describes how each resonance frequency, F_1 F_2 F_3, etc., varies with an incremental change of the area function A(x) at a coordinate x and allows for a linear summation of shifts from perturbations over the entire area function. The relative frequency shift $\delta F/F$ caused by a pertubation $\delta A(x)/A(x)$ is referred to as a 'sensitivity function'. We may also define a perturbation $\Delta x/\Delta x$ of the minimal length unit Δx of the area function which will produce local expansions and contractions of the resonator system. It has been shown by FANT [1975b] and FANT and PAULI [1975] that the sensitivity function for area perturbations of any A(x) is equal to the distribution with respect to x of the difference E_{kx}–E_{px} between the kinetic energy $E_{kx} = \frac{1}{2}L(x)\ U^2(x)$ and the potential energy $E_{xp} = \frac{1}{2}C(x)P^2(x)$ normalized by the totally stored energy in the system.

Figure 10 from SCHROEDER [1967] illustrates perturbations of a single tube resonator by changes in the area function derived from sinusoidal functions. These have been chosen to influence F_1 only (a), none of the formants (b), and F_2 only (c). The middle case is of special interest. There exists an infinite mumber of small perturbations applied symmetrically with respect to the midpoint of the single tube, which will have almost no influence on the formant pattern. In the general case of an

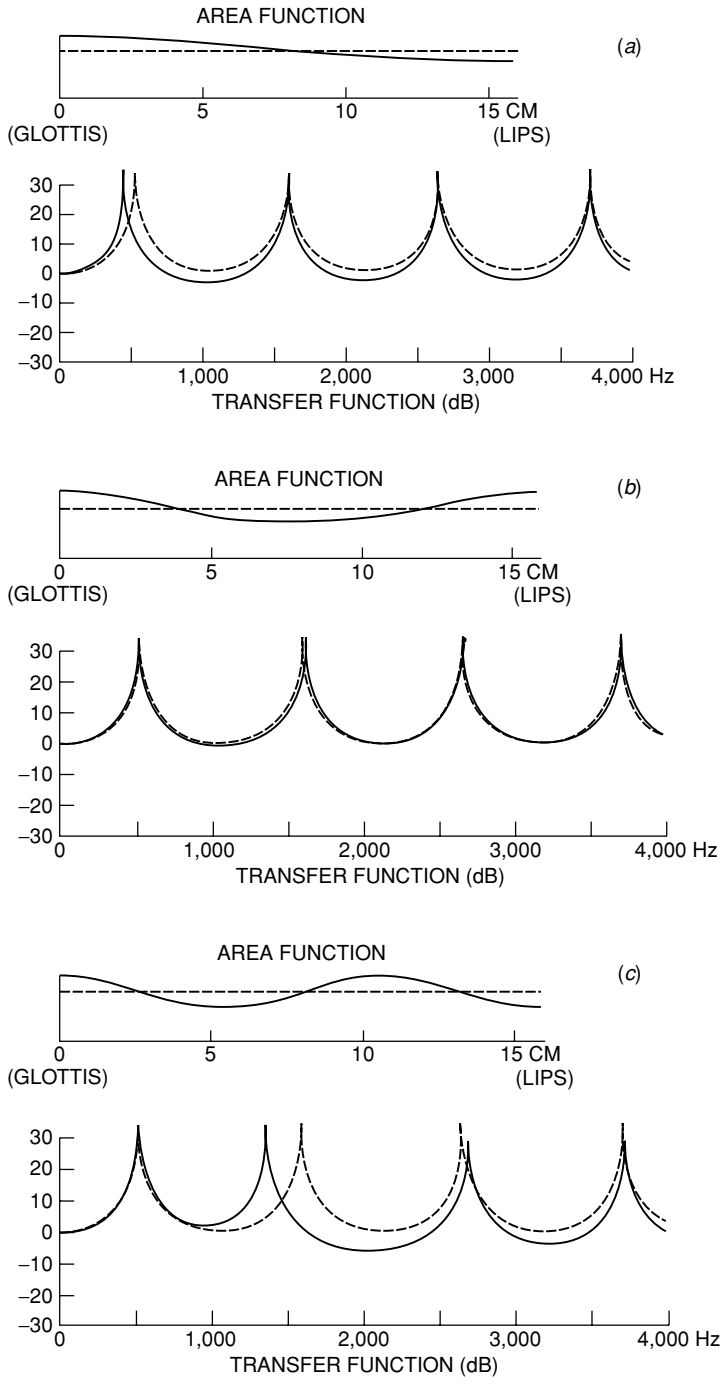


Figure 10. Perturbations of the single tube area function affecting F_1 only (a), having almost no influence (b), and affecting F_2 only (c) [after SCHROEDER, 1967].

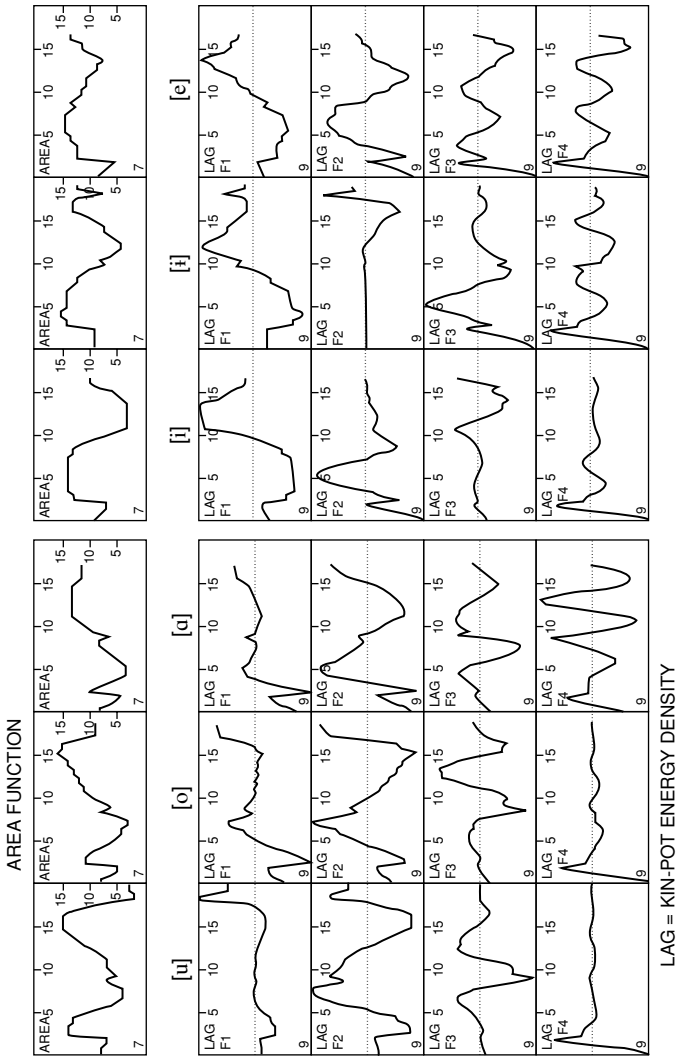


Figure 11. Sensitivity functions for area perturbations of the six Russian vowels [FANT, 1960]. From FANT [1975b]. The constriction coordinate is zero at the glottis.

arbitrary area function the rule of symmetry is upset [HEINZ , 1967] but there still exists a tendency of compensatory interaction between front and back parts [ÖHMAN and ZETTERLUND , 1975].

Sensitivity function for area perturbations of my six Russian vowels are shown in figure 11. This chart is useful as a reference for general use. Given the relative amount of area change, the corresponding relative frequency shift $\delta F_n/F_n$ is proportional to the product of $\frac{\delta A(X)}{A(X)}$ and the amplitude of the sensitivity function, E_{kx}—E_{px}. As an

example we may note that F_1 of the vowel [u] rises with increasing area at the lips, i.e. decreases with increasing degree of narrowing and that narrowing the tongue constriction of [u] causes F_2 to fall and F_3 to rise. A narrowing of the outlet of the larynx tube will apparently have the effect of tuning F_4 to a lower frequency.

With the area function sampled at intervals of Δx, e.g. $\Delta x = 0.5$ cm for practical use, we may ask what happens if we increase Δx at the coordinate x by the amount δx. The local expansion thus introduced causes a frequency shift $\delta F_n/F_n$, which is proportional to $-\delta(x)/[1 + \delta(x)]$ and to $(E_{kx} + E_{px})$ of resonance n.

The distribution of $(E_{kx} + E_{px})$ is uniform for a single tube resonator. The effect of a length increase is obvioulsly the same irrespective of where along the x-axis the tube is lengthened. An overall increase of the length by, say $\delta(x) = 0.2$, causes a shift of all resonances by a factor $-0.2/(1 + 0.2) = -0.17$. The same calculation performed directly from the resonance formula $(2n - 1)c/4l_t$, where l_t is the total length and $c = 35300$ cm/sec is the velocity of sound, would provide the same answer, i.e. a frequency ratio of $1/(1 + 0.2) = 0.83$.

The distribution $E_{kx} + E_{px}$ along the vocal tract is also a measure of the relative dependence of the particular resonance mode on various parts of the area function. This is the best definition we have of 'formant-cavity' affiliations. From figure 12 we may thus conclude that most of the energy of the second formant of [i] is stored in the pharynx, whilst the third formant of [i] 'belongs to' the front part of the system. F_3 of the back vowels [u] [o] and [a] are associated with a central part of the tract, and F_4 of all vowels has a substantial peak of energy located in the larynx tube. Expanding the length of the pharynx will have a large effect on F_2 of [i] and a small effect only on F_3 and vice versa for a length expansion of the mouth cavity. This analysis would apply to the relatively short pharynx of females compared to males.

If a perturbation of the entire area function is expressed as a function of as many parameters as there are formants, it is possible to calculate the change in area function from one F-pattern to another [FANT and PAULI, 1975]. This indirect technique has been used by MRAYATI *et al.* [1976] for deriving plausible area functions for French vowels. This procedure must be administered in steps of incremental size with a recalculation of the sensitivity function after each major step. It may involve length as well as area perturbations.

In practice, when aiming at direct transforms only, it may be easier to resort to a direct calculation of the response of the perturbed area functions than to derive it from the energy distributions. The perturbation formulas and especially their energy-based derivations are more useful for principal problems of vocal tract scaling or for gaining an approximate answer to a problem without consulting a computer program.

The area functions of male and female articulations of the Swedish vowels [i] and [u] and corresponding computed resonance mode pattern in figure 13 may serve to illustrate some findings and problems. The data are derived from tomographic studies in Stockholm many years ago in connection with the study of FANT [1965; 1966] and were published in FANT [1975a; 1976]. It is seen that in spite of the larger average spacing of formants in the female F-pattern related to the shorter overall vocal tract length, the female F_1 and F_2 of [u] and the F_3 of [i] are close to those of the male.

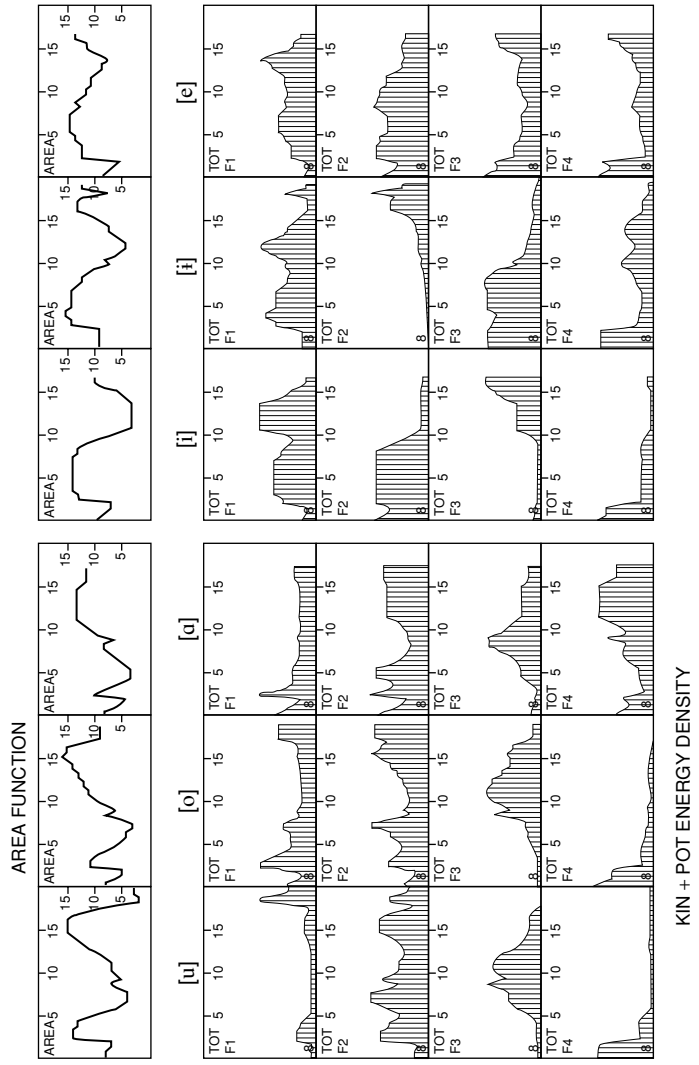


Figure 12. Sensitivity functions for length perturbations of the six Russian vowels [FANT, 1960]. From FANT [1975b]. The constriction coordinate is zero at the glottis.

This is an average trend earlier reported by FANT [1975a] (see fig. 14). Differences in perceptually important formants may thus be minimized by compensations in terms of place of articulation and in the extent of the area function narrowing. Such compensations are not possible for all formants and cannot be achieved in more open articulations. The great difference in F_2 of [i] is in part conditioned by the relatively short female pharynx but can in part be ascribed to the retracted place of articulation. It is also disputable whether this particular female articulation serves

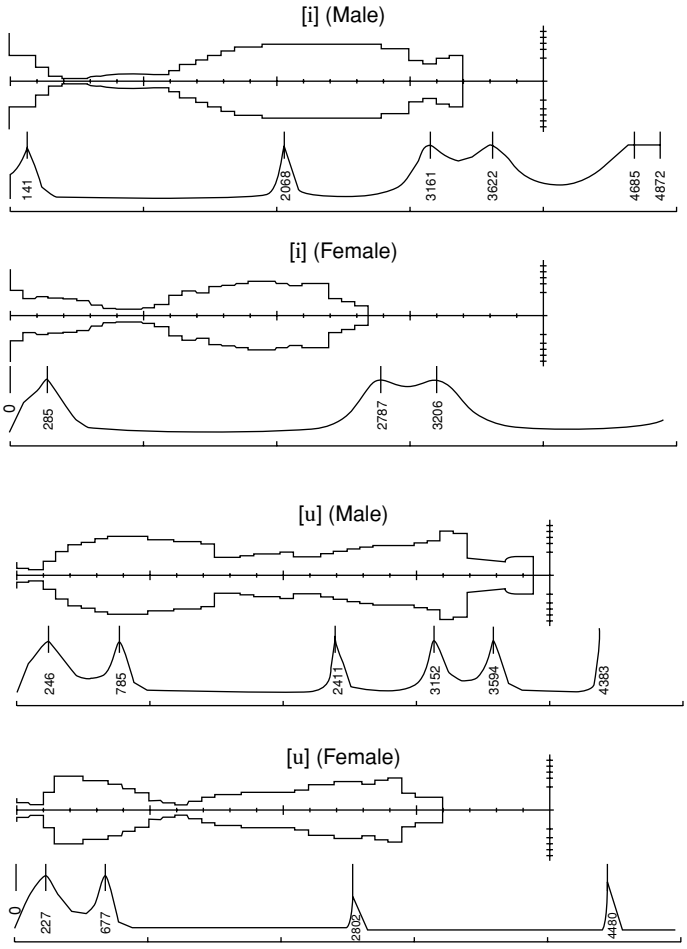


Figure 13. Male and female vocal tracts (equivalent tube representation) and corresponding F-patterns from the tomographic studies of FANT [1965].

to ensure an acceptable [i] or whether there is a dialectal trend towards [j]. Also, it is to be noted that X-ray tomography may impede the naturalness of articulations because of the abnormal head position required.

Much remains to be studied concerning how the vocal tract area functions of males, females, and children are scaled in actual speech and what kind of compensation occurs for minimizing perceptual differences or maybe the reverse, to mark contrasts between age and sex groups.

The lack of reference data on area functions is severe and the attempts to overcome this lack by means of area function scaling performed by NORDSTRÖM[1975] were not conclusive, except to support the general issue that the vowel and formant specific female-male differences, documented by FANT [1975a] (Fig. 14), do not always

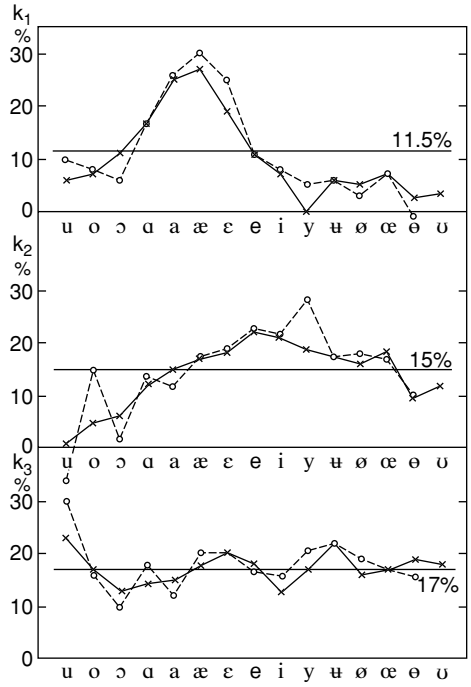


Figure 14. Female/male scale factor variation with vowels and the particular formant [FANT, 1975a]. o = FANT [1959]; x = 1 − 6 languages.

come out as a result of the particular scaling assumed. The agreement was good for F_3 and fair for F_2 and rather bad for F_1. The predictability of F_3 is expected in view of the high dependency of F_3 on length dimensions.

A weakness in the NORDSTRÖM study is that his [ae] and [ɛ] vowel area functions were interpolated from the Russian [a] and [e] vowel and accordingly attain a centralized quality not representative of the [a] and [ae] category vowels which normally display a very large female-to-male F_1 ratio (see fig. 14).

It is interesting to note that the nonuniform differences between females and males are paralleled by similar patterns comparing tenor and bass male singers. These vowel and formant specific trends are not only the automatic consequence of different anatomical scalings but also reveal compensations according to criteria that are not very well understood yet. A promising project on vocal tract modeling from anatomical data, now carried out at MIT [GOLDSTEIN, 1979], should provide us with fresh insight in female, male, and child differences.

From GOLDSTEIN's still unpublised graphs of vocal tract outlines I have noted that the length of the pharynx measured from the glottis to the roof of the soft palate grows from 3.3 cm in the newborn child to 7.6 cm for the female aged 21 and 10 cm for the male aged 21. The length of the mouth measured from the back wall of the upper pharynx to the front teeth (alveolar ridge for the newborn infant) grows

from 5.5 cm for the newborn infant to 8 cm for the female of 21 and 8.5 cm for the male of 21. The tendency of relatively small variations of mouth cavity length with sex and age is more apparent than anticipated from earlier studies and would tend to minimize the range of 'mouth cavity formant frequencies'. The radical variations in relative pharynx length suggest that the relative role of front and back parts of the vocal tract could be reversed for a small child, i.e. that F_2 of the vowel [i] would be a front cavity formant, whilst F_3 is more dependent on the shorter back cavity. When front and back cavities are of more equal length, the dependency is divided and the F_3/F_2 ratio smaller than for males, which is typical of females or children of an intermediate age.

THE INVERSE TRANSFORM

As noted already in the introduction, there has been a substantial amount of theoretical work directed towards the derivation of area functions from speech wave data. In practice, however, these techniques are limited to nonnasal, nonobstructed vocal productions and the accuracy has not been great enough to warrant their use in speech research as a substitute for cineradiographic techniques. In the following section I shall attempt to comment on some of the main issues and problems. The usual technique, e.g. WAKITA [1973], is to start out with an LPC analysis of the speech wave to derive the reflection coefficients which describe the analog complex resonator. The success of this method is dependent on how well the losses in the vocal tract are taken into account. Till now the assumptions concerning losses have been either incomplete or unrealistic. Also the processing requires that the source function be eliminated in preprocessing by a suitable deemphasis or by limiting the analysis to the glottal closed period. In spite of these difficulties the area functions derived by WAKITA [1973; 1979] preserve gross features.

In general, a set of formant frequencies can be produced from an infinite number of different resonators of different length. We know of many compensatory transformations, such as a symmetrical perturbation of the single-tube resonator. However, if we measure the input impedance at the lips [SCHROEDER, 1976] or calculate formant bandwidths, we may avoid the ambiguities. A technique for handling tubes with side branches has been proposed by ISHIZAKI [1975].

According to WAKITA [1979], the linear prediction method is capable of deriving an area function quantized into successive sections of equal and predetermined length providing the LPC analysis secures an analysis equivalent to M formants specified in terms of frequency and bandwidth.

An estimation of the total length and of the area scale factor requires additional analysis data. An incorrect length estimate automatically generates compensatory changes in the area function which may be appreciable.

LPC analysis is a simple and powerful method of analysis but it fails in naturalness of representing the production process and as such is a poor substitute for a lossy transmission line representation. With the fresh eyes of a nonexpert on the inverse transform, I would attempt to make the following suggestions. One is that M formants with associated bandwidths could have a greater predictive power than noted by

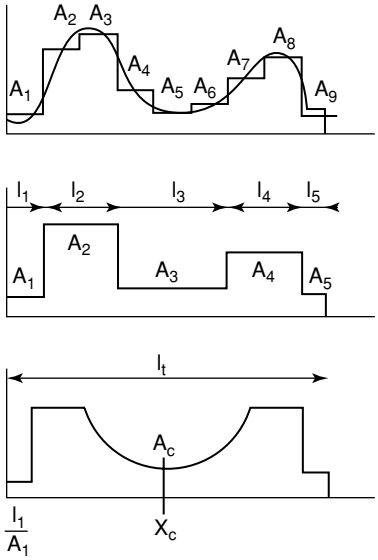


Figure 15. Continuous area function approximated by a constant larynx tube and 8 sections of equal length (top), by 4 sections of variable length and area (middle), and by a three-parameter model extended to include the total length (bottom). The constriction coordinate is zero at the lips.

WAKITA [1979]. The area scale factor could be included in addition to the 2M relative areas of his model. In general, with reservation for possible uniqueness problems, 2M formant parameters, including bandwidths but not necessarily as many bandwidths as frequencies, would suffice for predicting 2M independent area function parameters.

Thus, adding one more formant frequency to the M pairs of frequencies and bandwidths would suffice for estimating the total length of the 2M system. Alternatively, from the 2M formant measures we could derive a model quantized into M equivalent tubes, each specified by cross-sectional area and specific length, thus also predicting the total length (fig. 15). The rationale for this reasoning is that all losses in the transmission line analogs are unique functions of the area and length dimensions. One could also design a three-parameter model of the vocal tract as in figure 15 with a constant larynx tube. The four parameters (lip parameter A_1/l_1, x_c and A_c and the total length l) would hopefully be predictable from a specification of F_1, F_2, and F_3 and a bandwidth, say B_3, which appears to be more discriminating than B_1 and B_2. If we omit the total length and sacrifice the bandwidth, we have approached the articulatory modeling of LADEFOGED *et al.* [1978], which is based on correlational methods for deriving three articulatory parameters from F_1, F_2, and F_3.

In general, bandwidths have less predictive power than frequencies. They are to some extent predictable from formant frequencies [FANT 1972] (fig. 16). Furthermore, bandwidths vary with speaker, voice effort, and laryngeal articulations and are inherently diffcult to measure.

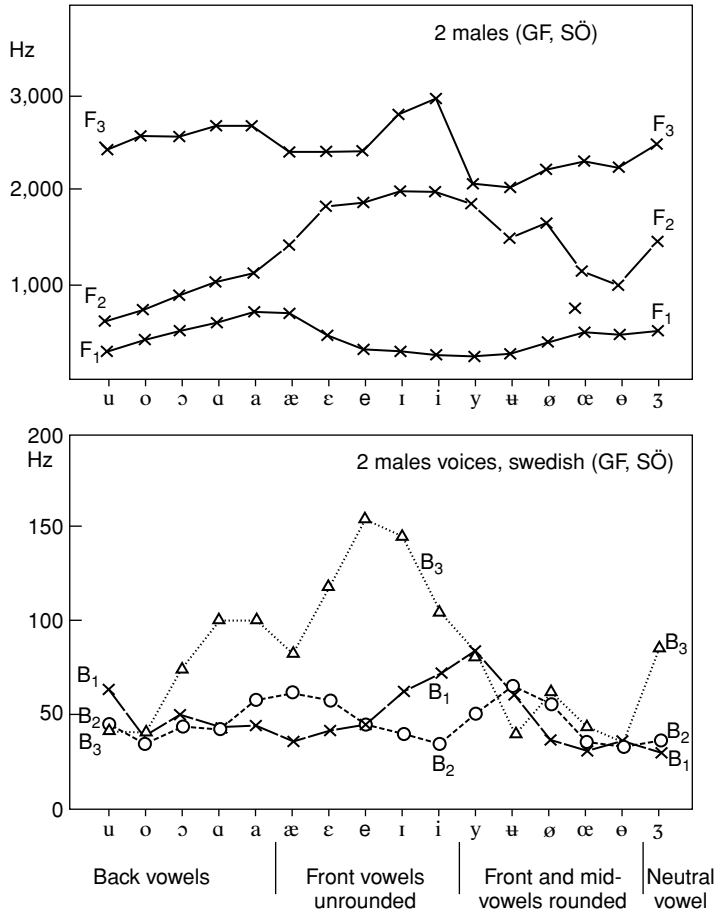


Figure 16. Frequency and band width patterns of Swedish vowels [FANT, 1972].

Still, I do not want to rule out the use of bandwidths. The following examples may serve to illustrate their predictive power and limitations. First, a test of the uniqueness in predicting 2M area function parameters from 2M formant data. Take the simple case of $M = 1$ which implies a single tube resonator. What are the length and cross-sectional area of a tube with a specified first resonance frequency and bandwidth? The length is immediately given by $F_1 = c/41$. As shown in figure 17, the area is a single-valued function of bandwidth providing only one loss element is postulated (as in LPC analysis). If we include both the internal surface losses of a hard-walled tube and the radiation resistance, the bandwidth versus area attains a minimum at 10 cm^2 and there are two alternative areas that fit the same bandwidth. The higher value could possibly be ruled out as being outside the possible range of human articulation. Similar ambiguities could also be expected in a more complex lossy transmission line model, as pointed out by ATAL *et al*. [1978]. However, one

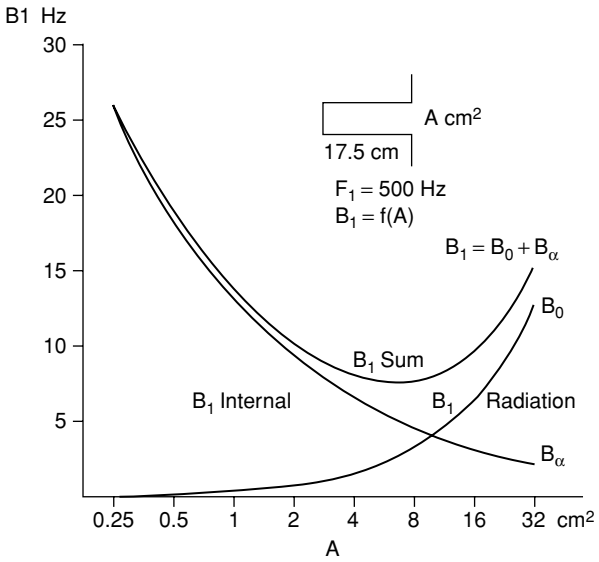


Figure 17. Bandwidth versus area of a single tube resonator taking into account internal losses and radiation load losses.

should note that their treatment of the invariance problem is not quite fair. They introduce more articulatory parameters than acoustic descriptors which obviously exaggerate the ambiguities. Next consider a two-tube approximation of the vocal tract (fig. 18A), with a back tube of length 8 cm and area 8 cm^2 and a front tube of length 6 cm and cross-sectional area 1 cm^2. The formant frequency pattern of $F_1 = 275$ Hz, $F_2 = 2132$ Hz, $F_3 = 2998$ Hz, $F_4 = 4412$ Hz and all higher formants is exactly the same as that of a two-tube system with the same areas but the lengths reversed, i.e. a front tube of length 8 cm and a back tube of length 6 cm (fig. 18B).

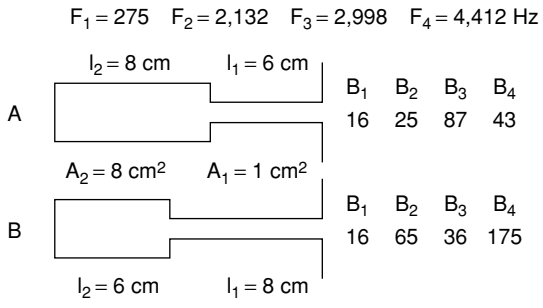


Figure 18. Two twin-tube resonators that provide the same F-pattern appropriate for the vowel [i], differing in terms of bandwidths.

This length ambiguity rule is apparent from the expression for resonance conditions

$$\frac{A_2}{A_1}\mathrm{tg}\frac{\omega l_1}{c} \times \mathrm{tg}\frac{\omega l_2}{c} = 1 \tag{5}$$

If bandwidths are calculated taking into account both the interior surface losses and the radiation resistance by formulas given by FANT [1960], we find that B_2 and B_4 of figure 18A are relatively low compared to B_3. In figure 18B, B_2 and B_4 are large compared to B_3. The different bandwidth patterns resolve the ambiguity. The physical explanation is that F_2 and F_4 of the first model are essentially determined by the back cavity and by the front cavity in the second model. The high damping associated with the surface losses in the narrow tube and the radiation resistance affect B_3 of (A) and B_2 and B_4 of (B).

The two models do not differ in terms of B_1. Theoretically it would be possible to choose the correct l_1, l_2, A_1, A_2 of the two-tube model from a specification of F_1, F_2, F_3 and either B_2 or B_3 or the ratio B_2/B_2 or B_4 or some combination of B_4 and other bandwidths, e.g. $(B_2 + B_4)/B_3$. In a real speech case the situation might be different if the glottal losses are large and execute high damping of the back tube resonances.

In practice it may take a ventriloquist to produce something similar to these two models. Possibly the one with a shorter back tube would fit into the vocal tract anatomy of a very small child, as suggested in the previous section.

In conclusion—to improve techniques for inferring vocal tract characteristics from speech wave data we need a better insight into vocal tract anatomy, area function constraints, and a continued experience of confronting models with reality—a balanced mixture of academic sophistications and pragmatic modeling.

REFERENCES

ATAL, B.S.; CHANG, J.J.; MATHEWS, M. V., and TUKEY, J. W.: Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J. acoust. Soc. Am. *63*: 1535–1555 (1978).

FANT, G.: Acoustic theory of speech production (Mouton, The Hague 1960; 2nd ed. 1970).

FANT, G.: Formants and cavities; in ZWIRNER, Proc. 5th Int. Congr. Phon. Sci., Munster 1964, pp. 120–141 (Karger, Basel 1965).

FANT, G.: A note on vocal tract size factors and non-uniform F-pattern scalings. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *4*: 22–30 (1966).

FANT, G.: Vocal tract wall effects, losses, and resonance bandwidths. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *2–3*: 28–52 (1972).

FANT, G.: Speech sounds and features (MIT Press, Cambridge 1973).

FANT, G.: Non-uniform vowel normalization. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *2–3*: 1–19 (1975a).

FANT, G.: Vocal-tract area and length perturbations. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *4*: 1–14 (1975b).

FANT, G.: Vocal tract energy functions and non-uniform scaling. J. acoust. Soc. Japan *11*: 1–18 (1976).

FANT, G.: Glottal source and excitation analysis. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *1*: 85–107 (1979).

FANT, G.: ISHIZAKA, K.; LINDQVIST, J., and SUNBERG, J.: Subglottal formants. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *1*: 1–12 (1972).
FANT, G.: and LILJENCRANTS, J.: Perception of vowels with truncated intraperiod decay envelopes. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *1*:79–84 (1979).
FANT, G.: NORD, L., and BRANDERUD, P.: A note on the vocal tract wall impedance. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *4*: 13–20 (1976).
FANT, G.: and PAULI, S.: Spatial characteristics of vocal tract resonance modes; in FANT, Proc. Speech Comm. Sem. 74. Speech communication, vol. 2, pp. 121–132 (Almqvist & Wiksell, Stockholm 1975).
FLANAGAN, J. L.: Speech analysis synthesis and perception (Springer, Berlin 1965; 2nd expanded ed. 1972).
FLANAGAN, J. L.; ISHIZAKA, K., and SHIPLEY, K.: Synthesis of speech from a dynamic model of the vocal cords and vocal tract. Bell Syst. tech. J. *54*: 485–506 (1975).
FUJIMURA, O. and LINDQVIST, J.: Sweep-tone measurements of vocal-tract characteristics. J. acoust. Soc. Am. *49*: 541–558 (1971).
GOLDSTEIN, U.: Modeling children's vocal tracts. J. acoust. Soc. Am. *65*: S25(A) (1979).
GUÉRIN, B.; MRAYATI, M., and CARRÉ, R.: A voice source taking into account of coupling with the supraglottal cavities. Rep. Lab. Communication Parlée, ENSERG, Grenoble (1975).
HEINZ , J. M.: Perturbation functions for the determination of vocal-tract area functions from vocal-tract eigenvalues. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *1*: 1–14 (1967).
ISHIZAKA, K.; FRENCH, J. C., and FLANAGAN, J. L.: Direct determination of vocal tract wall impedance. IEEE Trans. Acoust. Speech Signal Processing (ASSP) *23*: 370–373 (1975).
ISHIZAKI, S.: Analysis of speech based on stochastic process model. Bull. Electrotechn. Lab. *39*: 881–902 (1975).
JOSPA, P.: Effects de la dynamique du conduit vocal sur les modes de résonances. Rapp. Inst. Phonét., Universite Libre Bruxelles: 51–74 (1975).
LADEFOGED, P.; HARSHMAN, R.; GOLDSTEIN, L., and RICE, L.: Generating vocal tract shapes from formant frequencies. J. acoust Soc. Am. *64*: 1027–1035 (1978).
LINDBLOM, B. and SUNBERG, J.: A quantitative model of vowel production and the distinctive features of Swedish vowels. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *1*: 14–32 (1969).
LINDQVIST, J. and SUNBERG, J.: Acoustic properties of the nasal tract. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *1*: 13–17 (1972).
MRAYATI, M. et GUÉRIN, B.: Etude des caracteristiques acoustiques des voyelles orales franɔcaises par simulation du conduit vocal avec pertes. Revue Acoust. *36*: 18–32 (1976).
MRAYATI, M.; GUÉRIN, B. et BOË, L. J.: Etude de l'impédance du conduit vocal. Couplage source-conduit vocal. Acustica *35*: 330–340 (1976).
NORDSTRÖM, P.-E.: Attempts to simulate female and inFANT vocal tracts from male area functions. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *2–3*: 20–33 (1975).
ÖHMAN , S. E. G. and ZETTERLUND , S.: On symmetry in the vocal tract; in FANT, Proc. Speech Comm. Sem. 74. Speech communication, vol. 2, pp. 133–138 (Almqvist & Wiksell, Stockholm. 1975).
SCHROEDER, M. R.: Determination of the geometry of the human vocal tract by acoustic measurements. J. acoust. Soc. Am. *41*: 1002–1010 (1967).
SIDELL, R. S. and FREDBERG, J. J.: Noninvasive inference of airway network geometry from broad-band long reflection data. J. biomed. Engng *100*: 131–138 (1978).
SONDHI, M. M. and GOPINATH, B.: Determination of vocal tract shape from impulse response at the lips. J. acoust. Soc. Am. *49*: 1867–1873 (1971).
STEVENS, K. N.: Airflow and turbulence noise for fricative and stop consonants, static considerations. J. acoust. Soc. Am. *50*: 1180–1192 (1971).
STEVENS, K. N. and HOUSE, A. S.: Development of a quantitative description of vowel articulation. J. acoust. Soc. Am. *27*: 484–493 (1955).

WAKITA, H.: Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. IEEE Trans. Audio Electroacoust. (AU) *21*: 417–427 (1973).
WAKITA, H.: Estimation of vocal tract shapes from acoustical analysis of the speech wave: the state of the art. IEEE Trans. Acoust. Speech Signal Processing (ASSP) *27*: 281–285 (1979).
WAKITA, H. and FANT, G.: Toward a better vocal tract model. Speech Transm. Lab. Q. Prog. Status Rep. R. Inst. Technol., Stockh. *1*: 9–29 (1978).

Received: January 9, 1980; accepted: January 9, 1980.

GUNNAR FANT, Department of Speech Communication, Royal Institute of Technology, *S-Stockholm* (Sweden)

DISCUSSION

HISASHI WAKITA, RAYMOND DESCOUNT and PETER LADEFOGED opened the discussion.

HISASHI WAKITA: In determining the interrelationship between speech articulation and acoustics, we are particularly interested in the inverse problem, i.e. the estimate of vocal tract shapes from the acoustic waveform. There are various uncertain factors in deriving vocal tract area functions from the waveform, but it is an attractive method, because it is both the safest and easiest. (The problem with recent articulatory models for vocal tract shaping is that we do not yet know the exact parameters that control vocal tract shapes in terms of articulators, and we do not have sufficient methods to obtain the data.) One of the most promising methods is the linear prediction (LPC) method, to estimate area functions from acoustic data. We do not know to what extent we can describe the details of the vocal tract shape, but by combining the LPC method with physiological data, we hope to improve this method.

One problem is the nonuniqueness, i.e. we can generate an infinite number of shapes having exactly the same frequency spectrum within a limited frequency band. To solve the uniqueness problem we have to impose constraints, physiologically determined constraints, or constraints determined by the higher harmonic structure. So far, the LPC method has been using formant frequencies and bandwidths, and in fact the final area function is sometimes quite sensitive to bandwidth. But we would like to get rid of bandwidth in the calculations: From the first three formant frequencies we can obtain the midsagittal view of the vocal tract, as in PETER LADEFOGED's model, and to get at the unique shape of this midsagittal area function we may employ physiological constraints.

Another problem with LPC analysis is the vocal tract excitation and the losses, both within the vocal tract and at its boundaries, and these problems have to be solved in order to get more accurate vocal tract shapes. In fact, with the LPC method we can detect the closed glottis portion, where the interaction between sub- and supraglottal cavities is minimized, which makes for more accurate area functions. A further drawback of LPC is that we have to start from the very simple assumptions of a simple loss at the glottis and a lossless acoustic tube. On the other hand, you can make a production model as complex as you wish, you can add any realistic losses along the vocal tract or at the glottis that you like, but as an analysis model there is a strong limitation in incorporating losses and other factors. So at this

moment, the imminent problem is how to attack the loss problems and the source uncertainties.

RAYMOND DESCOUNT: Very little original data has accumulated on area functions, because collecting it is difficult, from a technical point of view. On the other hand, deriving vocal tract area functions from acoustic data has some disadvantages: with LPC techniques we only get pseudo-area functions, and with acoustic measurements, which I previously worked on, there is a great problem in dynamic measurements, especially. Further, interest has largely centered on the midsagittal view of the tract, but we need information about the frontal view as well, which may be obtained with the new techniques of computerized tomography. We need this information in order to turn the midsagittal view into a three-dimensional area function, and to determine the shape factors that are necessary for the introduction of losses in our models.

All the articulatory models proposed are based upon vowel configurations, and when we try to make dynamic simulations on the articulatory model, everything that we do not know about the consonants is put into a special coarticulation and transition rule. We need more information on the consonants.

The acoustic model of the vocal tract is derived from the propagation equations, based on assumptions of symmetrical, equal length sections,—but to do an inverse transform you really need a very appropriate model which includes the shape factors that are necessary for the loss calculations, because the mathematical technique involved in the transformation is stupid in the sense that the result will be adjusted according to mathematical criteria, but this may not result in a realistic vocal tract. Therefore, I think that doing inverse vocal tract transforms is premature: we must work first of all on the proposition of the best production model, including factors and losses, before trying to do inverse vocal tract transforms.

Due to the progress made in articulatory modelling and to the limitations of LPC techniques, we have witnessed a comeback of studies on vocal tract and vocal source simulations. To refine the articulatory model, we need further physiological data.

In conclusion: I do not think that LPC will give us a better understanding of speech production (it is, however, excellent for synthesis purpose). We need more studies on the relationship between articulatory parameters/area functions/vocal tract shapes.

PETER LADEFOGED: GUNNAR FANT showed us many years ago that what is important in characterizing speech are the first three formant frequencies, and you can even get a great deal of a speaker's personal quality with just three formant frequencies. But with the inverse transform, to get as far as eight tubes (which is only a coarse model of the vocal tract), you need at least four formant frequencies and their bandwidths, and with 18 tubes you need 9 formant frequencies and bandwidths, etc. Now something is wrong here: any phonetician can draw, more or less accurately, the midsagittal view of a given speaker's vowels, and we ought to be able to develop an algorithm that will go from the acoustics to the tract shape. There are, of course, problems—we do not actually observe the tract shape, only the midsagittal dimensions, and there are only very limited sets of data that tell us how to derive the tract shape from the sagital dimension.

The work of LINDBLOM and others has shown that you can produce an [i:] with your jaw in a more or less open position, i.e., one has the ability to control tract shapes

using different articulatory procedures, and it is of great interest to us to know how we exert that control and less interesting what the muscles do. Eventually, we have got to be able to go from acoustic structures, finding out what the tract shape is, and then deducing from that what the underlying signals must have been.

GUNNAR FANT: I agree with the main points of the discussants. Inverse transforms cannot make up for our great lack of physiological reference data. My suggestions for improving inverse transform techniques in part supported by the previous discussions are: (1) We should model the vocal tract in terms of lossy transmission line sections instead of the simplified LPC model. (2) We should not expect to generate a larger number of independent production parameters than we have independent and well-specified speech wave descriptors relating to the vocal tract transfer function. Overspecified area functions are necessarily nonunique, whereas a balanced specification can be, but need not be, unique. With proper model and parameter constraints, a 32-section area function model may be generated from a set of 3–6 articulatory parameters and controlled by the same number of acoustic parameters. It remains to be seen if we can extract more than four independent acoustic parameters. (3) The vocal tract total length should be derivable from one extra independent acoustic parameter.

Our discussion concerning bandwidths is still rather academic and we appear to share a doubt concerning the specificational value of bandwidths. Theoretically the set F_1 F_2 B_1 B_2 could suffice to specify a three-parameter model extended with a fourth parameter, e.g. the total length. This might hold for a resonator model only, but not for a true vocal tract with less predictable bandwidth sources and the limited accuracy in bandwidth measurements. A more efficient set of acoustic parameters would be F_1 F_2 F_3 and B_3. From my figure 16 illustrating bandwidths of Swedish vowels it is seen that B_3 is a good correlate of degree of lip opening and also mouth opening. However, vowel bandwidths including B_3 are to a high degree predictable from formant frequencies. The role of bandwidths in an LPC model is not the same as that of a true vocal tract model. This is an important distinction. The LPC bandwidths, e.g. B_3, may come out quite different from those of real speech or from simulations by an improved model. The bandwidths we need for the inverse LPC-based transforms are the bandwidths of a production model which has losses at the glottis only and lacks the cavity wall shunt. From the true formant frequencies and bandwidths we thus have to make a best guess of what bandwidths the LPC model would generate. This is in the line of the recent work of HISASHI WAKITA [1979].

KENNETH STEVENS: With regard to what a male speaker does in order to compensate relative to the [u:] of a female: if we define narrow vowels as having so narrow a constriction that turbulence is just not generated, is it conceivable then that males, who generate a greater airflow than women, cannot round the vowels as much as can women, and therefore the formants are not lower than those of women?

GUNNAR FANT: It could be, but in Swedish the vowel [u:] as well as [i:], [y:], and [ʉ:] are generally produced, by males and females alike, with a diphthongal glide passing through a relatively constricted phase in which some turbulence may be generated. I would rather expect different male and female articulations to be aimed at some criterion of perceptual invariance of which we do not know too much yet.

ANTTI SOVIJÄRVI asked GUNNAR FANT what his concept is about nasalized vowels.

GUNNAR FANT: An essential characteristic of nasalization independent of the specific resonances added is the reduced F_1 amplitude which is especially apparent in an oscillographic analysis. What appears to be a sub-F_1 nasal formant is often a voice source feature which is relatively reinforced because of the F_1 reduction.

HISASHI WAKITA: As long as the calculations are based on the first few formant frequencies, the problems in inverse transformation are rather equivalent with different methods. To uniquely determine a six-tube vocal tract shape, LPC uses the first three bandwidths. If you want a smooth area function, you have to specify one of the higher frequency characteristics, and to do that you have to impose some kind of constraint, which is what Dr LADEFOGED does. And whatever the method, if you do not want to use bandwidth you have to use some other kind of information to uniquely determine the spectra, and any information will do as long as you are able to reconstruct the original spectrum with its original bandwidths—so bandwidth is in fact a very important parameter.

GUNNAR FANT: It would be interesting to see how far you would get if you started out with F_1, F_2 and F_3, and then predicted B_1, B_2 and B_3 from the formulas that I have.

PETER LADEFOGED: I have tried using HISASHI WAKITA's formulae with GUNNAR FANT's type of predicted bandwidths (and other bandwidths from the literature), and it did not work,—I got absolutely impossible vocal tract shapes. Regarding ATAL's vocal tract shapes that produce identical formant frequencies: some of them are quite impossible, the tongue just cannot produce some of these shapes.

JOHN HOLMES: I wish to emphasize the difficulty of mathematically deriving the vocal tract from the speech waveform, because we know too little about the glottal source. GUNNAR FANT emphasized that the closed glottis portion is better suited than the open glottis portion to work out the supraglottal characteristics, but (as can be seen on the Farnsworth vocal chord movie of about 1940 and from TOM BAER's work), even when the vocal chords are closed, there is sufficient ripple and surface movement for there to be an effective volume velocity input into the vocal tract, which means that your resultant waveform is never a force-free response,—and this is one of the things that makes bandwidths so difficult to estimate, because it is quite possible that ripple in vocal chord surface could actually be causing the formant amplitude to be still building up even, in exceptional cases, during the closed glottis period. I think this supports the view that we have to work from much more basic information and use articulatory constraints rather than to derive vocal tracts by purely mathematical techniques from some artificial and unrealistic production model.

GUNNAR FANT: I can only agree with your statements. It is necessary to learn more about the human voice source in order to improve our method of inverse transforms.

OSAMU FUJIMURA: We can obtain cross-sectional vocal tract shapes with the regular computerized tomography, but only at great costs, because the X-ray dosage

is tremendously high, a requirement of brain diagnoses that demand a very good density solution. But I think the machine can be adjusted and the X-ray dosage reduced for our purposes, where we are really only interested in the distinction between matter and air.

MOHAN SONDHI at the Bell Laboratories has proposed an acoustic impedance measurement using an impulse-like excitation at the lips, which can give us complete information about the area function of the vocal tract, because we obtain two sets of infinite series, i.e. the poles and the zeroes of the impedance function that together uniquely determine the vocal tract shape, without having to assume or measure losses. I think that there is one major difficulty with this technique: the subject articulates silently, i.e. he has no auditory feedback, and we cannot be sure about the actual gestures. That problem can be overcome if we simultaneously monitor the vocal tract with e.g. the X-ray microbeam method.

GUNNAR FANT: The microbeam system will certainly provide us with excellent data about speech articulation, but will it provide us with all the details that we want about the vocal tract, like the exact dimensions of the pharynx and larynx cavities?

OSAMU FUJIMURA: We can obtain data on cross-sectional shapes, because we can place pellets also outside the midsagittal plane, the only constraint being that we cannot use too many pellets at the same time, which will increase the X-ray dosage, but it is not easy to place pellets on the pharyngeal walls, which is a limitation of the method. However, we have a new stereofiberscope which can be used for three-dimensional optical observations of the pharynx, and I hope in the future to be able to develop a technique that will supplement the X-ray technique with this kind of optical information.

RAYMOND DESCOUT: I am presently working with a prototype CT (computerized tomography) scanner, which scans in 5 sec, and we are trying to lower the X-ray dosage to 10% of the normal dosage, because all we need is to see the difference between air and flesh. There is still a problem with the CT technique, though, and that is determing exactly the position of the slice relative to the skin and the rest of the person.

CHAPTER 2.3

SWEDISH VOWELS AND A NEW THREE-PARAMETER MODEL

ABSTRACT

Vocal tract area functions of 13 Swedish vowels have been derived from midsagittal tracings of X-ray pictures, supported by a limited tomographic material. With a few exceptions, e.g. F2 of [u:], [o:] and [ɔ], calculated formant frequencies show a substantial agreement with data measured during the X-ray session. The sensitivity of formant frequencies to variations in vocal tract area functions have been studied.

Observed co-variation of overall vocal tract dimensions revealed a number of dependent relationships that enable a prediction of overall length, inter-incisor distance, and asymmetries of cavity shapes from the basic specification of location and area of tongue constriction and the degree of lip-rounding. The new model thus preserves physiological constraints that make it better suited for a future adaptation to consonantal modifications than earlier three-parameter models. Nomograms of formant frequencies for systematically varied model parameters are shown. Problems related to inverse mapping, from formant frequencies to model parameters, are discussed.

1. INTRODUCTION

In connection with an X-ray tomographic study of the vowels [u:], [ɑ:] and [i:] (Fant, 1964) processed by Sundberg (1969) a set of lateral views were obtained for the same male subject sustaining the vowels [u:] [o:] [ɔ] [ɑ:] [a:] [æ:] [ɛ:] [e:] [i:] [y:] [ʉ:] [ø:] [œ:] [ɵ:].

These have been used on various occasions to illustrate the Swedish vowel system, see [8], but have not been processed earlier. Now, with the growing interest in articulatory synthesis, (Lin,1990) this material should constitute a valuable source for articulatory modelling, adding to the rather meagre available data on vowel specific VT area functions which for a long time has been dominated by the Russian vowels, (Fant, 1960).

2. THE SWEDISH VOWEL SYSTEM

Swedish has a quite rich vowel system (Fant, 1973, 1983). There are three back vowels, /O/, /Å/ /A/, three front vowels /I/, /E/, /Ä/ and three rounded front vowels, /Y/, /U/, /Ö/. These occur in pairs of long and short vowels, thus in all 18 phonemes. Within a pair there usually exists a quality difference, which might be small or absent as in the pre-[r] allophones [ae:] and [ae] of the phoneme /Ä/ and in the pre-[r] allophones [oe:] and [oe] of the phoneme /Ö/. Of the 13 vowels selected for our study, see Figure 1, three are phonemically short, [ɔ], [a], [ɵ]. They have F-patterns significantly different from those of the corresponding long phonemes, and may accordingly be sustained. The individual vowels in Figure 1 have been oriented approximately according to their position in an F1 versus F2 plot.

Observe the front vowel character of [u:] articulated with a lip opening more narrow than in [y:] and with an extra apical elevation. Whilst the [ʉ:] historically has

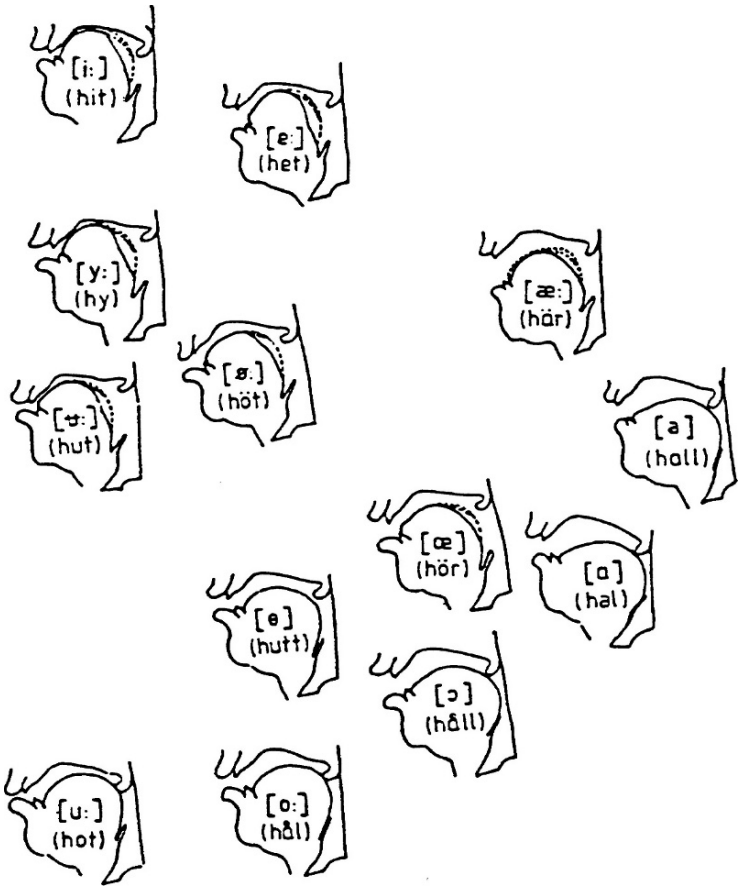


Figure 1. Lateral view of sustained Swedish vowels. From Fant (1964,1983).

advanced to an extreme high front vowel, its phonemical mate, the short vowel [ɵ] is quite close to the back vowel [ɔ]. In standard Swedish, the distance between the vowels [ʉ:] and [ɵ] is greater than within any other pair of a long and a short vowel.

The overall relations between vowels as seen in Figure 1 are typical. One exception is the fairly large lip opening of [y:]. One has also reason to suspect that the articulatory targets of [u:] [o:] [ɔ] attained during the X-ray exposures were somewhat relaxed compared to the subject's reference conditions. This pertains both to lip opening and back tongue constrictions, and could explain the small difference between [ɵ] and [ɔ].

2.1. *F-Pattern Calculations*

The subject was a man of age 30, an amateur singer with barytone voice, and as judged from the collected data, of average head size. For the processing of dimensions we followed the system of Fant (1960). The basic coordinate system in the sagittal

plane was thus not the usual one with horisontal layers slicing the pharynx and a joining system of radial lines for the mouth. Such a system may produce projection errors with respect to wavefronts. Instead, an outline connecting estimated center points of wavefront slices was constructed. As a zero coordinate along this center line we choose its intersection with a line through the upper and lower incisors.

The total length of the vocal tract, from the assumed plane of radiation at the lips to the glottis, was typically 19.5 cm for rounded vowels and 17.5 cm for completely unrounded vowels. A part of this difference, about 1 cm, was associated with a larynx lowering in rounded vowels, a well known phenomenon, see e.g. (Wood, 1986).

The conversion from distance d(x) in the sagittal plane to cross-sectional area A(x) was performed on the basis of power function expressions, in part derived from tomographic data for the subject's [u:] [ɑ:] and [i:], in part from earlier studies, (Sundberg, 1969; Sundberg et al. 1987; Lin, 1990).

$$A(x) = a\, d(x)^b \tag{1}$$

For the lip section in the range of $d < 1.7$ cm, we choose $a = 1.8$, and $b = 2.5$, and for $d > 1.7$ cm, $a = 5$ and $b = 0.6$.

For the mouth cavity we adopted an initial estimate of $a = 2.4$ and $b = 1.4$, which was derived from Fant (1960) and was found to be very close to the (Sundberg, Johansson, Wildebrand & Ytterberg, 1987) data. However, when tested against the tomographic data for the subject's [u:] and [ɑ:] we found that it was necessary to add a correction for air columns on both sides of the tongue. These are quite prominent, see pictures in Fant (1964). For the vowel [u:] they add about 35% to the mouth cavity volume. A similar correction was applied to all back vowels.

When applying the power function to pharynx measures it was found to be necessary to divide the range into three regions; one of $d < 1.75$ cm with $a = 2$ and $b = 1.6$, an intermediate region of $1.75 < d < 2.5$ cm with $a = 2.8$ and $b = 1$, and a limiting higher region of $d > 2.5$ cm with $a = 3.7$ and $b = 0.7$. Pharynx cross-sectional areas derived from these expressions are somewhat smaller than those reported in Sundberg (1969) but larger than those in Sundberg et al. (1987).

For the larynx tube we adopted area functions derived from the tomographic data of [u:] [ɑ:] and [i:], adjusted to the particular side views. The overall length varied from 2.2 to 2.5 cm and the input area was close 1.7 cm^2. A standard sinus piriformis cavities was inserted, see section 4.

Calculations were performed from detailed equivalent network representations of area functions (Lin, 1990). All known losses were incorporated. The wall impedance was introduced by an R, L shunt at a level 4 cm above the vocal folds, (Fant, Nord Branderud (1976). This simple representation appears to provide a more realistic overall function than a distributed impedance.

The sound recordings from the X-ray session were not complete in all details, and because of noise interference we had to rely on averages of phonations during the subject's rehearsal prior to exposure.

Formant frequencies were measured from broad-band spectrograms. Measured and calculated data are shown in Figure 2. Except for the back vowels [u:] [o:] [ɔ] the overall fit in $F_1$and F_2 is quite good with an error of predicted minus

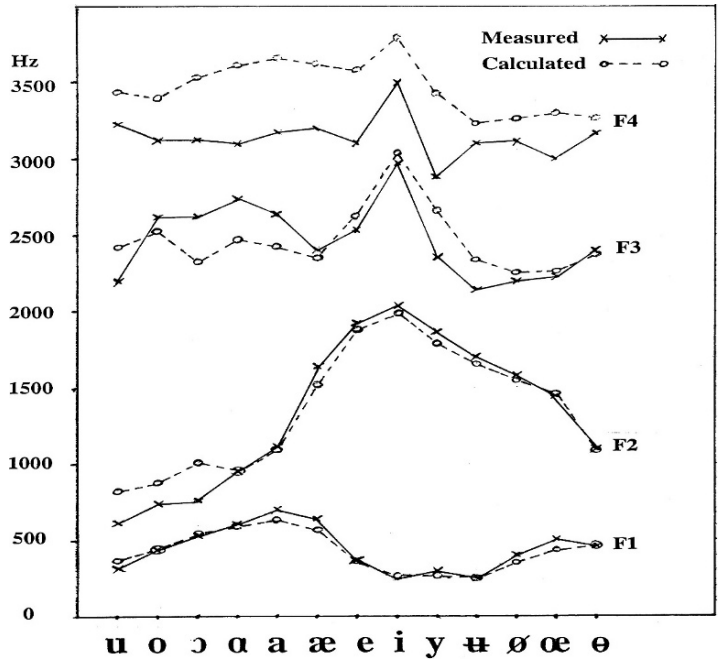


Figure 2. Measured and calculated formant frequencies

measured formant frequencies averaging −22 Hz (SD = 27 Hz) in F_1 and -42 Hz (SD = 42 Hz) in F_2.

Except for [u:] where F_3 is too high, calculated F_3 values of back vowels tend to come out too low. F_3 of the rounded front vowels [y:] and [ʉ:] are somewhat too high, reflecting insufficient lip rounding during the X-ray session. Calculated F_4 values were about 200 Hz too high, which is a general finding.

As judged from a perturbation analysis the errors in F_2 of the three back vowels, about 200 Hz, reflect a combination of insufficient lip rounding and insufficient narrowing at the tongue constriction of an order of a factor 1.5 to 2.0. These large differences can not likely be explained from systematic errors in the power functions for area determinations. The most likely explanation lies in a relaxed articulation during the exposure. A corresponding, but even greater difference in F_2 was observed in the MRI study of Baer, Gore, Gracco & Nye (1991).

Systematic analysis of factors related to assumptions concerning vocal tract configurations have been made. Thus the effect of neglecting the air columns on the sides of the tongue in the vowel [u:] is to increase F_2 by about 100 Hz and to decrease F_3 by about 200 Hz.

The sinus piriformis cavities cause a lowering of all formant frequencies by a small amount, usually less than 70 Hz, whilst the wall impedance has an opposite effect, essentially confined to F_1. The central grove in front vowels needs to be taken into account. If neglected, cross-section areas tend to be overestimated and

calculated F_2 values come out about 80 Hz too low. An increase of the length of the larynx tube by 2.5 mm or introducing a corresponding radiation inductance load at its termination will mainly effect F_4 by a lowering of the order of 100 Hz.

3. A NEW THREE-PARAMETER MODEL

Earlier three-parameter models of VT area functions, employing a symmetrical tongue hump constriction between front and back cavities of constant cross-sectional areas are rather stereotype. The constricted region is either given a constant area (Fant, 1960) or a parabolic (Stevens & House, 1955) or a catenoidal (Fant, 1960) or a cosine shape (Lin, 1990). The need for an asymmetry of the constrictied region was pointed out by Lin (1990).

Our ambition has been to include as much as possible of physiological realism retaining the three basic parameters, X_c and A_c specifying the location and minimum area of the tongue constriction, and l_o/A_o specifying the length over area ratio of the lip opening. A detailed study of the areafunctions of the thirteen vowels revealed a number of dependent relations to other descriptors such as jaw opening, ratios of front to back cavity maximum areas, overall length and constriction asymmetry.

As shown in Figure 3, the modelling employs somewhat different conditions for three major regions of the X_c scale, a "front" region of X_c located less than 4 cm from the teeth, a "mid" region at coordinates between $X_c = 4$cm and $X_c = 7$cm, and a "back" region at X_c greater than 7 cm. In the latter region we find all back vowels The vowel [u:] occupies a position at the border between the mid and back regions. The vowel [ɵ] is close to [o:] but has greater A_c. As expected, the vowel [æ] fitted best into the back category with the largest X_c, i.e. a constriction in the bottom of the pharynx.

The areafunctions have been divided into a number of segments, six for front vowels, eight for mid vowels and seven for back vowels. A_o extending from $X = -1$ to $X = 0$ is the lip area. Instead of the default value $l_o = 1$ cm one may choose an l_o covarying with A_o, retaining the desired l_o/A_o. The area between the front teeth, A_t, is derived from the sagital distance dt through a power function with $a = 2.4$ and $b = 1.4$ in Eq 1. A_m, at $X = X_m$, is the maximum mouth cavity area in mid and back vowels. In front vowels there is merely a gradual transition between A_t and A_c. The coordinate X_b is defined by $A(x) = 4.5\text{cm}^2$, which is the value approached for the entire areafunction when $A_c = 4.5\text{cm}^2$, i.e the neutral state. $A(x) = 4.5\text{cm}^2$ also defines the coordinate X_f, located posterior to X_m. In mid vowels the point (X_f, A_f) on the area function is located in the decent from maximum mouth area A_m to constriction minimum area A_c. The (X_p, A_p) point ensures a suitable shape of the pharynx cavity.

In front vowels, the teeth opening, dt was found to be correlated with increasing A_c and A_o. In back vowels we found a positive correlation of dt with X_c, i.e. the jaw opens as the tongue constriction moves towards the bottom of the pharynx. The linear regression equation

$$dt = 0.18X_c - 0.4 \quad (R = 0.94) \tag{2}$$

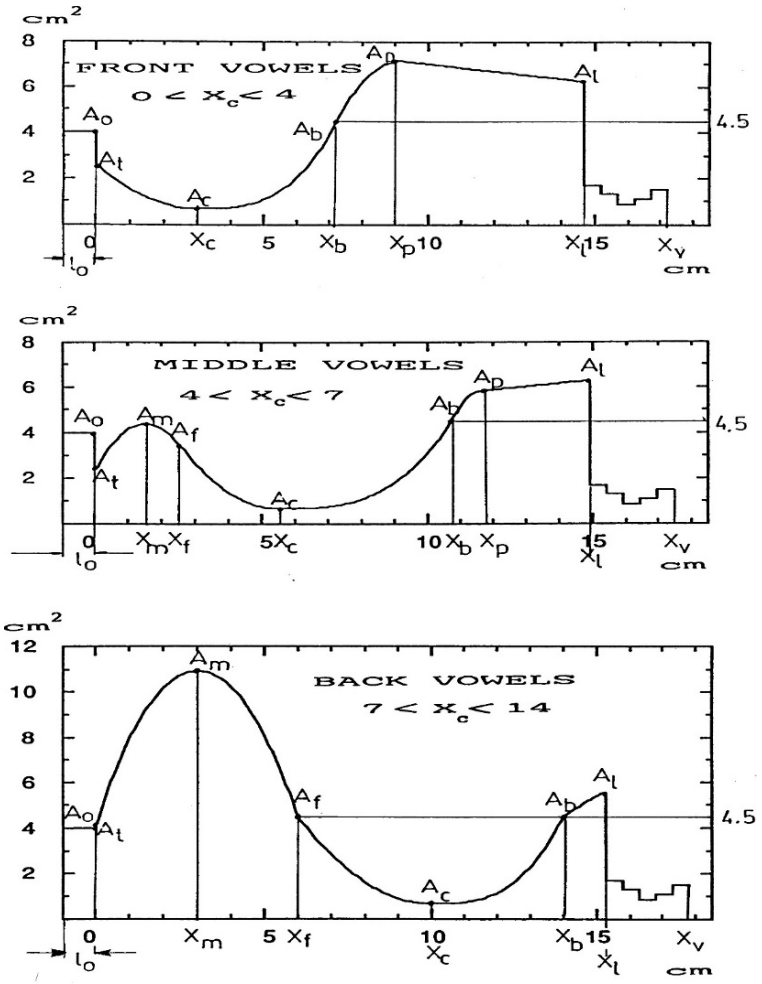


Figure 3. The new three-parameter model.

fits well except for [æ:] which was produced with 0.5 cm greater dt than predicted.

The coordinates X_f and X_b of back vowels show accurate linear relations to X_c, from which we may derive an asymmetry index

$$S_a = (X_c - X_f)/(X_b - X_c) \tag{3}$$

which covers a substantial range, $S_a = 0.37$ at $X_c = 7$ and 2.4 at $X_c = 13$.

The interior length of the vocal tract from the teeth to the entrance to the larynx tube, X_l, is kept constant 15,5 cm in back vowels and is varied somewhat with l_o/A_o in front vowels and also with X_c in midvowels. The larynx tube is represented by five 0.5 cm sections of standard values. At its outlet in the pharynx it is connected to a sinus piriformis cavity, also of 2.5 cm length, and with an area linearly varied from

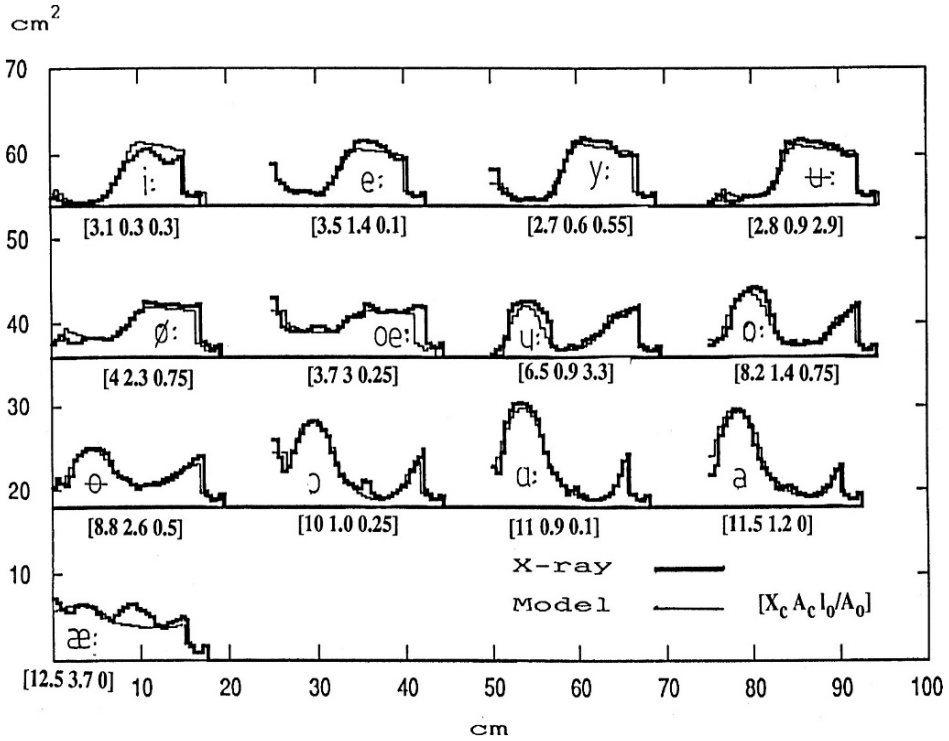


Figure 4. Area functions from the X-ray study and from the model.

3 to 0 cm^2. The individual segments within an areafunction are in most instances modelled by parabolic functions. An exception is the segment between X_c and X_b, where a third order power function was applied, and the intervall between X_p and X_l which was represented by a straight line.

The region of midvowels has on the whole been designed to provide a suitable transition between the very different front and back vowel regions. Some guidance was attained from the areafunction of a medio-palatal consonant [g], and from the [u:] located close to $X_c = 7$. A problem has been to ensure maximally continuity at the model boundaries $X_c = 4$ and $X_c = 7$, not only in absolute values but also with respect to derivatives. Some final adjustments remain to be made.

The result of a first order visual matching of model generated areafunctions to those derived from the X-ray data are shown in Figure 4. The match is on the whole good, but we observe a tendency of a minor underestimation of total length. In [æ:] we find an underestimate of the area around $X_c = 10$. The general appearance of the [æ:] area function with two internal minima is similar to what has been described by Boe, Perrier & Bailly (1992). In terms of X_c, the ordering of the front vowels is [y:] [ʉ:] [i:] [e:] [œ:] [ɵ] occupying a region of X_c from 2.7 to 4.0. No vowel was found well within the mid range. An exception was the [u:] at $X_c = 6.5$. The remaining back vowels form the sequence [o:] [ɵ] [ɔ] [ɑ:] [a] and [æ:] which is in essential agreement with Wood (1979).

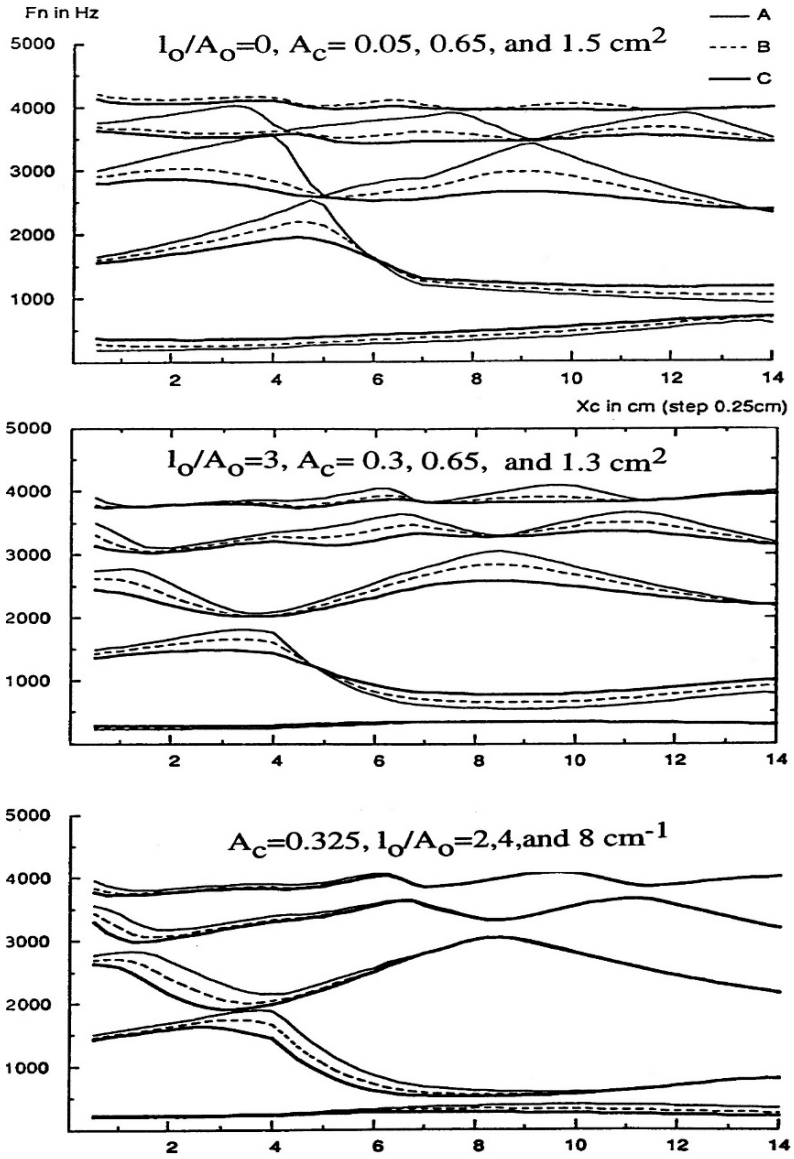


Figure 5. Nomograms of F_1, F_2, F_3, F_4, F_5 for varying X_c coordinates generated from the new three-parameter model.

Formant frequencies calculated from the model areafunctions agreed on the whole with those from the X-ray data. In 50% of the cases the model generated formant data matched the subject's phonation better than the X-ray derived data.

Figure 5 shows nomograms of F_1, F_2, F_3, F_4, and F_5 as function of X_c coordinates. The top figure pertains to $l_o/A_o = 0$, i.e. no lip rounding and $A_c = 0.05$, 0.65 and 1.5 cm^2. Here we note the maximally high $F_3 = 3500$ Hz at $X_c = 4$ and $A_c = 0.05$

typical of a [j] target. The middle diagram pertains to $l_o/A_o = 3$ and $A_c = 0.3, 0.65$ and 1.3 cm^2. The F_2-F_3 proximity point has now advanced to the left (Fant, 1960; Wood. 1986).At $X_c = 6.5$ the F-pattern is appropriate for an [u]. Here, the main effect of a decrease of A_c is to lower F_2. This illustrates the origin of the F_2 error of [u†] discussed in connection with Figure 2. There is apparently a large range of X_c locations that would provide a satisfactory [u:], as already pointed out by Baer et al. (1991). In the lower diagram A_c was set to a constant value of 0.325 cm^2 and $l_o/A_o = 2$, 4, and 8. Here we may note the influence of lip rounding on F_2 of [u:]. The F_2-F_3 proximity range is enlarged which make [u:] and [y:] insensitive to X_c. These are typical examples of stable regions as proposed by Stevens (1989).

We have applied the algorithms of Lin (1990) for inverse transformation, deriving X_c, A_c, and l_o/A_o from F_1, F_2 and F_3 of the subjects phonation. The automatic search was in most instances successful, but difficulties were encountered in achieving a correct F_3 of [u:], [ɔ] and [æ] while maintaining correct F_1 and F_2. The situation will probably improve by a systematic release of the constraints that tie overall length and mouth opening to the three basic parameters. The model has been extended to consonant articulations (Fant & Båvegård, 1997).

4. ACKNOWLEDGEMENTS

This work has in part been supported by grants from the Swedish Board for Technical Development, STU and STUF, and from a grant from Carl Tryggers Stiftelse. A considerable amount of calculation work was contributed by Qiguang Lin.

REFERENCES

Baer, T., Gore, J. C., Gracco, L. C. &. Nye, P. W. (1991) Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *J. Acous., Soc. Am.* **90** (2) 799–828.

Boe, L.-J., Perrier, P. & Bailly, G. (1992). The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, **20**, 27–38.

Fant, G. (1964). Formants and cavities, Proc.Vth Int Congress of Phonetic Sciences. Karger, Basel, 120–140.

Fant, G. (1960). *Acoustic Theory of Speech Production.* Mouton, The Hague.

Fant, G. Nord, L. & Branderud, P. (1976). A note on the vocal tract wall impedance. *STL-QPSR* 4-1976, 13–20.

Fant, G. (1973). *Speech sounds and features*. The MIT Press. Cambridge, Mass.

Fant, G. (1983) Feature analysis of Swedish vowels—a revisit. *STL-QPSR* 2–3 1983, 1–19.

Fant G. & Båvegård, M. (1997). Parametric model of the vocal tract area function: Vowels and consonants, *ESPRIT/BR SPEECHMAPS (6975). Delivery 28, WP2.2*, 1–30 (1995). Also published in *TMH-QPSR* 1/1997, 1–20.

Lin, Q. (1990). *Speech production theory and articulatory speech synthesis*. D.Sc. thesis. Royal Inst of Technology, KTH, Stockholm.

Sundberg, J. (1969). On the problem of obtaining area functions from lateral X-ray pictures of the vocal tract, *STL-QPSR* 1/1969. Royal Inst. of Tech. Stockholm, 43–45.

Sundberg, J. Johansson, J. C. Wilbrand, H. & Ytterberg, C. (1987). From sagittal distance to area. A study of transverse, vocal tract cross-sectional area, *Phonetica*, **44**, 76–90.

Stevens, K. N. & House, A. S. (1955). Development of a quantitative description of vowel articulation. *J. Acoust. Soc. Am.,* **27**, 484–493.
Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, **17,** 9–34.
Wood, S. (1986) The acoustical significance of tongue, lip, and larynx maneuvers in rounded palatal vowels. *J. Acoust. Soc. Am.* ,**8**0(2), 391–401.
Wood, S. (1979) A radiographic analysis of constriction locations for vowels*, Journal of Phonetics*, **7**, 25–4.

CHAPTER 2.4

1. INSTRUMENTATION FOR PARAMETRIC SYNTHESIS (OVE II)

Our formant coded speech synthesizer OVE II is organized as shown in the block diagram of Fig. II-1. There are three main synthesis filters which are operated independently labelled the F N and K systems. Each of these is designed as a cascade (series) connection of elementary zero and pole functions. These systems are fed from a voice source and a noise source through source amplitude modulating gates A_O A_N A_H A_C. The gate A_O determines the amplitude of voicing for the vocalic filter system and A_N is the corresponding amplitude of voice for the nasal system N. The notation A_H pertains to the noise amplitude for the vocalic filter system F and A_C is the amplitude of noise for the fricative filter K. The outputs of the F- N- and K-systems are added in a mixer stage.

A mechanical function generator provides means for voltage control of 12 parameters during 3 seconds of speech. Normally we utilize 11 of these, namely the voice fundamental frequency F_O, the four gating functions mentioned above and further 3 variable F-formants, 2 variable K-formants, and one variable K-anti-resonance. These control voltages are smoothed through low-pass filters, the time constants of which may be varied in octave steps from 2.5 msec to 40 msec.

Fig.:s II-2, II-3, II-4, and II-5 provide detail information on the design of the main building blocks, including the formant generators, the anti-resonance circuit, and the amplitude modulated gates.

The F-filter is intended for the synthesis of vowellike sounds and comprises the voltage controlled units F1 F2 F3 and the fixed position, manually controllable units F4 and F5′ and KH.

The voltage-frequency calibration of F1 F2 and F3 is linear which is a requirement for their sharing a common field on the function generator. With normal condenser settings in the resonance circuits the ranges are as follows:

F1	140 c/s	$\longrightarrow$	1000 c/s
F2	500 c/s	$\longrightarrow$	3000 c/s
F3	1100 c/s	$\longrightarrow$	4500 c/s

The alternative settings for F4 and for the higher pole correction circuits F5′ and KH (see ref. (1)) are as follows:

F4	2500 c/s	3000 c/s	3250 c/s	3500 c/s	4000 c/s
F5′	3.0 kc/s	3.5 kc/s	4.0 kc/s	4.5 kc/s	5.0 kc/s
KH	3.0 kc/s	3.5 kc/s	4.0 kc/s	4.5 kc/s	5.0 kc/s

Equivalent vocal tract length.

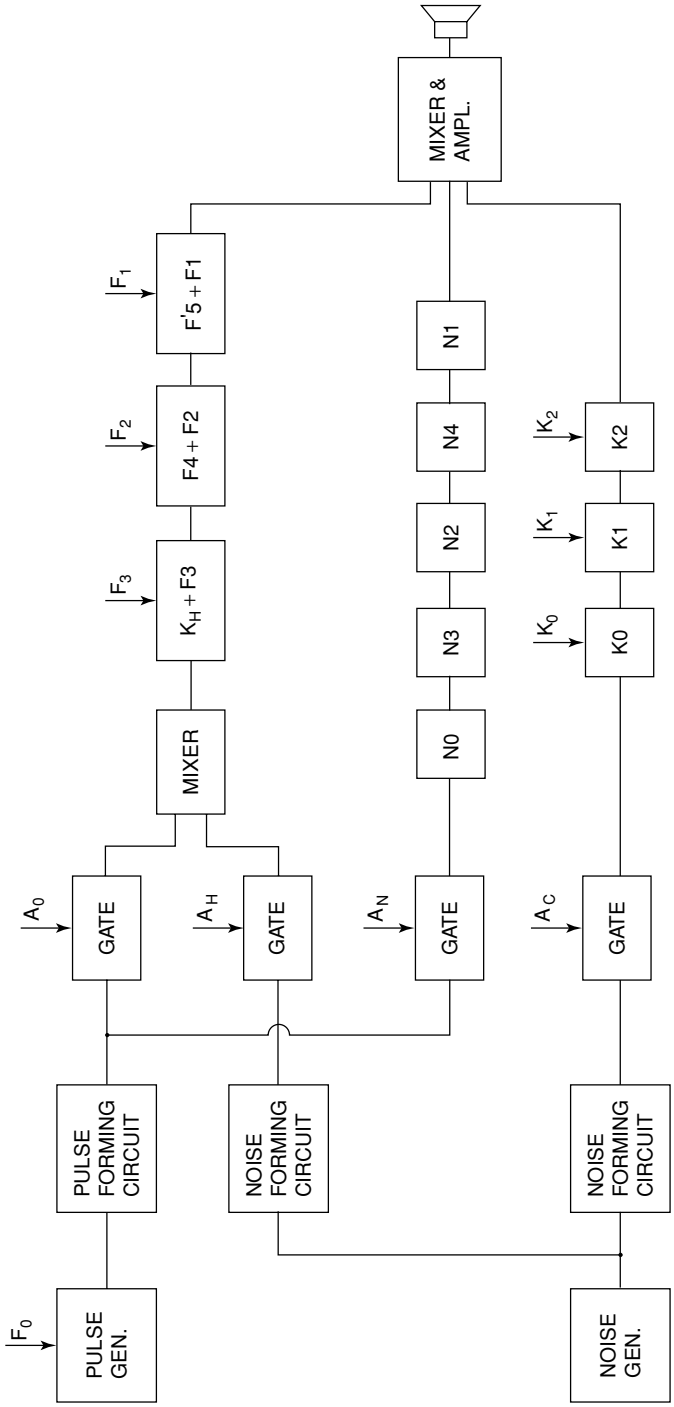


Figure 1. Block diagram of OVE II.

The bandwidth may be tuned in by hand for each formant. Unless otherwise specified we select the values:

$B_1 = 70$ c/s $\quad B_2 = 80$ c/s $\quad B_3 = 100$ c/s
$B_4 = 140$ c/s $\quad B_5 = 400$ c/s

The second filter unit N, is intended for nasal sounds and is not controlled from the function generator except by means of the nasal gate which connects or disconnects the nasal filter to the common mixer. This unit has 4 pole circuits:

N1, N2, N3, N4, and a zero circuit NO

Frequencies and bandwidths are manually controllable over wide ranges. Typical settings for simulating a nasal consonant are:

[m]	$N_1 = 225$ c/s	$N_2 = 700$ c/s	$N_3 = 1200$ c/s	$N_4 = 2200$ c/s
	$N_5 = 2700$ c/s	NO = 800 c/s		
[n]	$N_1 = 225$ c/s	$N_2 = 1100$ c/s	$N_3 = 2100$ c/s	$N_4 = 2500$ c/s
	$N_5 = 2700$ c/s	NO = 1500 c/s		

The third filter unit, K, is intended for synthesis of non-vowellike noise sounds, i.e. unvoiced fricatives, affricates, and short fricative sound segments within the burst of a stop sound. This filter comprises 2 poles and one zero which are voltage controlled parameters. These 2 resonances and the anti-resonance cover the following frequency ranges:

K_1	1000–6500 c/s
K_2	2500–10000 c/s
K_0	1000–6000 c/s

The bandwidths are preset before synthesis to values of the order of 200–400 c/s.

2. SYNTHESIS STRATEGY

The particular combination of source and filter functions are selected in order to make possible the following types of sound segments (compare the section on type features in Chapter IV of ref. (2)).

	Voice source		Noise source		Special requirement
	F Vocalic A_O	N Nasal A_N	F Vocalic A_H	K Fricative A_C	
Phonetic category					
Voiced vowel diphtong, glide [1] and [r] and noise-free variants of [v] [w] [j]	+				
Voiced occlusive	+				F_1 very low A_O-amplitude low
Whispered vowel			+		
Nasal consonant		+			
Nasalized vowel (Always next to nasal consonant)	+	+			
Fricative consonant including affricates and short initial transient and fricative phase of stop bursts			(+)	+	
Same as above with voicing	+			+	
[h]-like sounds including unvoiced onset of vowel after an unvoiced consonant			+		A_H-source spectrum shaped with low F1-level, low level above 3000 c/s, and −6 dB/oct average slope

The F_1 F_2 and F_3 data are traced directly from spectrograms produced with an expanded time scale (twice the normal Sonagram speed.) and expanded frequency scale. The F_O-curve may also be traced in this way from a narrow-band spectrogram following one of the harmonics. Alternatively, we produce an F_O-curve with an automatic pitch extractor recording it as a Mingogram with the appropriate time scale for transfer to the coding sheet of the function generator.

An A_O-curve is also transferred from a Mingogram. As a crude approximation to the inverse filtering we make use of a simple integration for deriving the source amplitude from the speech wave, this gives us $A_O + A_N$, a separation is made with the aid of a spectrogram. After the first trial of making a synthesis and comparing it with the original it is possible to make fine adjustments of these gating functions as well as all other functions.

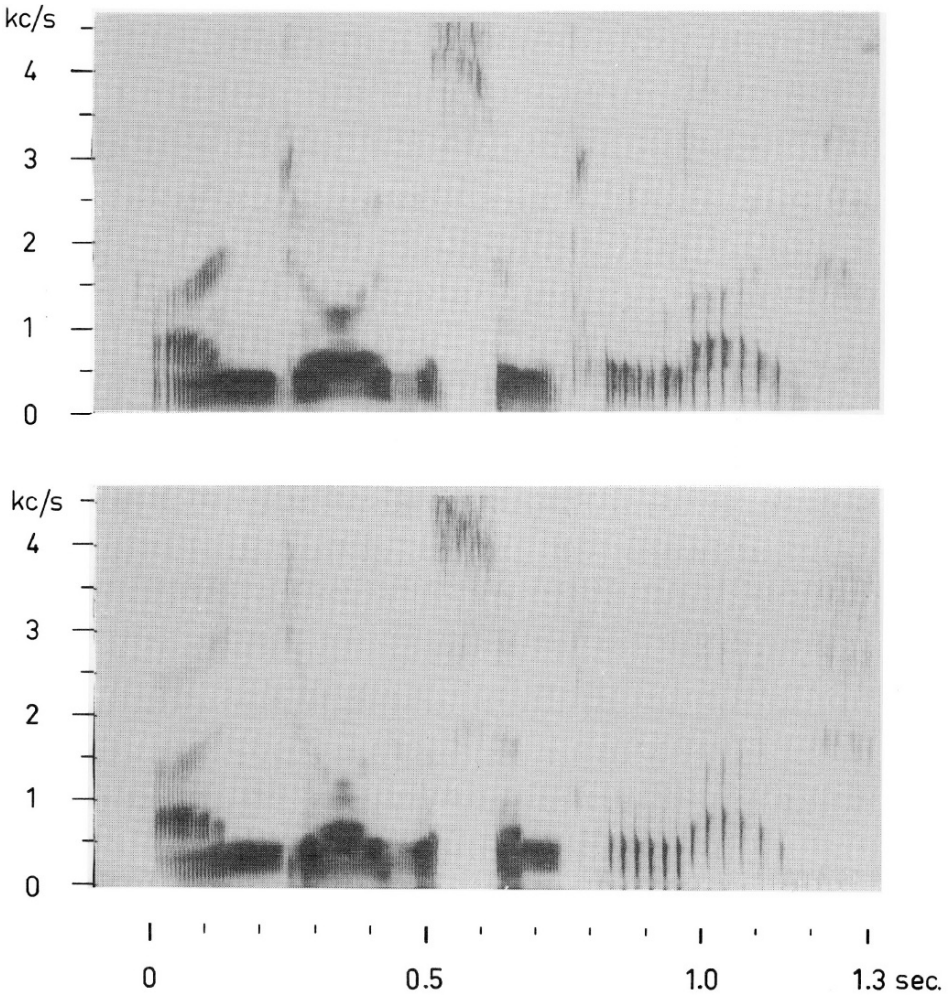


Figure 6. Spectrograms of the sentence "I enjoy the simple life" as produced by a human subject (above) and OVE II synthesis (below).

Fricatives of the type s and sh can generally be synthesized with a fairly good quality merely by placing K_1 K_2 and K_0 according to the visible evidence from the spectrogram. For the intense syllables such as [s] and [ʃ] and [ç] the formant K_1 is placed on the main peak, K_2 on the next higher peak of importance, and K_0 about an octave below K_1. In case of labiodental [f] K_1 and K_0 are placed rather close in the 1000–2000 c/s range whereas K_2 is placed at the upper extreme of the frequency scale in the vicinity of 8000 c/s and is well damped. The F-system activated through gate A_H is frequently added if judged necessary by the appearance of F2 F3 and F4 in addition to higher K-type formants. This would be common practice for an aspirated [t]-burst. The A_H and A_C sources are shaped from white noise. The A_C source is flat

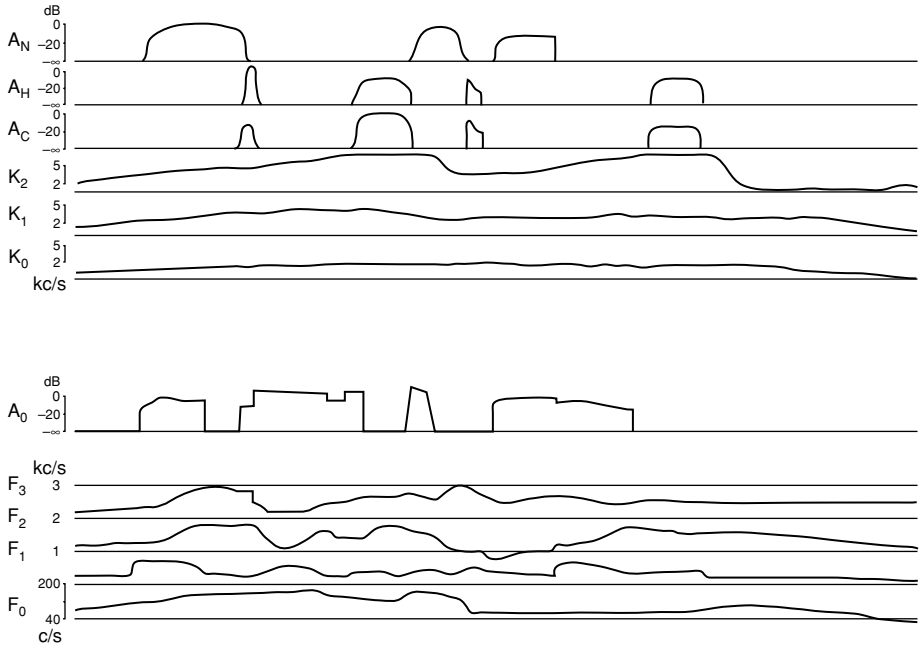


Figure 7. The time-variation of the 11 synthesis parameters within the sentence "I enjoy the simple life".

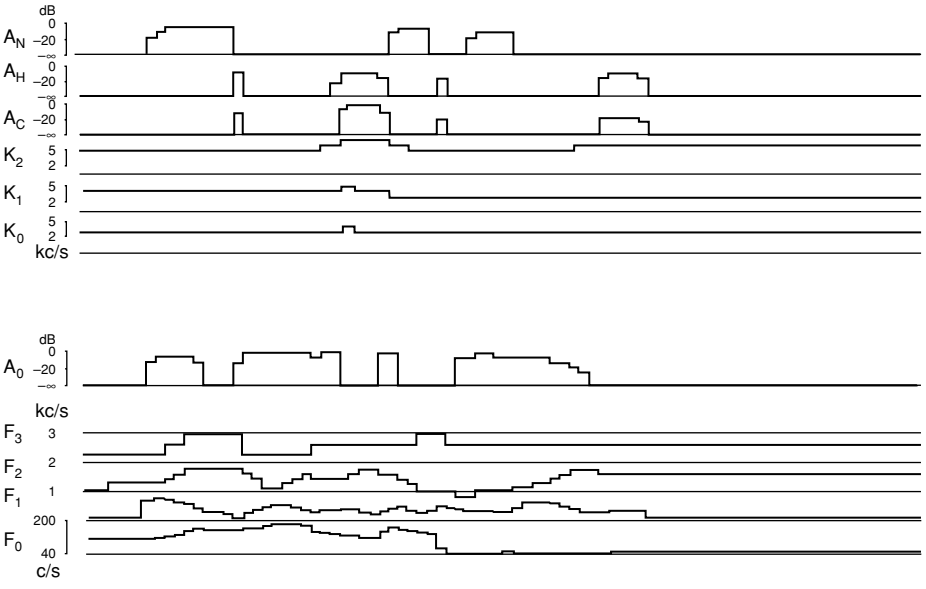


Figure 8. The time-variation of the 11 synthesis parameters within the sentence "I enjoy the simple life". The parameters have been sampled at a rate of 40 times per second and quantized.

and A_H is integrated (−6 dB/oct) and both have some degree of high-pass filtering to suppress the spectrum level below 500 c/s. The A_H-gate has in addition a low-pass cutoff at 3000 c/s. Without the latter precaution an [h]-sound might be confused with a [sh]-sound.

The A_0-source has an average slope of −12 dB/oct. The effects of radiation transfer are added by a differentiation (+6 dB/oct) in the A_0 shaping circuitry. Variations of the voice source are made possible by an additional conjugate complex zero and a pole and a few zeros and poles on the negative axis (simple RC circuitry).

Only in very special cases we have bothered to attempt a more detailed source match. In the sentence "I enjoy the simple life", see Fig. II-6, synthesized by J. Holmes ref. (3), the primary voice source was takne from an especially shaped triangular wave selected to match the voice source pulse shape as viewed from an inverse filtering set up.

The quality of the synthetic speech can be made to approach the human original very closely providing a careful match has been undertaken and the particular speaker has voice characteristics which are favorable for reporduction with OVE II. In general, basing the synthesis on a standard source, we may loose typical aspects of the speaker's voice timbre. However, apparent speaker characteristics are still retained in the faithful reproduction of the F-pattern (F_1 F_2 and F_3) and F_O.

The general impression of the OVE II speech is that it lacks the "harshness" quality typical of most channel vocoders, but that it also fails to reproduce the elements of "crispness" found in many human voices.

3. QUANTIZATION OF SYNTHESIS PARAMETERS*

A quantization scheme planned to suit the demands of a synthesis process in a formant vocoder was carried out by drawing staircase curves instead of continuously varying curves. Fig. II-7 illustrates this procedure which was carried through for two sentences ("I enjoy the simple life" and "He knows just what he wants"). The parameters were sampled at a rate of 40 times per second and were quantized as follows:

Parameters	Number of bits	Number of levels	Comments
F_O	4	16	First voiced sample following voicelessness was coded in 1/6 octave steps covering the range 60–340 c/s. Next and following samples within voiced portions of speech were coded in 1/24 octave steps of change vs the pitch of the previous sample. Total possible range of F_O 46.5–428 c/s.
F_1	4	16	The range of F_1 was 150–900 c/s covered in 50 c/s quantal steps.
F_2	4	16	The range of F_2 was 550–2800 c/s covered in 150 c/s quantal steps.

F_3	3	8	The range of F_3 was 1550–4000 c/s covered in steps of 350 c/s.
K_1	2	4	The range of K_1 was 3000–6000 c/s covered in quantal steps of 1000 c/s.
K_2–K_1	1	2	Two alternative values of K_2–K_1 = 2000 and 3000 c/s.
K_0/K_1	1	2	Two alternative values of $K_0/K_1 = 1/2$ and $1\sqrt{2}$ were adopted.
A_0	3	8	$A_0 = +5$, 0, -5, -10, -15, -20, -25, and $-\infty$ dB.
A_C	3	8	Same as A_0 above.
A_H	2	4	-5, -10, -20, and $-\infty$ dB.
A_N	2	4	Same as A_H above.
	Total	29 bits/sample	

Since it is theoretically possible to let the F_1 information occupy the same signal channel as the K_0 K_1 K_2 information (mutually exclusive variations) it would be possible to subtract 4 units from the number of bits 29 and conceive of the coding as requiring a channel of the capacity of transmitting 25 bits/sample and thus 1000 bits/second.

A special test was run on the effect of varying the outoff frequency and thus the time constant of the smoothing filters in each control signal channel connecting the function generator output of a channel and the corresponding input control terminal of the synthesizer. It was found that a time constant of 10 milliseconds was sufficient to smooth out the discontinuities of the control signals to the extent that the audible effects were eliminated. A time constant of 20 milliseconds did not noticeably affect the quality of the analog or quantized speech. These low-pass smoothing experiments were made with our standard 3rd order minimum overshoot low-pass filters of the transform:

$$H(s) = \frac{1.27w_I^3}{(s + 0.85w_I)[(s + 0.7w_I)^2 + w_I^2]} \tag{1}$$

In producing the analog speech the smoothing filters were set at a time constant of 10 milliseconds corresponding to a low-pass cutoff frequency of 40 c/s.

Special tests on varying the quantal steps in F_O showed that the 4 bits approximation with mixed absolute and differential code did not provide an improvement compared with a 3 bits approximation, the difference not being very great. However, these tests were made on a very limited speech material and should thus be regarded as merely indicative of practical coding demands.

G. Fant, J. Mártony

NOTE

* Section 3 contains old material, first reported in STL-QPSR 1/1961. It is included with the purpose of making chapter II a complete summary of present synthesis techniques.

REFERENCES

[1] FANT, G.: "Acoustic analysis and synthesis of speech with applications to Swodish", Ericsson Technics *15*, No. 1 (1959) pp. 3–108.
[2] FANT, G.: "Speech analysis and synthesis", Royal Institute of Technology, Div. of Telegraphy-Telephony, Speech Transmission Laboratory, Report No. 26 (June 1962).
[3] Holmes, J.N.: "Notes on synthesis work", *STL-QPSR* 1/1961, pp. 10–12.

CHAPTER 2.5

A NEW ALGORITHM FOR SPEECH SYNTHESIS BASED ON VOCAL TRACT MODELING*

ABSTRACT

A new algorithm for articulatory speech synthesis is described in this paper. The algorithm constitutes two main parts: a detailed frequency domain modeling of the vocal tract and a data transform from the frequency domain to the time domain. A computer model was developed to simulate the vocal tract acoustics. The model incorporates all known important components of the vocal system and computes the transfer function between the lip/nostril output and the acoustic source. By decomposing the obtained transfer function into its numerator and denominator, frequencies and bandwidths of resonances (and of anti-resonances if any) can be determined. The transfer function can next be written as a partial fraction expansion series in terms of calculated residues at the poles and can be approximated by retaining the first few terms in the series. Usually, each of these terms is a second-order module and corresponds to an elementary formant resonance. Formants are thus connected in parallel. The time-domain output is obtained by the inverse Laplace transform. Compared with other synthesis methods, this vocal tract oriented synthesis strategy has a number of advantages. For instance, the frequency dependency of loss elements of the vocal tract is preserved and accurate frequency responses can be reproduced. It is also computationally efficient relative to a direct convolution method. Examples of spectrum matching are presented to discuss the properties of the proposed algorithm. The algorithm has been incorporated in an articulatory-based speech synthesis system currently under development at KTH.

1. ACOUSTIC MODELING OF THE SPEECH TRANSMISSION SYSTEM

The speech transmission system consists of the tracheal tubes, larynx, pharyngeal cavities, vocal chambers, and nasal passages. Conventionally, the system of varying cross-sectional dimensions can be approximated by consecutive cylinders, and a planar wave propagation inside the system is assumed. Under this assumption, the transmission properties of each cylinder can be represented in analog to homogeneous transmission line, usually in terms of a T-network. The vocal/nasal tracts are terminated by a radiation impedance at their front end. See Fig. 1. This figure also shows a few additional shunt arms. They have been incorporated to simulate nasal sinuses to provide a good picture of the acoustic characteristics of the nasal system and to simulate the effects of the yielding wall.

A computer model for dealing with the vocal system shown in Fig. 1 has been developed (Lin, 1990), based on the design of Badin & Fant (1984). The system is simulated in the frequency domain, taking advantage of the more convenient and accurate modeling of losses and radiation. Some major features of the model can be summarized as follows [refer to Lin (1990) for more details]:

a) The numerator and the denominator of the calculated transfer function are decomposed so that data of poles (and zeros if any) can be correctly determined. For an ideal all-pole model, the numerator is a constant. Otherwise it becomes a function of frequency, for instance, when there are extra elements in the system. For a shunting element, the zero of the transmission occurs at frequency where the shunt has a zero, and for a serial

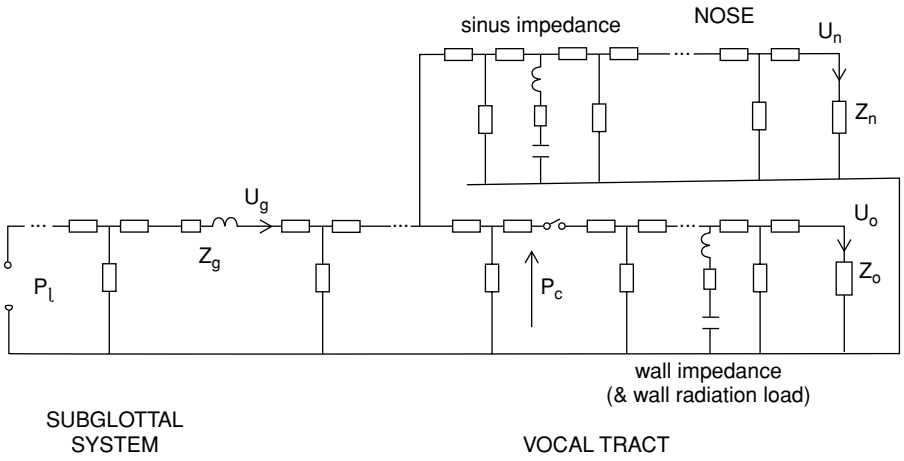


Figure 1. Network representation of the human speech-producing system.

element, the zero of the transmission occurs at frequency where the element has a pole. The location of the transmission zeros will be changed if one sums up the simultaneous outputs from different ports.

b) An on-line graphic display facility is provided, see Fig. 2. After each running, it is able to alter some of the simulation conditions and an updated transfer function is computed. This is a convenient tool to study the relations between articulation and acoustics. For example, it is clearly seen in Fig. 2 that a small perturbation of area function in the vicinity of the velum affects only the location of higher formants.

c) Different models for the radiation impedance and for the wall impedance have been implemented. For a simultaneous radiation occurring at various ports (the lips, the nostrils, and the vibrating walls) the volume velocities are superposed linearly, disregarding the spatial phase difference.

d) The residue data at the poles of the transmission are determined. They will be used later to expand the transfer function into a partial fraction series (see further Section 2).

Based the X-ray tracing data from Fant (1960), the model has been used to simulate various sounds, such as vowels, fricatives, liquids, nasal murmurs and nasalized sounds. Representative results have been achieved. An example is shown in Figs. 3, 4, and 5, where the transition from /m/ to /a/ is simulated. Fig. 3 shows the underlying area functions, while Figs. 4 and 5 are plots of the spectra and the zero/pole distribution, respectively.

The particular spectral envelope around 500 Hz for Curve A in Fig. 4 is the result of a zero-pole-zero combination. These two zeros are located closely, so a small frequency scanning increment should be used so as not to miss any of them. In Fig. 5, the positions of the first and third zeros are rather stable from A to B to C. These zeros are related to the zeros of the nasal sinuses. The other two zeros are caused by the entire branched system. Their trajectories vary as the configuration of the system changes.

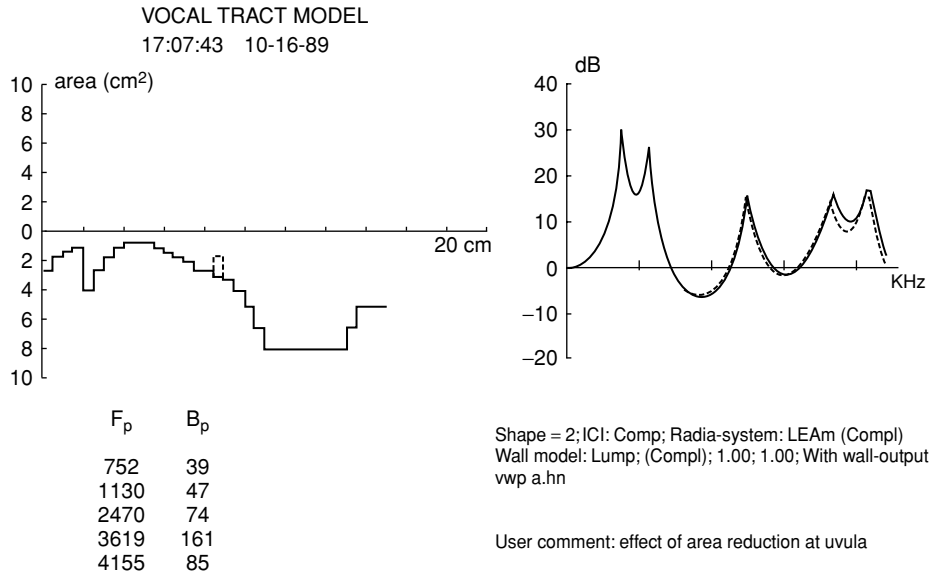


Figure 2. An example of on-line graphic display when running program TRACT. Note that F_4 and F_5 are shifted upwards when the area in the vicinity of the velum is enlarged.

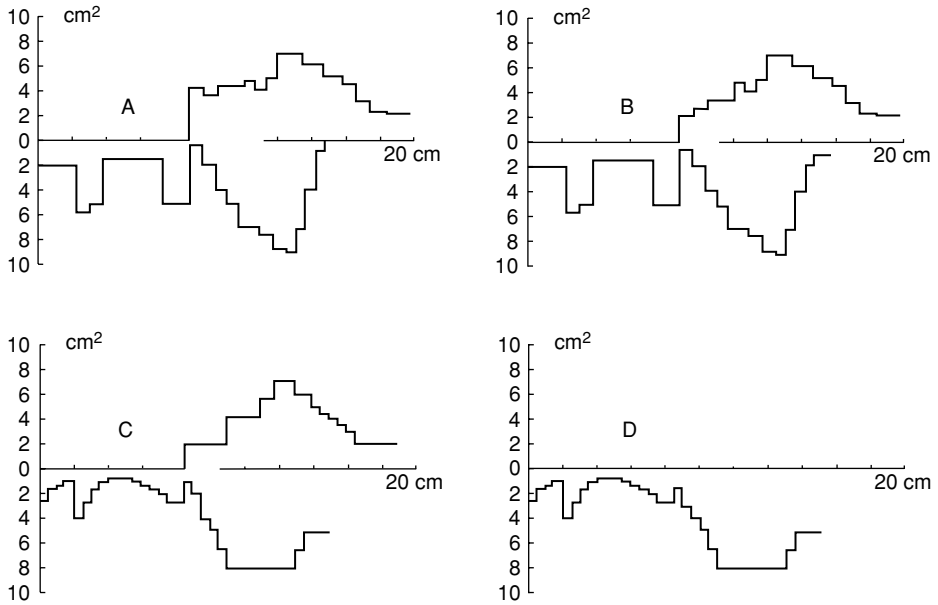


Figure 3. Plausible area functions of the syllable [ma], sampled at four instants in time. The abscissa, 2 cm per division, has its origin at the glottal end. A: An ideal nasal consonant [m]; B and C: intermediate phases of transition; and D: an ideal oral vowel [ɑ:]

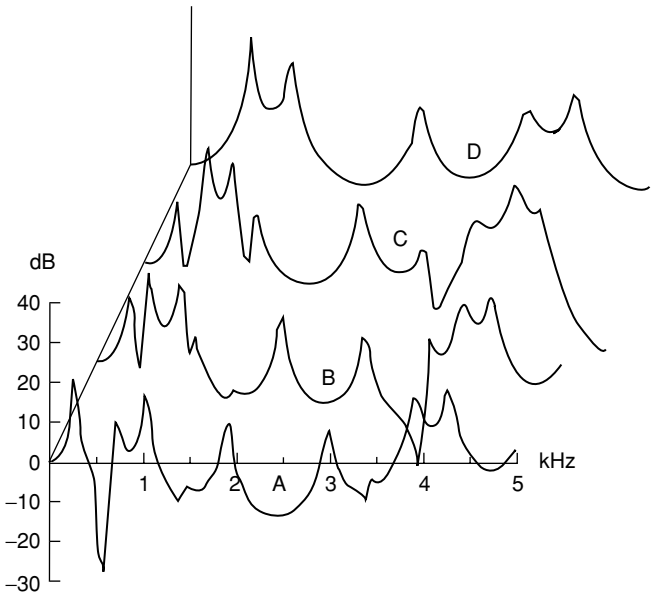


Figure 4. Transfer functions of the syllable [ma], refer to Fig. 3.

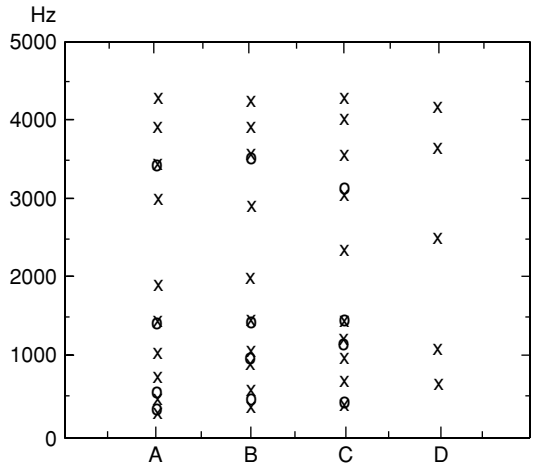


Figure 5. Plot of zeros (o) and poles (x), based on Fig. 4.

When the velum is raised and thereby discouples the nasal tract from the system, all of the zeros will then be cancelled by their bound poles, see D in Fig. 5. It can also be seen that the coupled nasal- and vocal tracts have more poles than the vocal tract alone does. When excluding the pole-zero pairs, there are 6 poles for A, B, and C below 5000 Hz and 5 poles for D. This result is expected since the effective length is different.

2. THE NEW ALGORITHM FOR SYNTHESIS IN THE TIME DOMAIN

A new algorithm to interface the above frequency-domain analysis with the time-domain synthesis has been developed. Once the data of poles and the associated residues are known, the transfer function can be written as a partial fraction expansion series. Theoretically, there are infinitely many terms in the series. However, within a limited frequency band of interest, the transfer function can be approximated by explicitly retaining the first few terms. If the order of the numerator of the transfer function is lower than that of the denominator and if there are no real poles, then each of the terms corresponds to an elementary formant resonance. Formants are thus connected in parallel. By performing inverse Laplace-transform, the time-domain output is obtained.

Mathematically, the calculation of residues at poles can, in the S-plane, be interpreted as vector products as depicted in Fig. 6. The desired residue equals the product of vectors from all zeros to the pole under investigation divided by the product of vectors from all other poles to that pole. If there is no zero, the dividend is of course equal to 1.

However, the transfer function is calculated along the jω-axis. All relevant vectors are pointing at $j\omega_n$ instead at $s_n = \sigma_n + j\omega_n$, which brings about error in the residue calculation. The error is marginal only if the neighbouring poles/zeros are located quite apart (relative to the distnace between s_n and $j\omega_n$) and if the real part of the current pole is relatively small. Otherwise a proper correction is needed to move back the shifted vectors. It is found that a correction from the nearest pole/zero is usually sufficient. Note that if no such corrections are necessary, it is then unnecessary to determine the zeros, Lin (1990).

Let $A_n = \alpha_n + j\beta_n$ be the (complex) residue at the nth pole, we then have:

$$H(s) = \sum_{n=1}^{N}[H_n(s)] = \sum_{n=1}^{N} \frac{2 \cdot \alpha_n \cdot (s - \sigma_n) - 2 \cdot \beta_n \cdot \omega_n}{s^2 - 2 \cdot \sigma_n \cdot s + {\omega_{on}}^2}, \tag{1}$$

with

$$H_n(s) = \frac{2 \cdot \alpha_n \cdot (s - \sigma_n) - 2 \cdot \beta_n \cdot \omega_n}{s^2 - 2 \cdot \sigma_n \cdot s + {\omega_{on}}^2}, \tag{2}$$

where ${\omega_{on}}^2 = {\sigma_n}^2 + {\omega_n}^2$, and N is the number of the formants retained.

The correction factor is given by:

$$C_{i,n} = \frac{s_n - s_i}{j\omega_n - s_i} = \frac{(\sigma_n - \sigma_i) + j(\omega_n - \omega_i)}{-\sigma_i + j(\omega_n - \omega_i)}, \tag{3}$$

where $s_i = \sigma_i + j\omega_i$ may be a neighbouring zero or pole to the current pole. The denominator of Eq. (3) is constructed to eliminate the shifted vector, and the numerator specifies the substituting vector, or the original one.

The correction factor $C_{i,n}$ can also be used to estimate the transfer function for a lossy system based on the residue data calculated from the corresponding lossless system (Lin, 1990). This can save computation time, provided the lost information on the bandwidths can be estimated by some empirical formulations.

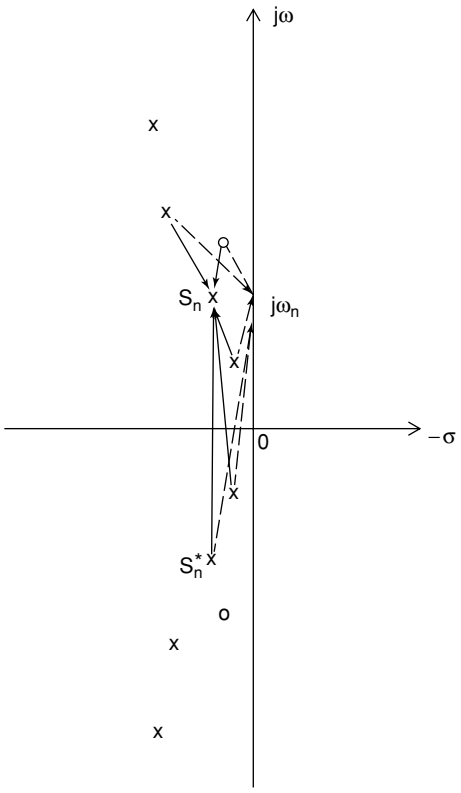


Figure 6. Vector representation of the residue calculation. Vectors in solid lines are the actual ones, and vectors in dashed lines are those obtainable from a direct vocal tract computation.

It is found, at least for the transfer function of vowels, that the complex residue $A_n = \alpha_n + j\beta_n$ can be approximated by its dominating part $j\beta_n$. When α_n is ignored, Eq. (2) reduces to:

$$H_n(s) = -\frac{2 \cdot \beta_n \cdot \omega_n}{s^2 - 2 \cdot \sigma_n \cdot s + {\omega_{on}}^2}. \tag{4}$$

or:

$$H_n(s) = (-1)^n \cdot \frac{-2 \cdot |H(\omega_n)| \cdot \sigma_n \cdot \omega_n}{s^2 - 2 \cdot \sigma_n \cdot s + {\omega_{on}}^2}, \tag{5}$$

where $|H(\omega_n)|$ denotes the spectrum level at ω_n (the nth formant level), which is intimately related to the associated residue. The numerator of $H_n(s)$ is now independent of the frequency. Such independence was actually the basic assumption of the Holmes parallel synthesis system, see Holmes (1983) for the detail of his design. Eq. (5) indicates that the simplified parallel system is specified exclusively by the formant frequency, bandwidth, and its amplitude. They can be estimated from a

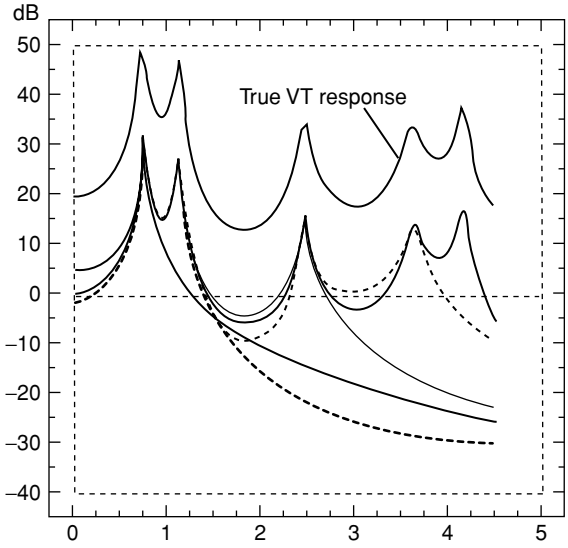


Figure 7. Magnitude spectrum of vowel [a]. Accumulative output of formants, in comparison with the true vocal tract response.

section display of a narrow-band spectrogram, by means of an analysis-by-synthesis routine. Note that the polarity alternates with pole number n in Eq. (5). When there enter zeros in the transfer function, the polarity alternates with pole plus zero number.

Figs. 7 and 8 give two examples of spectrum matching result, one for a vocalic sound and the other for a nasalized vowel. It is shown in Fig. 8 that a better fit is obtained when the correction is applied. More examples can be found in Lin (1990).

When formant circuits are connected in cascade, the spectrum of a vowel-like sound is defined solely by the formant frequencies and bandwidths, in addition to the term of higher pole correction, HPC (Fant, 1960). In terms of residue calculation, the cascade connection may be changed to an equivalent parallel connection. Fig. 9 presents such an example. Thus, by this technique it is able, though not necessary, to avoid the cascade configuration in a formant speech synthesizer. The control strategy remains, however, the same as if a cascade structure of formant was used.

3. COMPARISON WITH OTHER TYPES OF SYNTHESIS SYSTEM

In the Holmes design (Holmes, 1983), the spectrum match is based on the spectrum of radiated speech. The present system models, however, the transfer function only. The residues can be correctly determined (assuming that the underlying area function is known). Therefore, no formant shaping filters are required to avoid spectrum distortion. On the other hand, Holmes' design allows a formant amplitude to vary

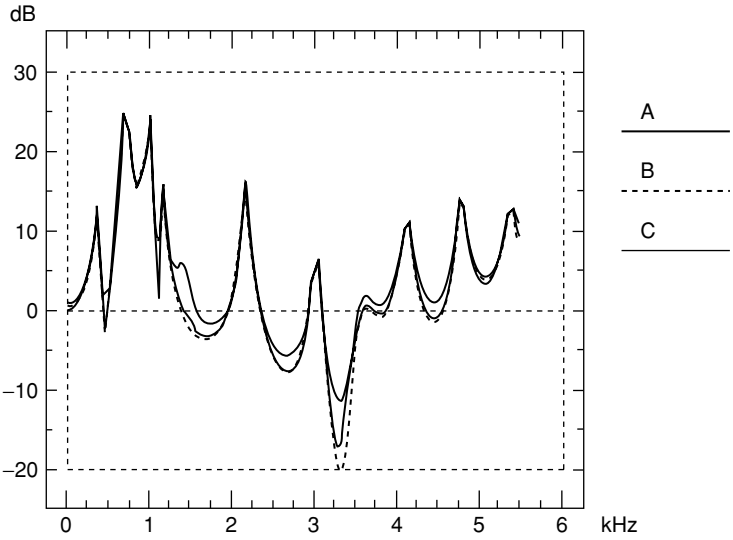


Figure 8. Transfer functions for a nasalized vowel. The first 11 poles are included. Curve A: Calculated transfer function; Curve B: Resynthesized transfer function, with the residues being corrected; Curve C: Same as Curve B but with no residue correction. Observe that the false hump around 1.4 kHz is removed by applying the correction term $C_{i,n}$, see text.

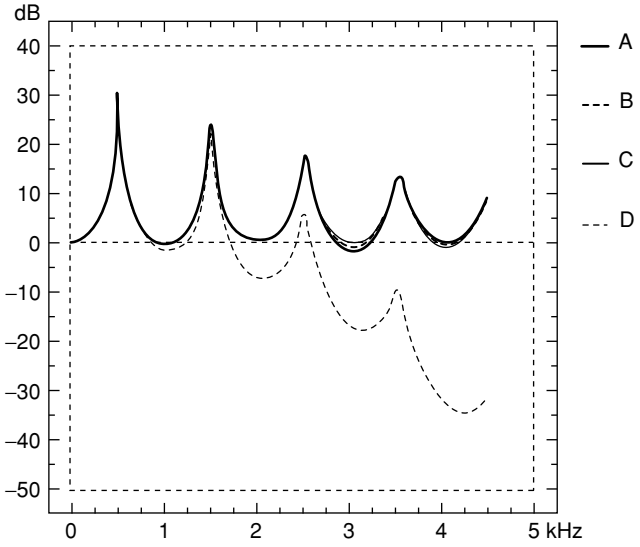


Figure 9. Transfer functions of the neutral tube, generated by the following methods: A: Synthesized transfer function by the parallel system; the residue data have been directly determined from a vocal tract computation; B: Same as A, but the residue data have been estimated based on a cascade model of formants. C: Synthesized transfer function by a cascade model; The effect of the higher pole contribution is included; D: Same as C, but without the higher pole contribution.

over a certain range thanks to the introduction of shaping filters. The present system does not generally have this feasibility, but by introducing a notch filter one can alter formant levels without disturbing the spectrum (Lin, 1990).

Information of all poles and zeros has been included when the residues are calculated. The higher pole/zero correction is therefore inherently and accurately preserved. This is an advantage over the conventional cascade formant synthesizer, where one has to specify the HPC term, explicitly or implicitly. Alternatively, the algorithm suggests a new method for specifying the HPC term.

The proposed system can also be compared with time-domain analogs, such as the transmission line analog or the reflection-type line analog, which will directly generate the time-domain functions. The present system may have difficulty in following rapid dynamic aspects of the articulation events. However, it has the advantage of i) preserving the accurate frequency-dependency of loss elements and of boundary conditions; ii) no constraints on the variation of the overall length of the vocal tract; iii) less computation time.

The computational efficiency is also a merit of the present system in comparison with a direct convolution method, see, for instance, Sondhi and Schroeter (1986). In the convolution method, the transfer function is first calculated by a frequency domain analysis and then its impulse response is determined by the inverse Fourier transform. The speech output is obtained by convoluting the resultant sequence with the source of excitation. In the proposed algorithm, the convolution is accomplished by a filtering process.

4. Concluding Remarks

A new algorithm for speech synthesis based on the vocal tract simulation has been described. Satisfactory results of spectrum match have been achieved. But it remains to be extended so that it can also deal with real poles. Such an extension may be of importance for the simulation of fricatives.

The algorithm can also be utilized to specify the driving-point impedance seen downstream from the glottis in terms of formant frequencies, bandwidths and residues at the poles. This specification is useful in studying the effects of the interaction between the glottal source flow and the acoustic load provided by the vocal tract during phonation.

The algorithm has been incorporated in an articulatory-based speech synthesis system currently under development at KTH.

Acknowledgements

The work was in part supported by the grants from the Swedish Board for Technical Development. The scholarship from the L M Ericsson Telephone Company (Stiftelsen för främjande av elektroteknisk forskning) to Qiguang Lin is gratefully acknowledged.

Qiguang Lin and Gunnar Fant

NOTE

* Paper presented at the 120th Meeting of the Acoustical Society of America, San Diego, November, 1990.

REFERENCES

Badin, P. & Fant, G. (1984): "Notes on vocal tract computation," *STL-QPSR* No. 2–3, pp. 53–107.
Fant, G. (1960): *Acoustic Theory of Speech Production*, Mouton, The Hague.
Holmes, J.N. (1983): "Formant synthesizers cascade or parallel?" *Speech Comm*. 2, pp. 251–273.
Lin, Q. (1990): *Speech Production Theory and Articulatory Speech Synthesis*, Ph.D. thesis, Dept. of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm.
Sondhi, M.M. & Schroeter, J. (1986): "A nonlinear articulatory speech synthesizer using both time- and frequency-domain elements", *Proc. ICASSP-Tokyo*, pp. 1999–2002.

APPENDIX I

TOMOGRAPHIC DATA

Fant, G. (1964). Formants and cavities. In E. Zwirner and W. Bethge, (eds.) *Proc. of the Fifth Intl. Congr. of Phonetic Sciences*, Munster. Basel: S Karger, 120–141.

[u]

[*a*]

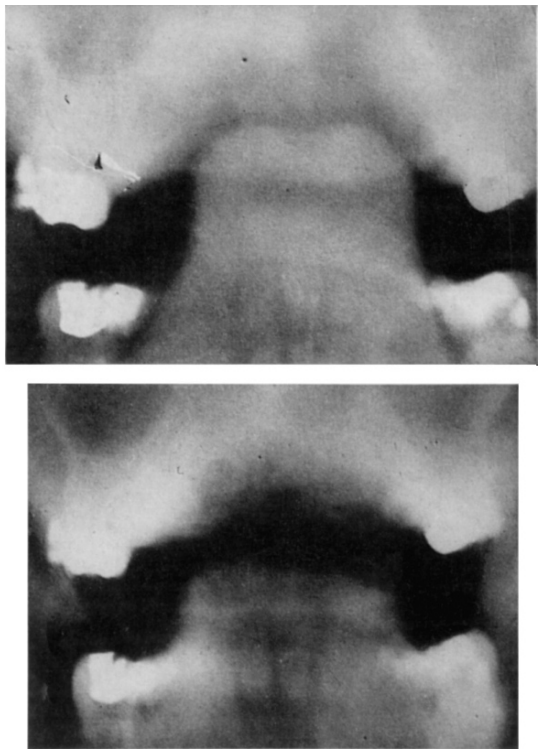


Figure A1. Vertical cuts through the mouth of the vowels [u:] and [ɑ:] and through the larynx. Horisontal cuts through pharynx for [ɑ:], [u:] and [i:].

[ɑ] [u]

[i]

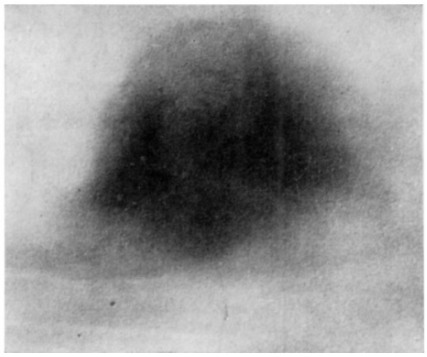


Figure A1. (*Continued*)

APPENDIX II

FEMALE/MALE FORMANT DATA

This is an abstract from Fant (1975B), a study of female/male vowel formant data from in all six languages, processed to bring out scale factors for each formant in 19 different vowels. Because of differences in vowel inventories, average values of a particular vowel have been derived from languages which posses comparable phonetic units. Thus, the vowel [ʉ:] is unique for Swedish.

The following languages have been exploited: Swedish, American English, Danish, Estonian, Dutch, Serbo-Croatian, see the original article for details.

Figure A1 shows F_2 versus F_1 and F_3 versus F_1 with female/male scale factors in percent for each vowel included. Scale factors ordered in a phonetic sequence for each of F_1, F_2 and F_3 appear in Figure 14 of Chapter 2.2. The overall average is 17 % higher formant frequencies in females than in males. A characteristic feature noted is the high scale factor for F_1 of [a] and the low values for both F_1 and F_2 of [u]. Low scale factors are also found for all close vowels, i.e. vowels with a low F_1. Typical of the vowel [i] is a low scale factor in F_1, and that the scale factor for F_2 is larger than for F_3. Major aspects of these relations can be explained from vocal tract anatomy, in the first place by the relatively shorter pharynx in females than in males. Similar anatomical differences exist comparing tenors versus bass singers, see Figure A1.2.

The pharynx length is the main determinant of F_2 of [i] with a scale factor of 21%, whilst F_3 of [i] is closer related to the mouth cavity and a scale factor of

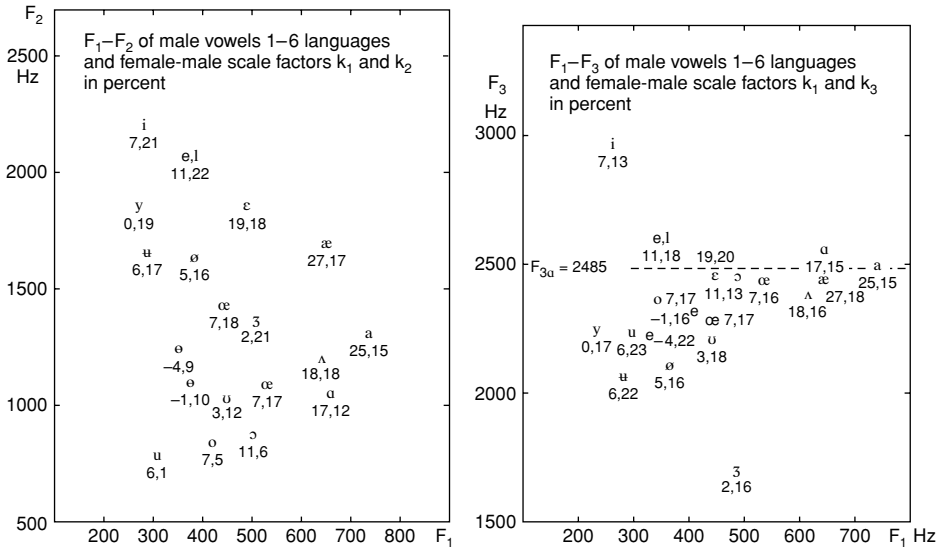


Figure A2.1. F_2 versus F_1 and F_3 versus F_1 of 19 vowels from 6 languages. Female/male scale factors are inserted.

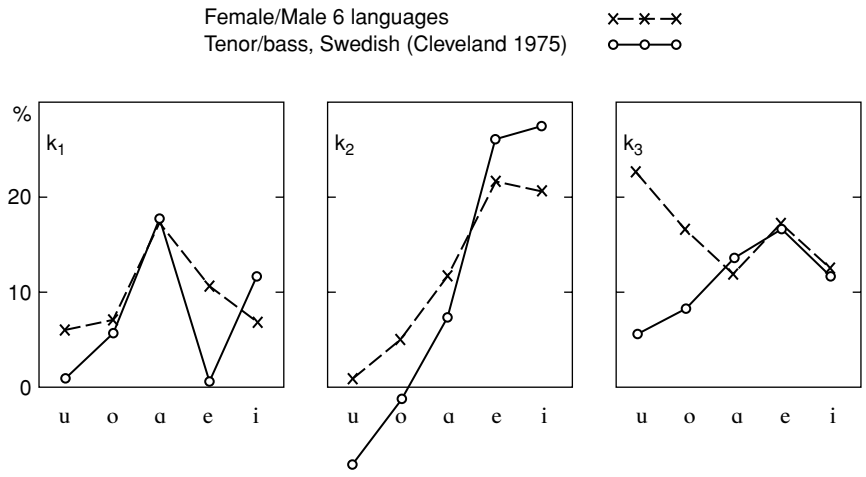


Figure A2.2. Tenor/bass formant frequency scale factors are similar to those of females/males. Data from T. Cleveland cited in Fant (1975B).

13%. We also know that formant F_3 of [i] has a greater perceptual importance than F_2.

The low scale factors of low F_1 vowels may be explained by similarities in vocal tract closing rather than in overall dimensions. The greater span between close and open vowels in female vowels to be related to anatomy rather than to perceptual demands.

A problem brought out in Fant (1975) is how to normalize vowel formant from individual speakers or from groups of speakers in order to remove anatomical constraints that can be deduced from formant frequency patterns. The first step is to normalize with respect to an average of F_3 in open vowels, which reflects the overall vocal tract length. An improvement may be achieved by a non-linear procedure, to take into account vowel category specific scale values.

This is an alternative to attempts of a direct perceptual approach. We still lack reliable methods of assessing absolute phonetic quality. Our normalization procedure has been tested in a study of two French dialects (Mettas and Fant, 1977).

REFERENCES

Fant, G. (1975 B). Non-uniform vowel normalization. *STL-QPSR* 2–3/1975, 1–19.
Mettas, O. and Fant, G. (1977). Front vowels in Parisian sociolects. *STL-QPSR* 2–3/1977, 1–7.

APPENDIX III

DIVER SPEECH

In 1964 I was consulted by a Swedish navy physiologist, Bertil Sonesson, about the speech of divers when inserted in a decompression chamber for gradual adjustment to normal pressure. He had observed a typical nasal quality and arranged for an X-ray experimental check on board the Navy submarine support vessel Belos. However, the nasal velar functions appeared to be normal, no objective signs of nasalization.

In a detailed theoretical study, supported by speech spectrograms, I was able to track down the cause to the vocal tract cavity walls participating in the tuning of formants. With increased air pressure there is a loss of impedance mismatch between the walls and the enclosed air. At normal air pressures the mass of the cavity walls and the enclosed air accounts for a finite resonance frequency of the closed tract at about 180 Hz. This is the lowest possible value of F_1 at complete closure. With increasing air pressure there in an increase of density and thus a decrease of the acoustic capacitance of the enclosed air. As a result F_1 of non-open vowels is raised, which accounts for the perceived nasal twang.

Under normal operation, divers breathe a gas mixture with helium as a substantial component together with oxygen and some minor part nitrogen. The origin of the

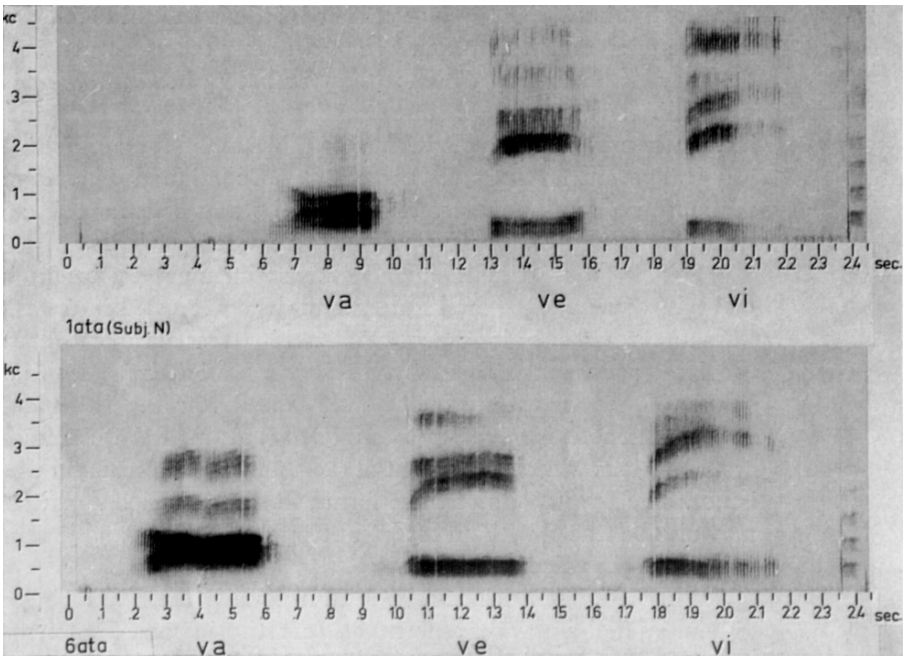


Figure A3. Speech recorded in a decompression chamber at 0 and 6 ata air pressure.

"Donald Duck" effect in helium speech is well known as a linear transposition of formants due to the increased velocity of sound.

I was now able to account for the combined effect of gas mixture and pressure, which included a prediction of conditions when the non-linearity of spectral shifts is minimized by an optimal combination of gas mixture and diving depth. These findings have been used for choice of gas mixtures and in the design of technical systems for speech restoration, so called helium speech unscramblers.

Fant, G. and Sonesson, B. (1964). Speech at high ambient air-pressure.*STL-QPSR* 2/1964, 9–21.
Fant, G. and Lindqvist-Gauffin, J. (1968). Pressure and gas mixture effects on divers' speech. *STL-QPSR* 1/1968, 7–17.

CHAPTER 3

THE VOICE SOURCE

A major field of study has been the voice source. Three articles are included here, supplemented by references to additional reading. The first article is the original publication of the LF-model (Fant, Liljencrants and Lin, 1985), which has found a wide use in synthesis applications. A more recent account is that of Fant (1995), see also article 4 in Chapter 6.

The second article (Fant and Lin, 1987) describes fundamental aspects of glottal source—vocal tract interaction. The third article (Fant and Lin, 1988) has a broader base of source-filter system function analysis, focusing on time and frequency domain properties and problems in inverse filtering. An extra formant in a vowel [ɛ] at about 1600 Hz for a female subject could be predicted from a production modelling assuming a finite glottal leakage.

An early study that deserves some attention is Fant (1979), in which formant excitation strength at instances of glottal opening and closing were derived by means of Laplace transform circuit theory.

Another study of general interest (Fant, 1982) contains data on glissando phonation, frequency domain matching of vowels, and data from recordings of true sub- and supraglottal pressure in connected speech.

Several articles have dealt with interaction phenomena. There is an acoustic interaction due to finite coupling between the sub- and the supraglottal systems, which accounts for the presence of extra poles and zeroes (resonances and antiresonances) in the overall system function.

As a consequence, time domain regeneration of glottal flow by means of inverse filtering often shows superimposed ripple of extra peaks and minima, see also article number 2 in chapter 4.

However, even when viewed in isolation, the spectrum of a glottal pulse may show a superimposed spectral ripple, especially when lacking prominent discontinuities.

Formant amplitudes are proportional to the supra-glottal pressure at the instant of excitation. A superimposed F1 oscillation will cause an acoustic interaction affecting formant amplitudes (Fant, 1986; Fant and Lin 1987). The interference can be appreciable. Thus, at a high F0 equal to F1, there will be a reinforcement of excitation above that of the mere centering of the fundamental at the resonance peak, and alternatively a reduced air consumption (Fant, 1986). With F_1 situated between harmonics there is accordingly a reduction.

Less radical but noticeable occurences of fluctuating formant amplitudes and spectral ripple are found in glissando phonation, i.e. when sustaining a vowel at increasing F0 (Fant 1982; Fant and Lin, 1987). These effects were originally documented by Fant et al. (1965). It is still unclear to what extent the observed ripple, in addition to acoustical interaction, is also caused by mechanical interaction from the F1 oscillation executing a finite force on the vocal folds.

A study of contextual variations of the voice source are found in Fant (1997).

SELECTED ARTICLES

[3.1] Fant, G., Liljencrants, J., and Lin, Q. (1985). A four-parameter model of glottal flow, *STL-QPSR* 4/1985, 1–13.

[3.2] Fant, G. and Lin, Q. (1987). Glottal source—vocal tract acoustic interaction. *STL-QPSR* 1/1987, 13–27.

[3.3] Fant, G. and Lin, Q. (1988). Frequency domain interpretation and derivation of glottal flow parameters. *STL-QPSR* 2–3/1988, 1–21.

ADDITIONAL READING

Fant, G., Fintoft, K., Liljencrants, J., Lindblom, B. and Martony, J. (1963). Formant-amplitude measurements. *J. Acoust. Soc. Amer* 35, 1753–1761.

Fant, G. (1979). Glottal source and excitation analysis. *STL-QPSR* 1/1979, 85–107.

Fant, G. (1982). Preliminaries to analysis of the human voice source. *STL-QPSR* 4/1982 1–27.

Fant, G. (1986). Glottal flow, models and interaction. *Journal of Phonetics,* **4** (3/4) Theme issue, Voice Acoustics and Dysphonia. Gotland, Sweden, August 1985, 393–399.

Lin, Q. (1987). Nonlinear interaction in voice production. *STL-QPSR* 1/1987, 1–12.

Fant, G. (1995). The LF-model revisited. Transformations and frequency domain analysis. *STL-QPSR* 2–3/1995, 119–155.

Fant, G. (1997). The voice source in connected speech. Speech communication 22, 125–139.

CHAPTER 3.1

A FOUR-PARAMETER MODEL OF GLOTTAL FLOW*

ABSTRACT

A glottal flow model with four independent parameters is described. It is referred to as the LF-model. Three of these pertain to the frequency, amplitude, and the exponential growth constant of a sinusoid. The fourth parameter is the time constant of an exponential recovery, i.e., return phase, from the point of maximum closing discontinuity towards maximum closure. The four parameters are interrelated by a condition of net flow gain within a fundamental period which is usually set to zero. The finite return phase with a time constant t_a is partly equivalent to a first order low-pass filtering with cutoff frequency $F_a = (2\pi t_a)^{-1}$.

The LF-model is optimal for non-interactive flow parameterization in the sense that it ensures an overall fit to commonly encountered wave shapes with a minimum number of parameters and is flexible in its ability to match extreme phonations. Apart from analytically complicated parameter interdependencies, it should lend itself to simple digital implementations.

THE L-MODEL

The four-parameter model, here referred to as the LF-model, has developed in two stages. The first stage was a three-parameter model of flow derivative, introduced by Liljencrants.

$$\frac{dU_g(t)}{dt} = E(t) = E_0 e^{\alpha t} \sin\omega_g t \tag{3.1.1}$$

It will be referred to as the L-model. It has the advantage of continuity whilst the early Fant (1979) model is composed of two parts, a rising branch:

$$\left.\begin{aligned} &U_g(t) = \frac{1}{2}U_0(1-\cos\omega_g t) \\ &\text{with derivative} \\ &E_1(t) = \frac{\omega_g U_0}{2}\sin\omega_g t \end{aligned}\right\} \quad \begin{aligned} &0 < t < t_p = \frac{1}{2F_g} \\ &\\ &\omega_g = 2\pi F_g \end{aligned} \tag{3.1.2}$$

and a descending branch at $t_p < t < t_p + \frac{\arccos\frac{K+1}{K}}{\omega_g}$

$$\begin{aligned} U_g(t) &= U_0[K\cos\omega_g(t-t_p) - K + 1] \\ E_2(t) &= -\omega_g K U_0 \sin\omega_g(t-t_p) \end{aligned} \tag{3.1.3}$$

This F-model has a discontinuity at the flow peak which adds a secondary weak excitation, see Fant (1979). One potential advantage of the L-model, Eq. (1), is that it can be implemented with a standard second-order digital filter with positive exponent (negative damping). The generated time function is interrupted at a time t_e when the flow

$$U(t) = E_0[e^{\alpha t}(\alpha\sin\omega_g t - \omega_g\cos\omega_g t) + \omega_g]/(\alpha^2 + \omega_g^2)] \tag{3.1.4}$$

has reached zero.

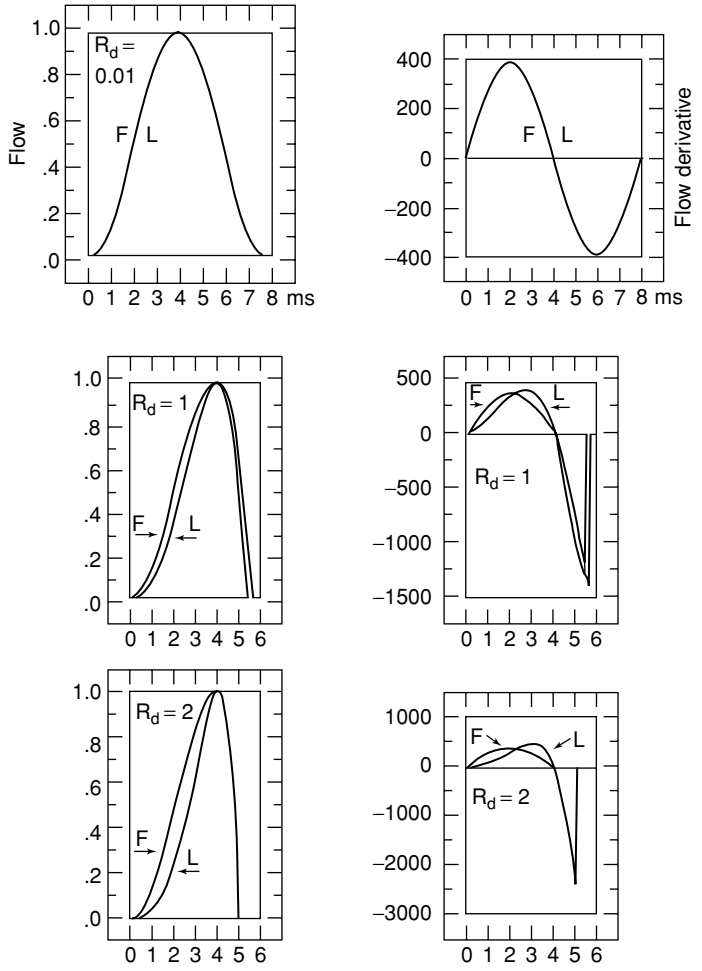


Figure 1A. Time-domain comparison of the F- and L-models.

The three algebraic parameters of the L-model, Eq. (1), E_0, α, and ω_g map on to the three basic flow derivative parameters

$$t_p = \frac{1}{2F_g}, \quad \omega_g = 2\pi F_g$$

t_e from Eq. (4)

$$E_e = -E_0 e^{\alpha t} \quad \sin\omega_g t_e \tag{3.1.5}$$

The F- and L-models share the parameter t_p or $F_g = 1/2\,t_p$. As shown in Fig. 1, the L-model displays a more gradual rise than the F-model given the constraint of equal t_e and E_e/E_i. This asymmetry increases with increasing E_e/E_i. The spectral

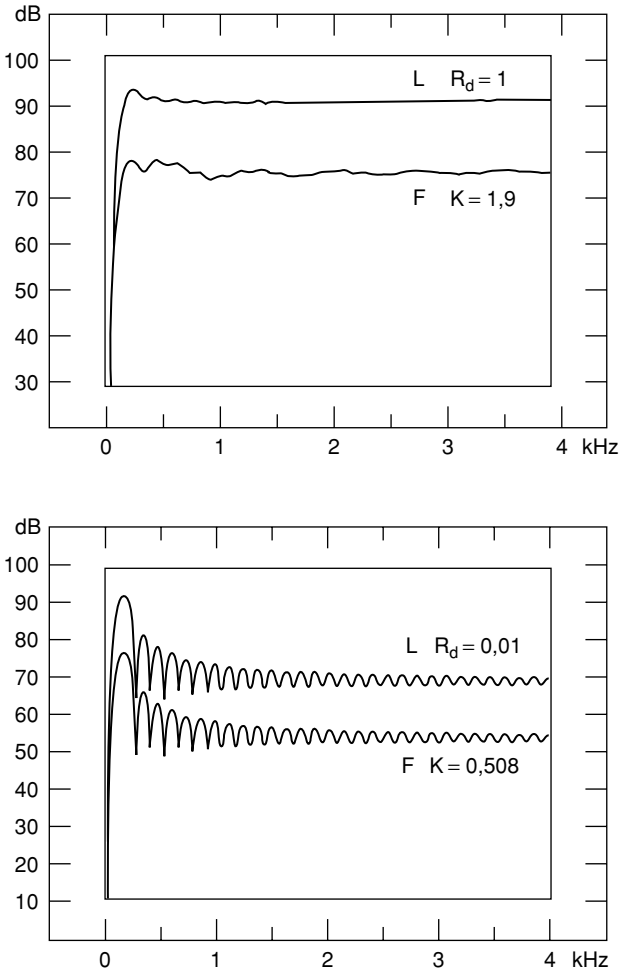


Figure 1B. Frequency-domain comparison of the F- and L-models. Flow spectrum +12 dB/oct pre-emphasis.

differences comparing the L- and F-models are not great. The L-model exhibits a lower degree of spectral ripple which is an advantage.

It is convenient to introduce the dimension-less parameters:

$$R_d = \frac{2\alpha}{\omega_g} = \frac{-B}{F_g}$$
$$R_k = \frac{t_e - t_p}{t_p} = \frac{t_n}{t_p} \tag{3.1.6}$$

Here, B is the "negative bandwidth" of the L-model exponent. R_d and R_k and E_e/E_i are mutually dependent shape parameters. The three-parameter L-model may

thus be conceived of having one shape parameter, one temporal scale factor, and one amplitude scale factor.

The shape factor of the F-model (Fant, 1979) is

$$K = \frac{1}{8}\left(\frac{E_e}{E_i}\right)^2 + \frac{1}{2} \tag{3.1.7}$$

Referring to a basic shape parameter, E_e/E_i, we may thus compare the F- and L-models as follows:

E_e/E_i	K	$R_d = \frac{2\alpha}{\omega_g}$	$R_k = \frac{t_e - t_p}{t_p}$	
0	0.5	0	1	(sine wave)
1	0.625	0.12	0.73	
2	1	0.43	0.54	
3	1.625	0.84	0.42	
4	2.5	1.35	0.33	

The location t_i of the flow derivative positive maximum E_i is

$$t_i = (t_p/2)[1 + (2/\pi)\text{artg}(R_d/2)] \tag{3.1.8}$$

which defines an asymmetry factor

$$R_i = (t_i - t_p/2)/(t_p/2) = (2/\pi)\text{artg}(R_d/2) \tag{3.1.9}$$

At sufficiently small K-values in the F-model, $K < 1$, there is a negative smooth turning point prior to the closure step. In the L-model a corresponding negative smooth turning point occurs at

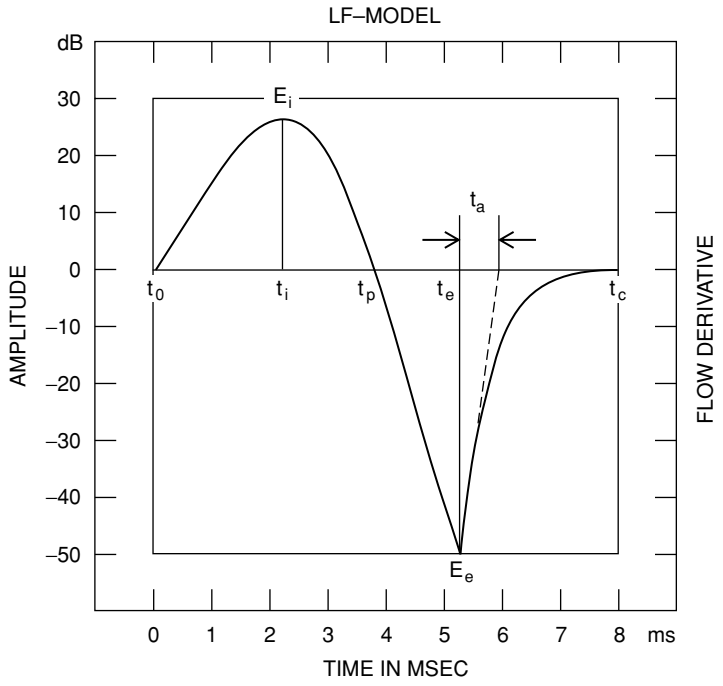
$$t = (3t_p/2)[1 + (2/3\pi)\text{artg}(R_d/2)] \tag{3.1.10}$$

which requires $R_k > 0.54$ or $R_d < 0.43$. This property allows both the F- and the L-model in a limit to approach a sine wave, $K = 0.5$ and $R_d = 0$, which is of some relevance to the termination of voicing.

THE LF-MODEL

The obvious shortcoming of a model with abrupt flow termination is that it does not allow for an incomplete closure or for a residual phase of progressing closure after the major discontinuity. Either and generally both of these conditions account for a residual phase of decreasing flow after the discontinuity. In voiced h-sounds, i.e., in breathy phonations, the discontinuity may occur in the middle of the descending branch at the flow followed by a less steep descent and a final trailing off corner effect (see Fant, 1980, 1982). In less breathy phonations, the discontinuity is found further down towards the base line of the flow.

Accumulated experimental evidence the last years has shown that this corner effect is more a rule than an exception. In his five-parameter model,



$$E(t) = E_0\, e^{\alpha t} \sin \omega g t \qquad (t < t_e)$$

$$E(t) = \frac{-E_0}{\varepsilon t_a} \cdot \left[e^{-\varepsilon (t - t_e)} - e^{-\varepsilon (t_c - t_e)} \right] \qquad (t_e < t < t_c)$$

Figure 2. The LF-model of differentiated glottal flow. The four wave-shape parameters t_p, t_e, t_a, and E_e uniquely determine the pulse ($t_c = T_0 = 1/F_0$).

Ananthapadmanabha (1984) has accordingly included a terminal return phase which is modeled as a parabolic function. We have chosen to adopt an exponential function

$$E_2(t) = \frac{-E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right] \qquad (3.1.11)$$
$$t_e < t < t_c$$

As shown in Fig. 2, the time t_a is the projection of the derivative of $E_2(t)$ at $t = t_e$ on the time axis. For a small t_a, $\varepsilon = 1/t_a$, and otherwise

$$\varepsilon t_a = \left[1 - e^{-\varepsilon(t_c-t_e)}\right] \qquad (3.1.12)$$

In practice it is convenient to set $t_c = T0$, i.e., the complete fundamental period.

The sole remaining independent parameter of the return phase is, thus, its experimentally determined effective duration t_a. The area under the return branch, i.e., the residual flow at t_e is written:

$$U_e = \frac{E_e \cdot t_a}{2} \cdot K_a \tag{3.1.13}$$

When t_a is small, $K_a = 2$, otherwise an integration of Eq. (11) may be approximated by the following procedure. Calculate the parameter

$$R_a = \frac{t_a}{t_c - t_e} \tag{3.1.14}$$

$$\left.\begin{aligned} &\text{If } R_a < 0.5 \\ &\qquad K_a = 2 - 2.34\, R_a^2 + 1.34\, R_a^4 \\ &\text{If } R_a > 0.5 \\ &\qquad K_a = 2.16 - 1.32\, R_a + 0.64(R_a - 0.5)^2 \\ &\text{If } R_a < 0.1 \\ &\qquad K_a = 2.0 \end{aligned}\right\} \tag{3.1.15}$$

We now know the residual flow U_e. By imposing the final requirement of area balance, i.e., zero net gain of flow during a fundamental period

$$\int_0^{T0} E(t) = 0$$

we may, accordingly, start out from the four waveform parameters, t_p, t_e, t_a, and E_e and solve for the α or $R_d = 2\alpha/\omega_g$ of Eqs. (4) and (5) to provide a flow matching that of Eq. (13).

Prior to this operation it is convenient to normalize flow functions by a division by E_e and t_p. The scale factor E_0 for synthesis can now be determined from Eq. (5).

The amplitude E_i of the positive maximum is found from

$$\frac{E_i}{E_e} = \frac{-e^{\frac{R_d}{2}\left(-\frac{\pi}{2} + \operatorname{artg}\frac{R_d}{2} - \pi R_k\right)} \cdot \sin\left(\frac{\pi}{2} + \operatorname{artg}\frac{R_d}{2}\right)}{\sin \omega_g t_e} \tag{3.1.16}$$

The computer program for implementation is organized as follows.

Perform inverse filtering. Shift, if necessary the base line to achieve area balance, i.e., zero net change of flow. If the zero line is judged reliable, note the residual net flow for further use. Mark with a cursor the starting time $t_0 = 0$ and the subsequent locations of t_p, t_e, $t_e + t_a = t_r$ and the complete period length T0. Measure E_e.

Go through the procedure above, Eqs. (4)(5), and (13) with normalized flow functions $U_e/(E_e \cdot t_p)$. Apply an iterative solution starting from zero return flow to

find R_d and then E_0 and E_i. Synthezize the model curve and compare it with the inverse filtered curve. If E_e/E_i does not fit, try to shift the starting point by adding a fixed quantity to t_p and t_e. Update the parameters and check the result.

The results so far have shown this procedure to provide not less reliable results than the five-parameter model of Ananthapadmanabha (1984) and will be adopted as our standard. It also has the advantage of peak continuity and that it better matches highly aspirated waveforms and allows a sinusoid as an extreme limit.

PROPERTIES OF THE LF-MODEL

An introduction of dynamic leakage, i.e., of a finite return phase, has to be compensated by a decrease of α, i.e., of $R_d = 2\alpha/\omega_g$ and, thus, also of the peak ratio $A_e = E_e/E_i$ in order to maintain the area balance, i.e., the zero net gain of flow within a fundamental period. These trading relations are apparent from the nomograms of Figs. 3 and 4. which can be used for quantifying parameter interdependencies.

The effect on the flow derivative and flow pulse shape on the second derivative flow spectrum (+12 dB/oct versus flow spectrum) of a finite t_a is illustrated in Fig. 5. Here $R_k = (t_e - t_c)/t_p = 0.5$. Increasing t_a causes an increased high frequency deemphasis, as noted by Ananthapadmanabha (1984). Due to the exponential wave shape of the return phase, it may be modeled as an additional first-order low-pass function of cutoff frequency

$$F_a = \frac{1}{2\pi \cdot t_a} \tag{3.1.17}$$

As evidenced from Fig. 5, this is a good approximation for assessing the main spectral consequence of dynamic leakage. Even a very small departure from abrupt termination causes a significant spectrum roll-off in addition to the standard −12 dB/oct glottal flow spectrum. Thus, $t_a = 0.15$ ms accounts for a 3-dB fall at $F_a = 1060$ Hz and 12 dB at 4000 Hz. In general, the loss is

$$\Delta L = 10\log(1 + f^2/F_a^2)\,\text{dB}$$

A variation of $A_e = E_e/E_i$ at constant E_e and $t_a = 0.1$ms and $t_p = 2.7$ms is illustrated in Fig. 6. Observe the significance of the negative spike E_e to set the spectrum level at frequencies above Fg. The lowpass filter effect is approximately independent of A_e. The spectral ripple increases with decreasing A_e and with increasing t_a. Increasing A_e at constant excitation E_e is related to a more "pressed" voice which attains a "thinner" quality due to the relative lower low-frequency level.

Fig. 7 shows a sequence of two complete periods illustrating typical wave shapes and spectra of breathy phonation as in voiced [h], $t_a = 1.3$ ms, and an intermediate wave shape, $t_a = 0.2$ ms in addition to abrupt closure, $t_a = 0$. The constants have been chosen to provide similarity with the inverse filter data of Fig. 7 in Fant (1980).

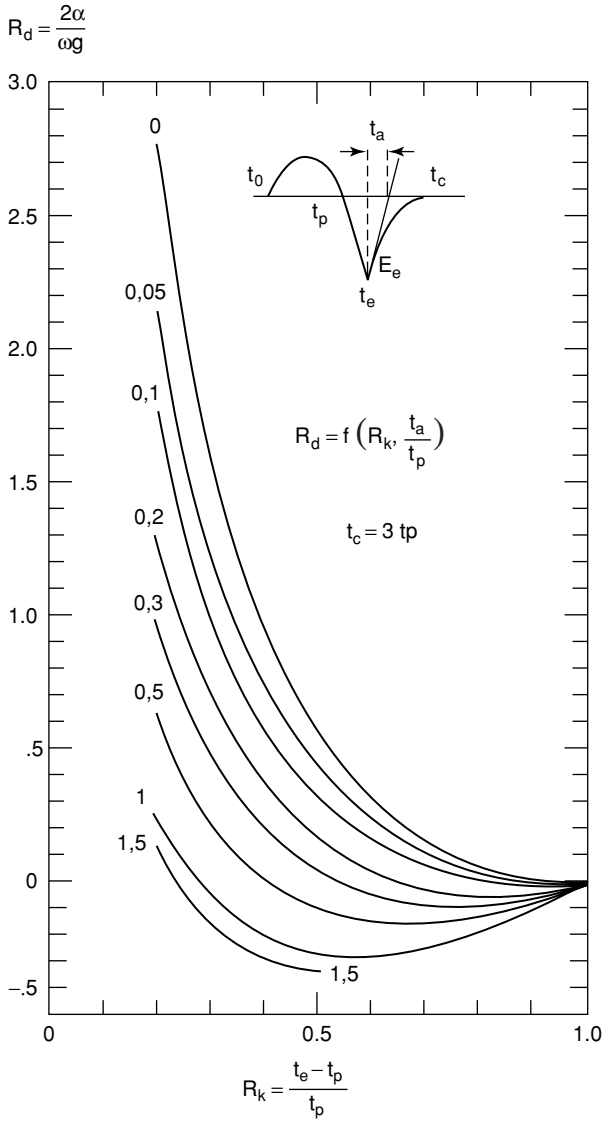


Figure 3. LF-model nomograms relating the exponent to the relative excitation timing.

Results from our present laboratory set-up for inverse filtering and glottal source analysis developed by Ananthapadmanabha (1984) are shown in Fig. 8. The object is a male [a] which has assimilated some breathiness from a following aspirated stop. Here there is a combination of an extreme large peak ratio $A_e = 4$ and a significant leakage, $T_a = 0.43$ ms, similar to that of Fig. 7. Under the oscillogram follows a five-parameter match with the A-model and then the four-parameter LF-model. In

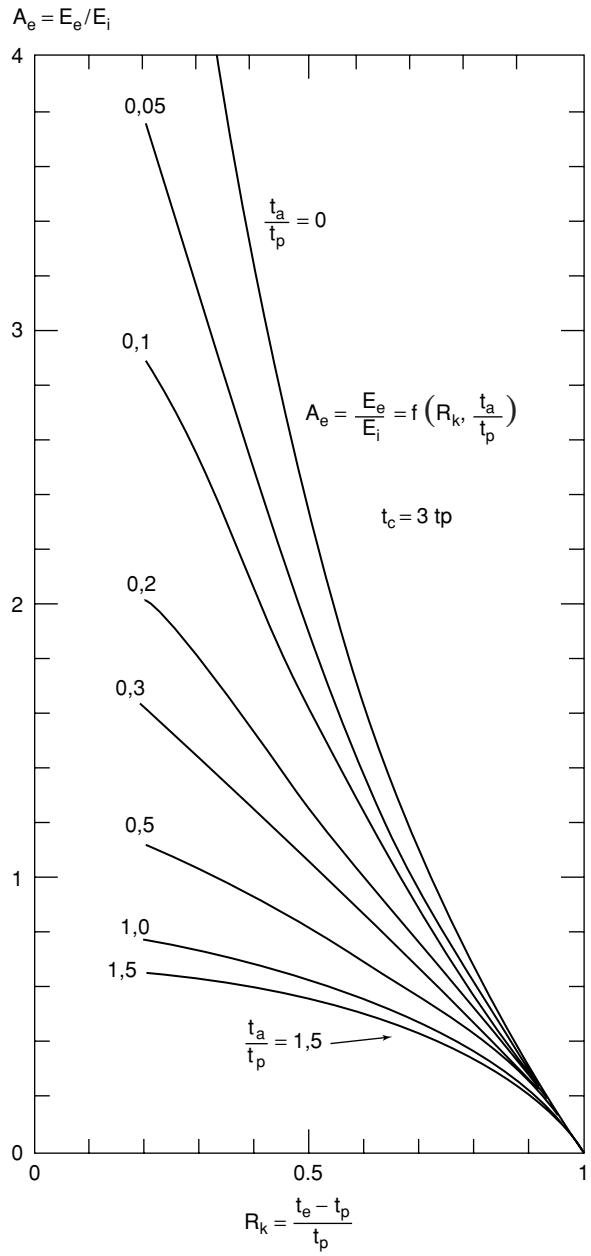


Figure 4. LF-model nomograms relating the negative to positive peak ratio to the relative excitation timing.

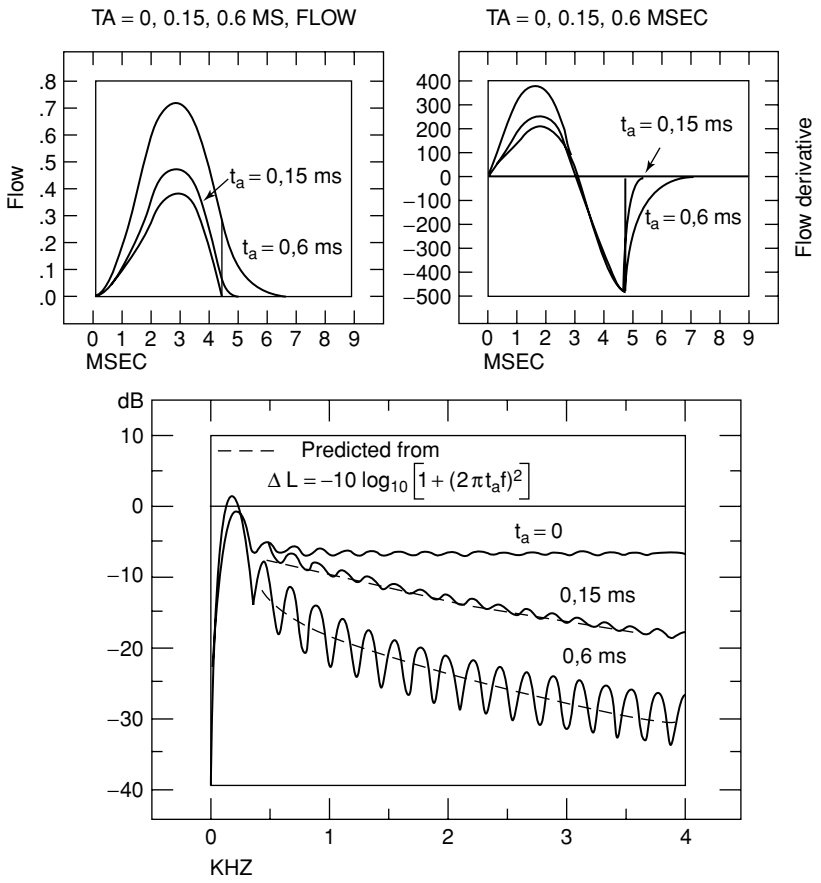


Figure 5. Waveshape and spectral correlates of increasing the return time constant t_a.

this case the LF-model provides a better overall fit. In less extreme cases the two models perform equally well. The initial LPC-spectrum analysis prior to inverse filtering does not fit the spectrum very well. This is in part due to the low sampling frequency in this special case.

The LF-model is mainly intended for a noninteractive modeling of the voice source but it could, like the Fant model, also be applied to the description of glottal area functions. A discussion of interactive phenomena is found in Ananthapadmanabha & Fant (1982), Fant & Ananthapadmanabha (1982), and Fant (1982). A perceptual evaluation of interaction has been reported by Nord, Ananthapadmanabha, & Fant (1984).

Gunnar Fant, Johan Liljencrants, and Qi-guang Lin

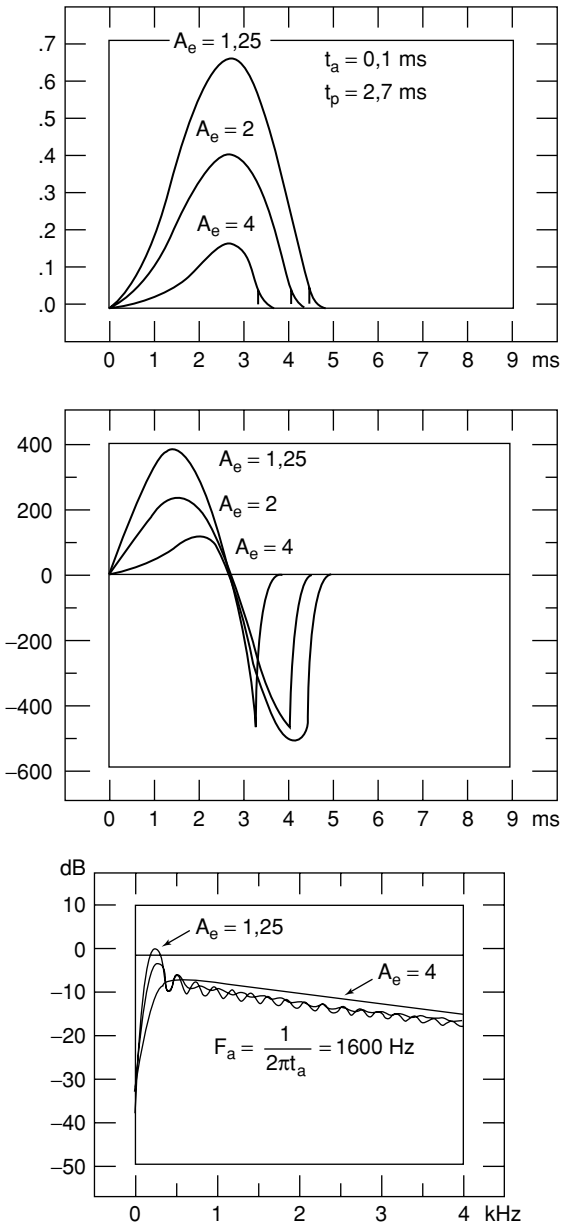


Figure 6. Increasing the peak factor A_e at constant excitation amplitude E_e and a constant return time constant $t_a = 0.1$ ms.

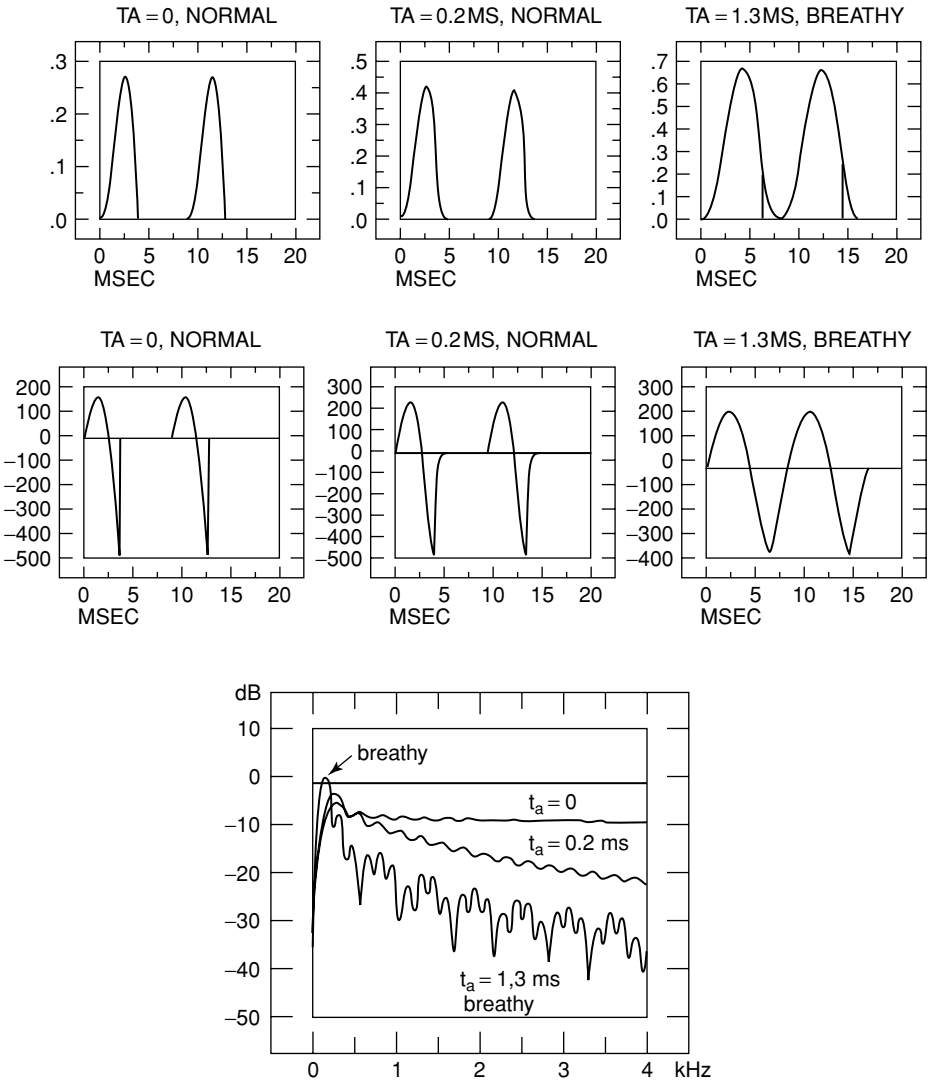


Figure 7. LF-modeling of breathy voicing.

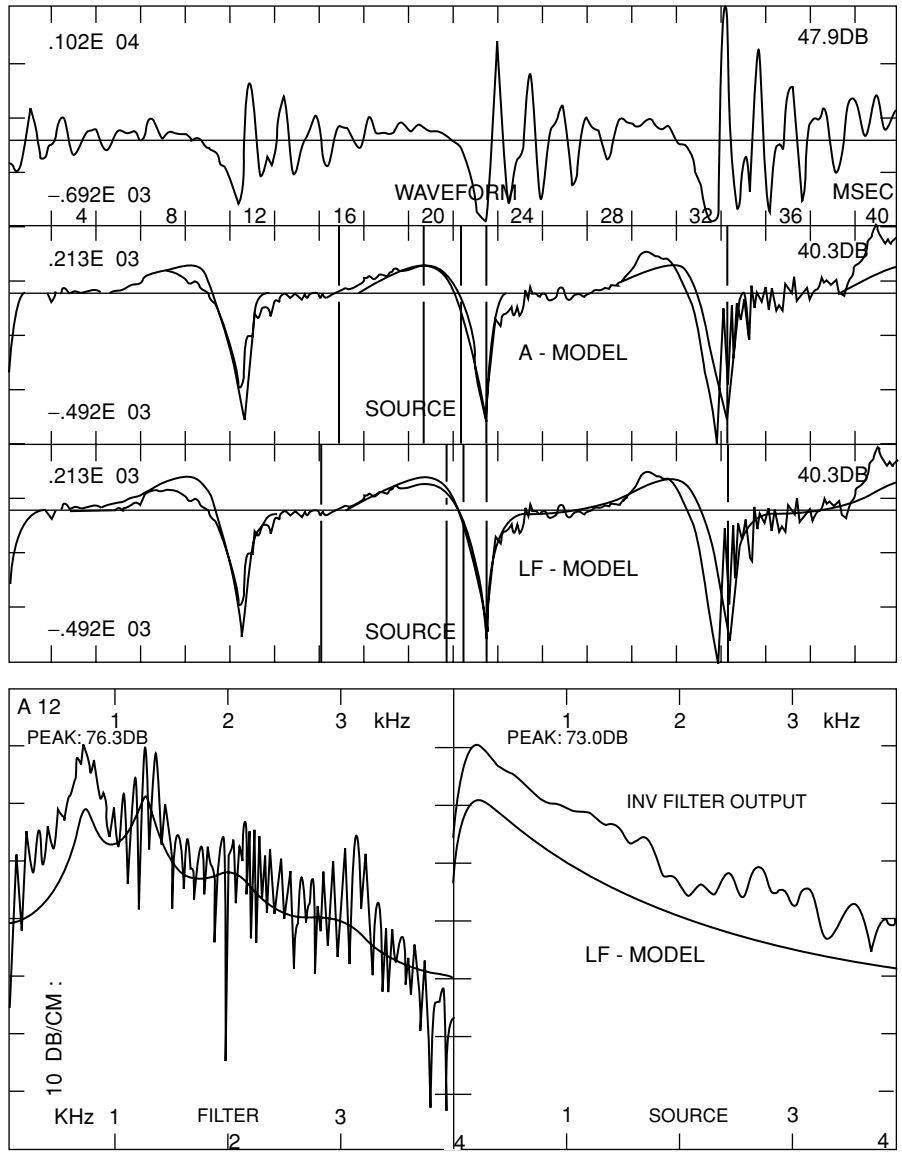


Figure 8. Waveform matching with A-model and LF-model.

NOTE

* Paper presented at the French-Swedish Symposium, Grenoble, April 22–24, 1985.

REFERENCES

Ananthapadmamnabha, T.V. (1984): "Acoustic analysis of voice source dynamics", STL-QPSR 2–3/1984, pp. 1–24.

Ananthapadmanabha, T.V. & Fant, G. (1982): "Calculation of true glottal flow and its components", STL-QPSR 1/1982, pp. 1–30; also in Speech Communication *1* (1982), pp. 167–184.

Fant, G. (1979): "Vocal source analysis—a progress report", STL-QPSR 3–4/1979, pp. 31–54.

Fant, G. (1980): "Voice source dynamics", STL-QPSR 2–3/1980, pp. 17–37.

Fant, G. (1982): "Preliminaries to the analysis of the human voice source", STL-QPSR 4/1982, pp. 1–27.

Fant, G. & Ananthapadmanabha, T.V. (1982): "Truncation and superposition", STL-QPSR 2–3/1982, pp. 1–17.

Nord, L., Ananthapadmanabha, T.V., & Fant, G. (1984): "Signal analysis and perceptual tests of vowel responses with an interactive source filter model", STL-QPSR 2–3/1984, pp. 25–52.

CHAPTER 3.2

GLOTTAL SOURCE—VOCAL TRACT ACOUSTIC INTERACTION*

ABSTRACT

Recent developments within our group of voice source—vocal tract acoustic interaction are reviewed. Special emphasis is layed on non-linear superposition phenomena, i.e., how the excitation within a period is dependent on the past history of vocal tract oscillations and their residual components within the transglottal pressure. A study of breathy phonation shows that constant leakage affects the voice source slope less than does the dynamic leakage in terms of a residual closing phase. A simulation of a female voice source is attempted.

INTRODUCTION

Acoustic interaction we define as the dependency of the glottal volume velocity flow $U_g(t)$ on supraglottal articulations under the constraints of a prescribed glottal area function $A_g(t)$. Mechanical interaction, on the other hand, is the dependency of glottal vibratory patterns and thus of $A_g(t)$ on the overall state and aerodynamics. The complete interaction thus has a mechanical and an acoustic component.

The source filter model we employ for simulations of interaction as well as for inverse filtering states that a voiced speech sound is the convolution of the true glottal flow and the supraglottal transfer function. It is thus the combination of a complex and dependent source and a linear, theoretically well defined transfer function. A practical complication in inverse filtering is the difficulty of attaining an accurate formant tracking and to estimate "closed glottal conditions" even if they do not exist as in breathy voicing.

The objectives of our simulations presented here are to contribute to an insight in the factors that determine acoustic interaction. The glottal flow $U_g(t)$ and its derivative $U_g'(t)$ and the spectrum of flow derivative $U_g'(f)$ have been calculated with a variation of vocal tract network parameters and glottal area functions. A main finding is the great dependency of interaction on the past history of the pressure-velocity state within the vocal tract. Basic theory and more detail data have been presented in earlier publications. This is to be considered an informal progress report of results from simulations supplementing earlier work; see the list of references.

We have not yet exploited the full potentialities of the theoretical modeling. Much remains to be done in systematic studies of the influence of the many components involved and of different voice types including female and childen's voices. Also we need more experience from confronting the model with data from inverse filtering of real speech and evaluating the perceptual importance of various aspects of interaction.

THEORY

The basic theory, see Ananthapadmanabha & Fant (1982) and more recently, Fant, Liljencrants, & Lin (1985a); Fant, Lin, & Gobl (1985b); Lin (1987), focuses on the instantaneous transglottal pressure.

$$\Delta P = P_\ell - (P_{in} + P_{sg}) = \frac{k\varrho v_0^2 |x| x}{2} + \frac{12\mu d \ell^2}{A_g^2} v_0 x + \varrho d v_0 x' \tag{3.2.1}$$

The three terms at the right represent pressure drops associated with glottal kinetics, viscosity and inductance. We have introduced the normalized particle velocity x.

$$\left.\begin{aligned} &U_g(t) = A_g(t) \cdot v_0 \cdot x(t) \\ &x = v_g(t)/v_0 \quad v_0 = (2P_\ell/k\varrho)^{\frac{1}{2}} \quad d = A_g(t) \cdot L_g(t) \end{aligned}\right\} \tag{3.2.2}$$

P_ℓ is lung pressure, d denotes glottal depth and ℓ glottal length. The glottal shape factor k is set to 1.

The convolution $P_i(t) = U_g(t)^* Z_i U_g(t)^* Z_i d(t)/(t)$ is handled by treating $Z_i(t)$ as the inverse Lapalce transform of the partial fraction expansion of Z(s) up to five formants and a realization with standard second order digital filters. Similarly, three formants are included for the subglottal impedance. Eq. (1) is discretized and solved as a quadratic equation without any iterations which is an improvement compared to the iterative procedure employed by Ananthapdamanabha & Fant (1982), see also Fant et al. (1985a; 1985b). In order to allow negative going flow in the glottal impedance, which is not unrealistic, we have substituted $|x|x$ for x^2.

The introduction of the relative particle velocity x allows a greater computational accuracy, simplifies the expression for the pressure drop across the glottal inductance and is a direct compound measure of interaction. If we temporarily neglect sub- and supraglottal loads, and the glottal inductance L_g and the viscosity term R_g, we end up with a constant $x = 1$ and a constant reference particle velocity

$$v_0 = 3709(P_\ell/8)^{0.5} \tag{3.2.3}$$

where P_ℓ is the lung pressure in cm H_2O.

The flow pulse

$$U_g(t) = A_g(t) \cdot v_0 \cdot x(t) = A_g(t) \cdot v_0 \tag{3.2.4}$$

is now fully proportional to the glottal area function. This is the so-called "short circuit" flow $U_{sc}(t)$ when the complete load is introduced. The function x(t) varies significantly within the glottal open period and accounts for:

(1) Pulse-skewing, generally to the right due to total vocal tract and glottal inductance. The flow is delayed with respect to the glottal area. This effect has been closely studied by Rothenberg (1981).

(2) Overlayed ripple caused by short-time transglottal pressure variations from formant oscillations evoked at the pulse onset and in previous glottal periods. As a rule, the contribution from excitations in a previous pulse overrides components originating from the particular flow pulse under observation. The quadratic kinetic term accounts

for a non-linear conversion from instantaneous transglottal pressure to flow. We thus have a nonlinear superposition, the past history of vocal tract oscillations adding to the details of a glottal source pulse.

(3) Short-term variations of formant bandwidth and frequency within the time span of a glottal pulse. These effects may be hard to separate from multiple excitations within a flow pulse adding to discontinuities of instantaneous frequencies and time-domain envelopes. A heavy damping is referred to as truncation. Ripple and truncation effects have been extensively studied by Ananthapadmanabha & Fant (1982); Fant & Ananthapadmanabha (1982); Fant & et al. (1985a; 1985b).

In the present study we find significant ripple and truncation effects even in higher formants, F_2, F_3 and F_4. All these interaction effects add to a random fine structure of shape amplitude and excitation time locations within a glottal pulse train. In the frequency domain, they are reflected by a few extra humps and valleys in the source flow spectrum. These features, as well as the mechanical component of interaction, seem to contribute to the naturalness of speech.

One finding of relevance to the characteristics of children's and female voices is that source spectrum slope is less influenced by the presence of a leakage than by concommittant changes in glottal area function shape, specifically the dynamic leakage, i.e., the residual closing phase.

Gunnar Fant and Qiguang Lin

NOTE

* Paper EE24, 113th Meeting of the Acoustical Society of America, May, 1987.

REFERENCES

Ananthapadmanabha, T.V. & Fant, G. (1982): "Calculation of true glottal flow and its components", Speech Communication *1*, pp. 167–184.

Fant, G. (1982): "Preliminaries to the analysis of the human voice source", STL-QPSR 4/1982, pp. 1–27.

Fant, G. (1986): "Glottal flow: Models and interaction", pp. 393–399 in (B. Fritzell & G. Fant, eds.), Voice Acoustics and Dysphonia, Theme Issue, J. of Phonetics *14*, No. 3/4.

Fant, G. & Ananthapadmanabha, T.V. (1982): "Truncation and superposition", STL-QPSR 2–3/1982, pp. 1–17.

Fant, G., Liljencrants, J., & Lin, Q. (1985a): "A four-parameter model of glottal flow", STL-QPSR 4/1985, pp. 1–13.

Fant, G., Lin, Q., & Gobl, C. (1985b): "Notes on glottal flow interaction", STL-QPSR 2–3/1985, pp. 21–45.

Lin, Q. (1987): "Nonlinear interaction in human voice production", paper to be presented at the Int. Conf. of Chinese Information Processing, Beijing, P.R. of China.

Rothenberg, M. (1981): "An interactive model for the voice source", STL-QPSR 4/1985, pp. 1–17.

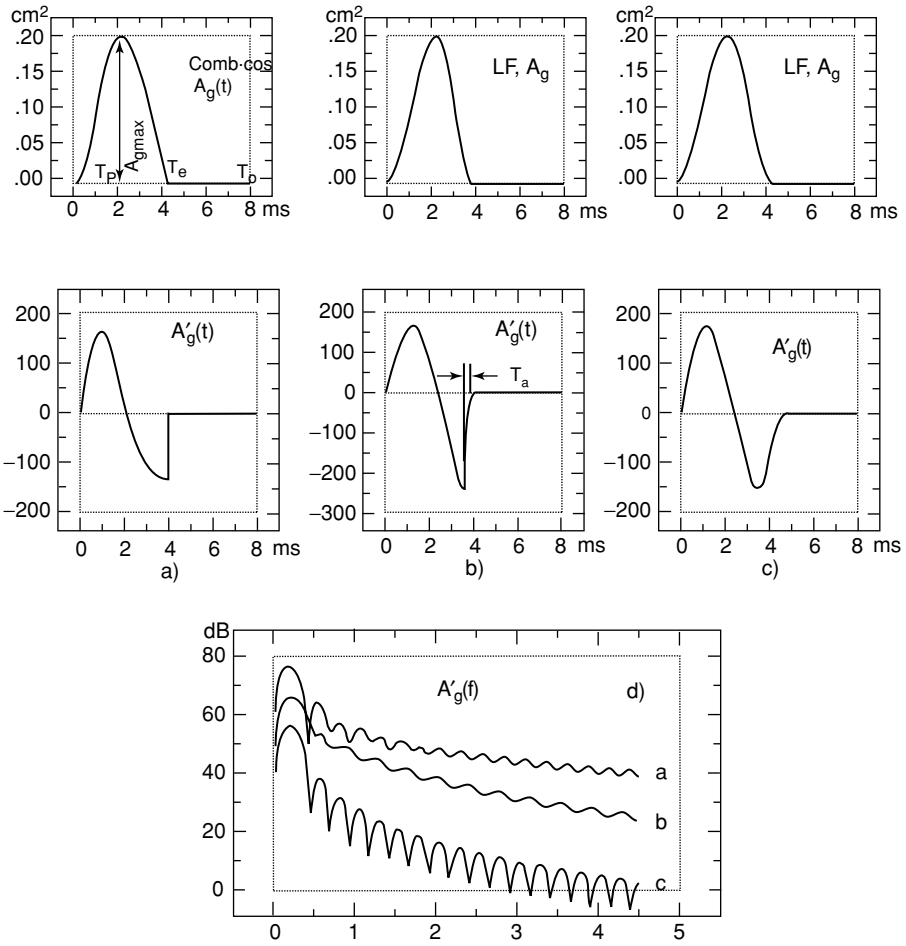


Figure 1. Examples of the glottal area function models employed in this study.

a) A raised cosine waveform and its derivative form;
b) An LF model and its derivative form whose main characteristic is the presence of the gradual return phase after the major closure which we used for simulating breathy voice and female and children's voices.
c) Another example of the LF model is shown.
d) The corresponding spectra, $A'_g(f)$.

The open quotient is defined as the ratio of T_e/T_0, and the symmetry factor as $T_p/(T_e - T_p)$.

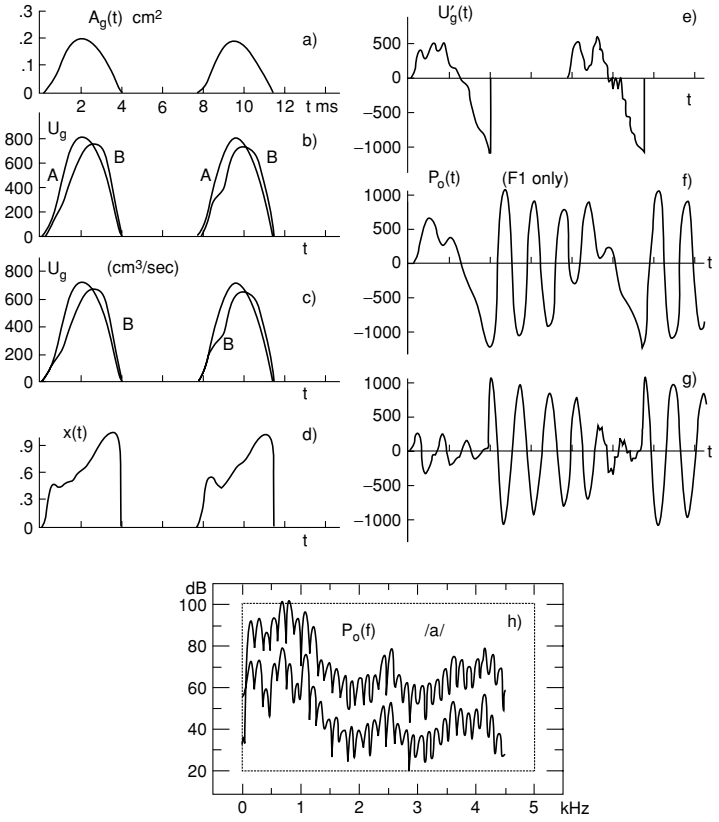


Figure 2.

a) Tracing of glottal area function (a combined cosine wave). Here $T_p = 2$ ms, $T_e = 4$ ms, $T_0 = 8$ ms, $A_{gmax} = 0.2\text{cm}^2$.
b) Glottal source flow. A comparison is made between a short circuit (glottal inductance and glottal viscosity are set to zero) and a compound vowel, /a/, loading. Curve A pertains to the short circuit case, while Curve B corresponds to the interactive model. It can be seen that Curves A and B differ in SKEWNESS: Curve A is skewed to the right side and RIPPLE: overlayed ripple components are seen in Curve A. These effects are the main consequences of interaction.
c) The same as b) except for the presence of a finite glottal inductance and glottal viscosity when computing the glottal flow for the short circuit case. Curve B is identical to Curve B in b). It is shown by comparing b) and c) that under the short circuit circumstance, the finite glottal inductance and viscosity will make the glottal flow onset more gradual, i.e., a corner-rounding effect.
d) $x = V_g / V_0$, the normalized particle velocity, see text for details; x is an indicator of interaction. In a conventional linear source model, it is constant, $x = 1$ when the glottis is open and $x = 0$ when the glottis is closed. The product $x(t)A_g(t)V_0$ is the computed true glottal flow.
e) Differentiated glottal flow, $U'_g(t)$. Here the ripple is more apparent than in $U_g(t)$. The negative peak amplitude becomes a scale factor of formant excitation.
f) The lip pressure output (defined as the derivative of the volume flow at the lips) after selected inverse-filtering (without cancellation of the F1 oscillation).
g) The same as f) except that the source component is eliminated here by a double-differentiation. The F1 oscillation decays faster as the glottis is open due to the coupling to the glottal internal impedance.
h) FFT spectrum of the sound pressure output, vowel /a/. The Hamming window covers a time span of three fundamental periods. A close comparison is made between the interactive (top curve) and the linear models (bottom). It is seen that the peaks of F1, F2, F4 and F5 are broadened in the top curve and that the valleys between F0 and F1 and beween F1 and F2 have been "filled in" in the interaction case. See Fig. 3 for further discussions on spectral details. The interaction has also increased the formant amplitudes, viz., the fundamental and the second harmonic amplitude which derive from the skewing effect.

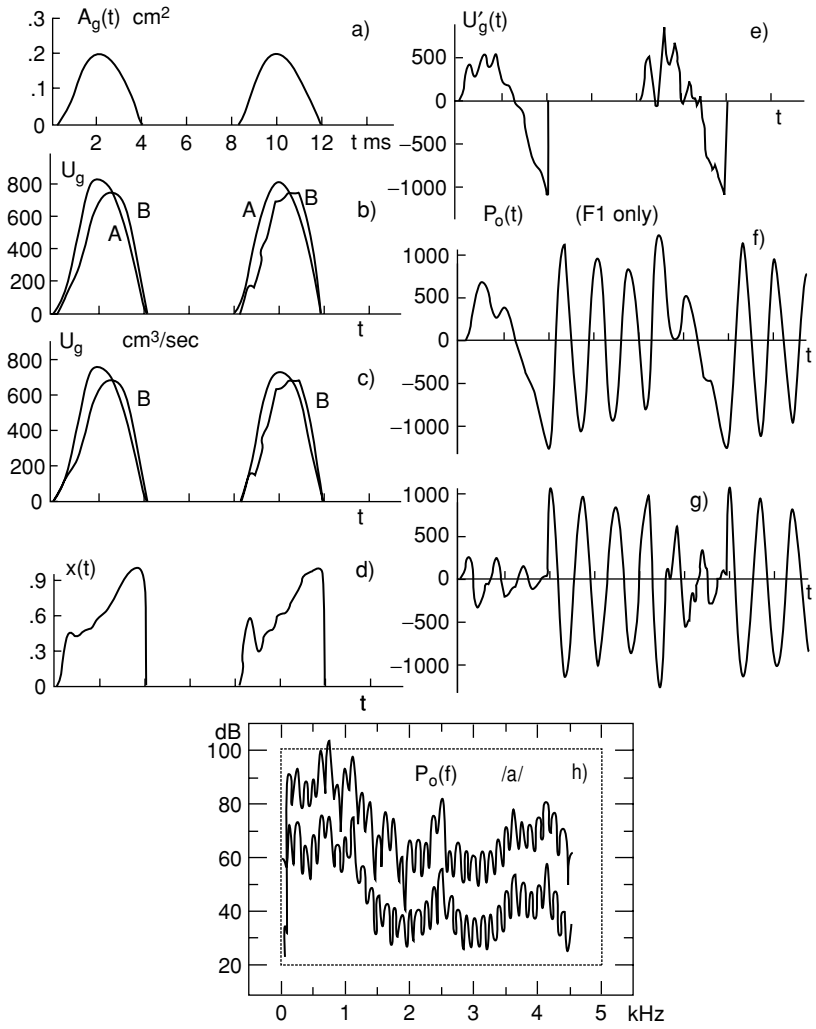


Figure 3. The same simulation condition as in Fig. 2 except for $T_0 = 8$ ms instead of 7.5 ms as in Fig. 2.

It is observed that a pseudo-peak between F2 and F3 appears in the sound pressure spectrum of the interaction model (the top curve in the bottom graph). Such a peak is not visible when $T_0 = 7.5$ ms. This means that the nonlinear superposition effect is very critical. This is especially true when some formants of a sound have a small bandwidth.

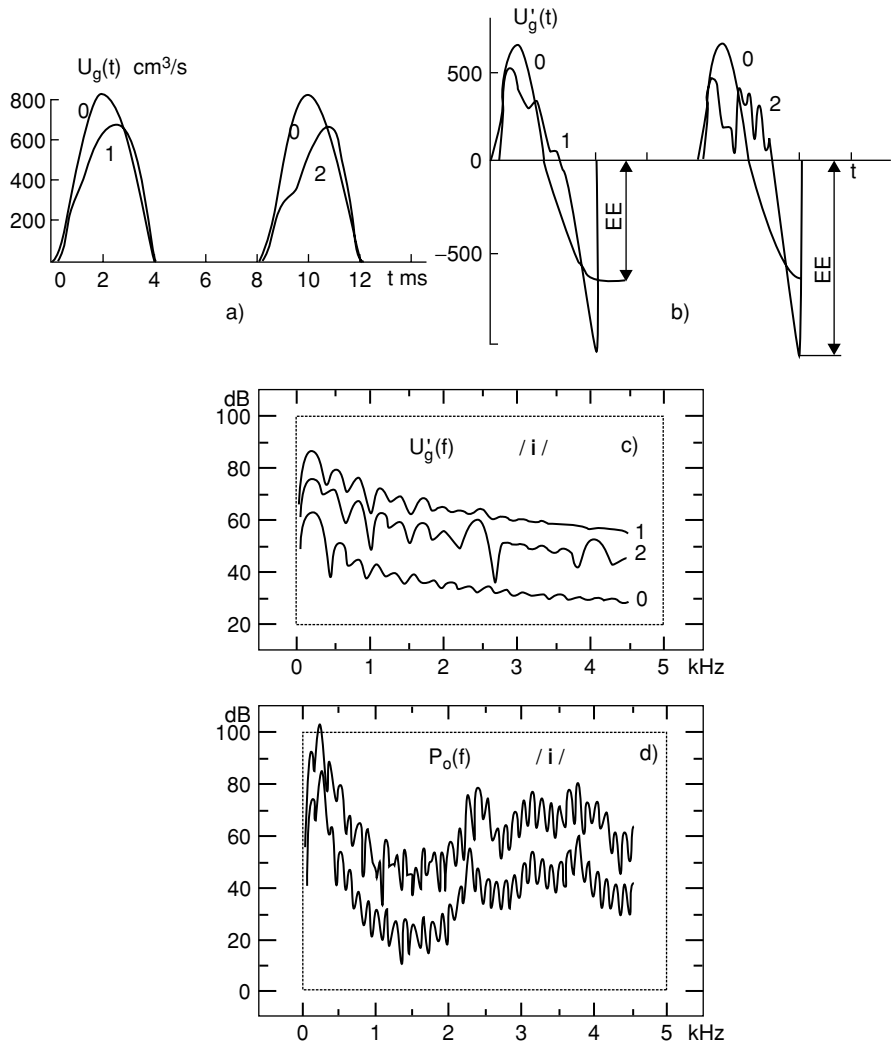


Figure 4.

a) The vowel /i/. A comparison between the glottal flow in the linear model, (curves labelled 0) and the interactive model, (curves labelled 1 and 2, the initial and second pulse, respectively).
b) Corresponding flow derivatives. The excitation level defined as the maximum negative peak at closure, EE, is about 4 dB higher in the interactive model than its counterpart in the linear source model.
c) The corresponding FFT spectrum of the curves in b). A couple of extra humps are clearly seen in the source spectrum due to the interaction effect.
d) FFT spectrum of the sound pressure within a frame of three complete periods. The interaction (top curve) causes a broadening of F1 and F2 and a small positive shift in the F2 frequency.

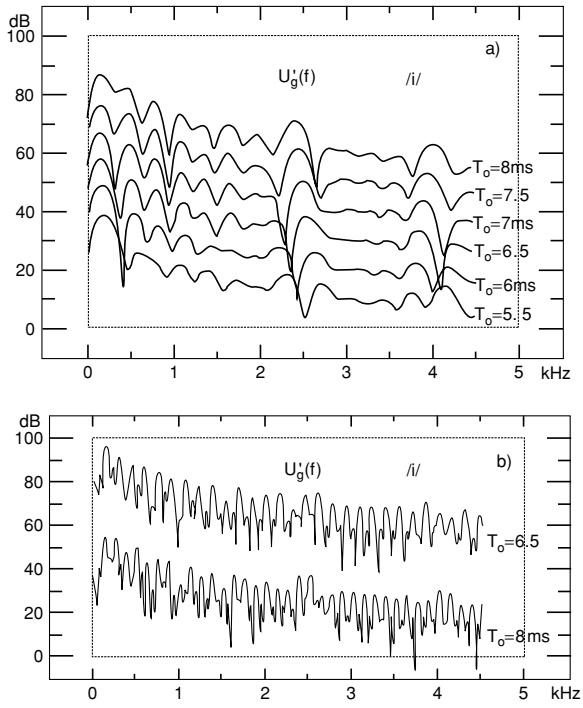


Figure 5.

a) The vowel /i/. Nonlinear superposition as a function of a varying duration of the closed phase. It should be noted that through $T_0 = 7$ ms, the notch changes its location relative to the extra peak in the vicinity of 2.2 kHz.

b) $U_g'(f)$, the Hamming window covering three voice cycles displays the harmonic structure. Top curve: $T_0 = 6.5$ ms; Bottom curve: $T_0 = 8$ ms.

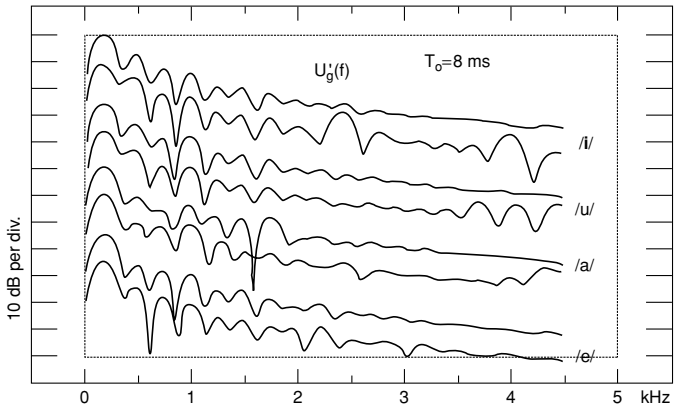


Figure 6. The first and second pulse source spectra of different vowels: /i/, /u/, /a/, and /e/. The curve shift between the vowels is 15 dB. The first and second pulses are separated 10 dB. A small inherent difference in excitation level exists. It increases within the series /e/, /u/, /i/ and finally to /a/ with a total span of 1.5 dB. It is observed that the source spectra are inter-periodically different for each vowel.

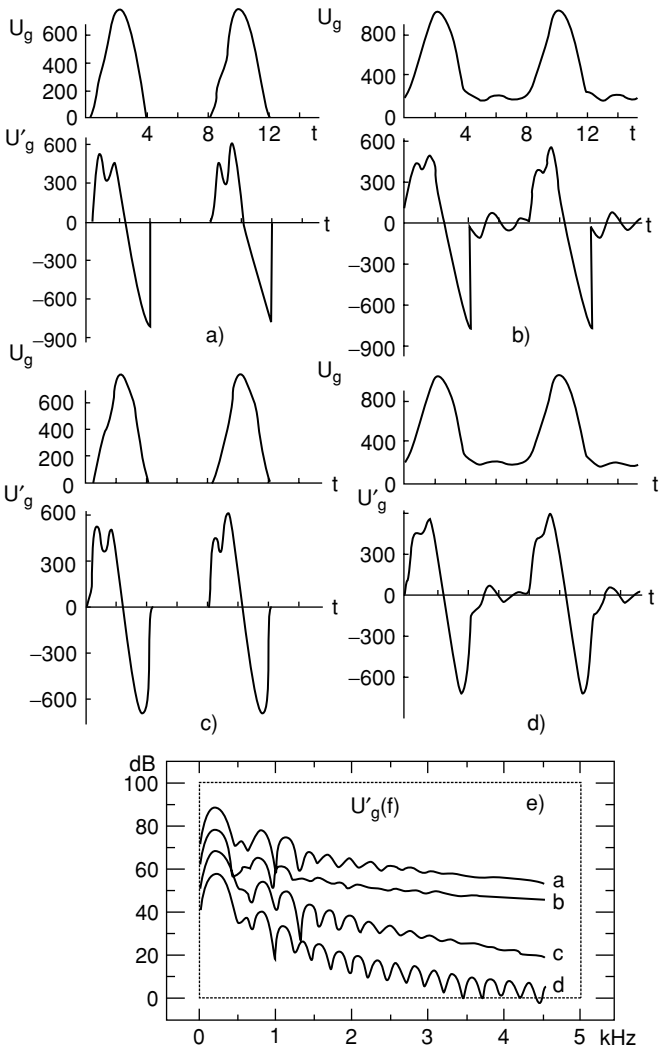


Figure 7A. Leakage simulation. The acoustic load is modelled by α single formant. F/B = 500/50, L = 5 mH.

a) No leakage. The glottal area function is a raised cosine; $T_o = 8$ ms, $T_e = 4$ ms, $T_p = 4$ ms, $A_{gmax} = 0.2$ cm^2.
b) A constant leakage (the leaky area is 25% of the glottal peak area) by adding the leaky area to the glottal function given in a). The load is the same as in a). Note the reduction of ripple components during the glottal open phase and the appearence of formant oscillation in the "closed" phase.
c) A dynamic leakage, the glottal area function is simulated by an LF model. $T_e = 4$ ms, $T_p = 2.2$ ms, $T_a = 0.15$ ms, Ee = 80. The maximum glottal area is approximately the same as in a).
d) The same as c) except that an addition of a constant leakage area is introduced to simulate both constant and dynamic leakage. Observe the remaining Fl ripples in the closed phase.
e) The corresponding source flow derivative spectra.

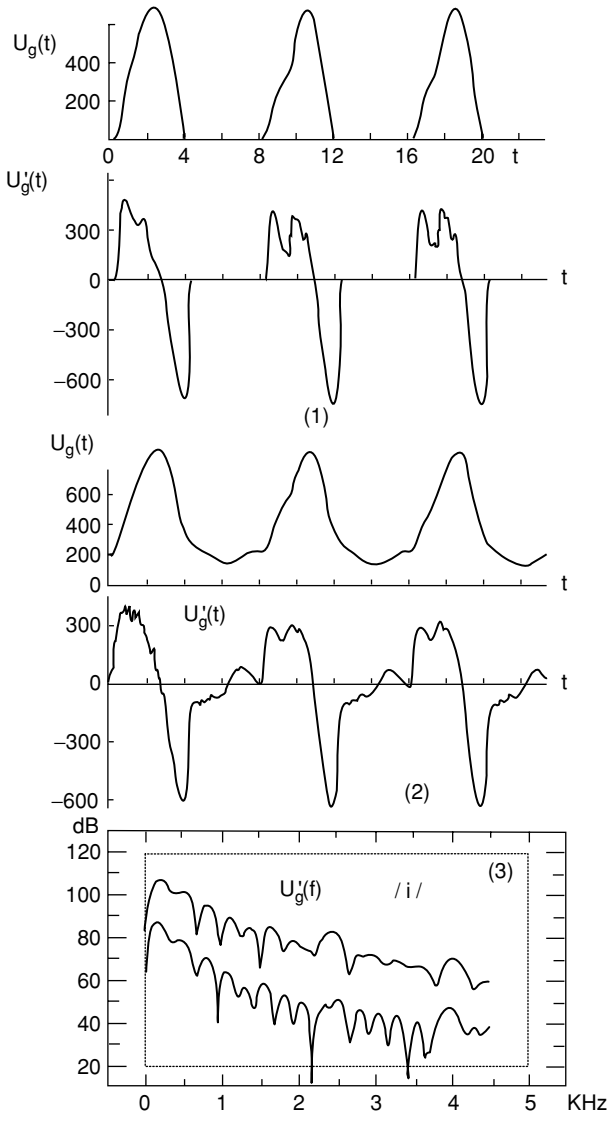


Figure 7B. Leakage simulation. The acoustic load is a compound vowel /i/.

1) Dynamic leakage only, A_g the same as in Fig. 7A:c.
2) Both dynamic and constant leak. Glottal area conditions are the same as in Fig. 7A:d.
3) The corresponding source spectra of 1) and 2). For the second pulse only.

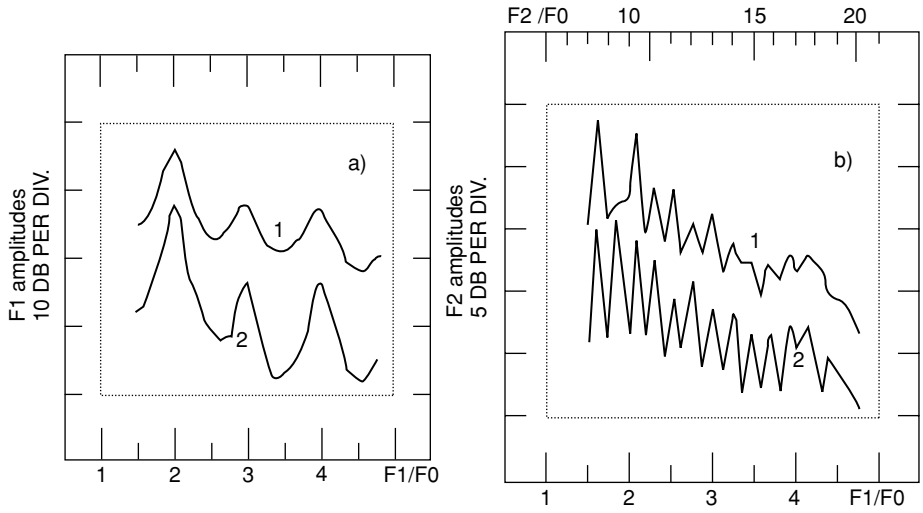


Figure 8.

a) Comparison of F1 amplitudes between 1: interactive and 2: linear models as a function of F1/F0.
b) Comparison of F2 amplitudes between 1: interactive and 2: linear models as a function of F1/F0 and F2/F0.

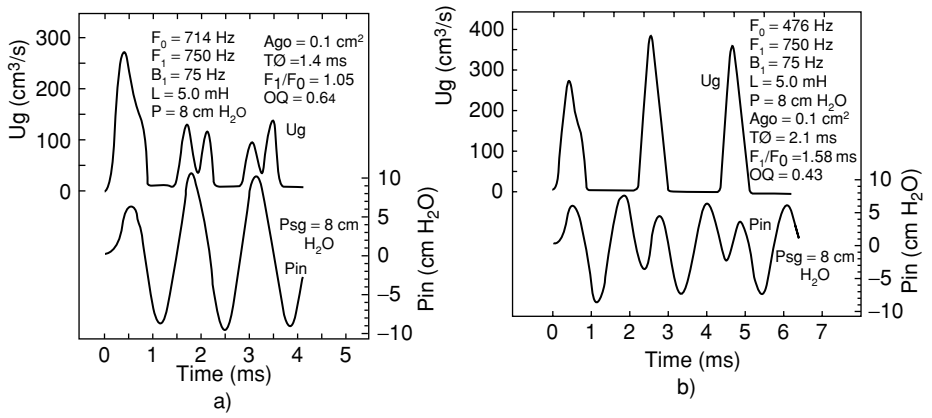


Figure 9. Glottal flow and supraglottal pressure assuming a one-formant loading.

a) F1 close to F0, F1/F0 = 1.05 simulating a soprano. This is an optimum condition for larger output and low air consumption.
b) Glottal flow and supraglottal pressure at a lower F0 = F1/1.58. Observe the larger air consumption (from Fant et al., 1985b).

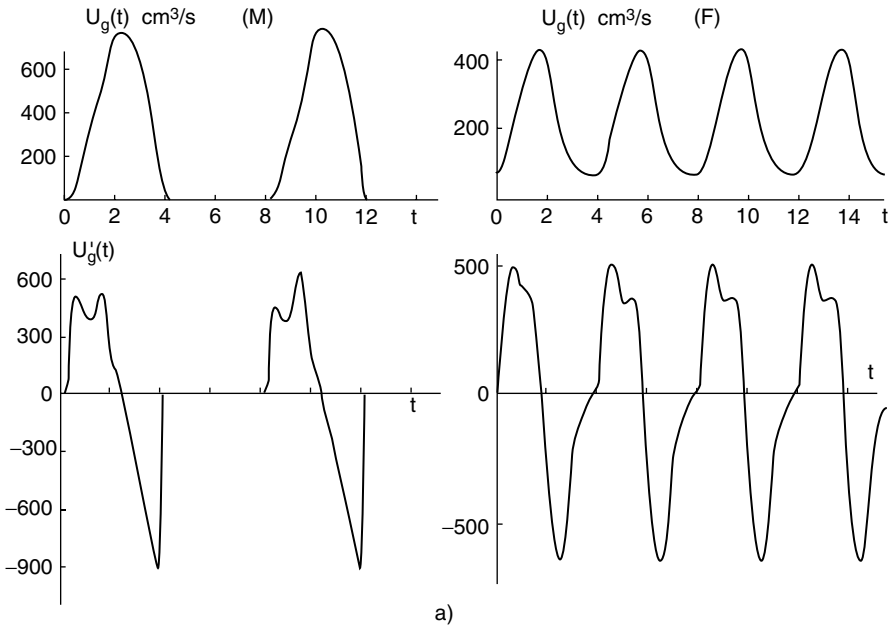


Figure 10. An attempt to simulate male and female phonation of a neutral vowel. Male: A_g is modelled by a raised cosine function, $T_0 = 8$ ms, open quotient is 50%, $T_p/T_e = 50\%$. $A_{gmax} = 0.2\,cm^2$. Female: A_g is simulated by an LF-model, $T_e = 2.8$ ms, $T_p = 1.5$ ms, $T_a = 0.2$ ms, $T_0 = 4$ ms. $A_{gmax} = 0.1\,cm^2$. This could represent a somewhat breathy phonation.
Formant frequencies and bandwidths of the female sample are adjusted to representative values.

a) Glottal flow and flow derivatives.
b) Glottal flow derivatives spectra, $U_g'(f)$, computed for the second pulse only.
c) The same as b) except that $U_g'(f)$ covers the first three periods which enhances the harmonic fine structure.
d) Sound pressure spectra. The length of the Hamming window is one and the same, 24 ms, for both sexes.

COMMENTS: The difference in source spectrum slope is primarily due to the "dynamic leakage", i.e., longer return phase of the female simulation whilst the added constant leakage has a smaller spectral effect, compare Fig. 7. Although female voices differ much, the prolonged return phase has been verified in human speech samples. The constant leakage mainly adds formant bandwidth increase, accounts for residual F1 ripple in the closed phase and adds somewhat to the spectral slope if combined with dynamic leakage.

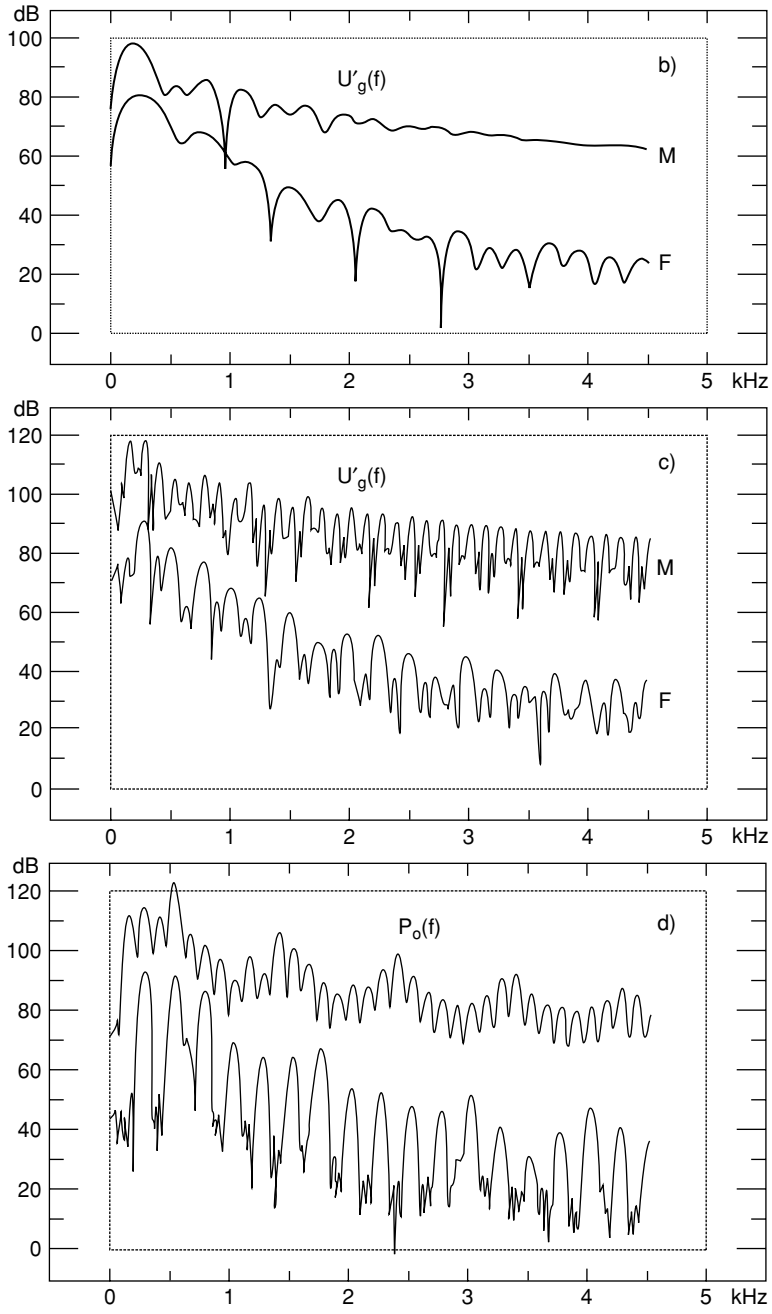


Figure 10. (Continued)

CHAPTER 3.3

FREQUENCY DOMAIN INTERPRETATION AND DERIVATION OF GLOTTAL FLOW PARAMETERS

ABSTRACT

Glottal flow parameters are generally defined as time-domain entities that specify the shape of glottal pulses or their derivatives. The present study is concerned with the relations of glottal parameters to frequency-domain properties in order to bring out perceptually important aspects. The analysis also aims at techniques to extract frequency- as well as time-domain parameters from frequency-domain representations. This involves frequency-domain inverse filtering, analytical transformations, and analysis-by-synthesis procedures. Frequency-domain processing is recommended as a complement to or a substitute to conventional time-domain analysis. An advantage is the less severe demands on low-frequency recording fidelity. Moreover, already available narrow-band spectral sections may be processed in order to derive major voice source parameters. The frequency-domain matching ensures optimal-conditions for Hi-Fi resynthesis. The theoretical analysis also sheds light on time-domain processing techniques suitable to support frequency-domain processing, e.g., selective inverse filtering. The frequency-domain analysis includes studies of covarying formant bandwidths and subglottal coupling effects which become especially apparent in breathy voicing.

INTRODUCTION

The demands on high-quality synthesis have provoked a renewed interest in voice source analysis and modelling. The model proposed by Fant (1979) and the later LF-model (Fant, Liljencrants, & Lin, 1985) have been exploited by several research groups. At KTH, the LF-model has recently been applied to studies of female speech (Karlsson, 1988) and to studies of the temporal variation of voice source parameters in connected speech Gobl, 1988).

In the course of our work, we have found it necessary to pay a close attention to the complete source-filter analysis and to the final demands of a maximally accurate resynthesis. It is ultimately the combination of source and filter function which is decisive. Since the process of separating source and filter functions is seldom unambiguous, there often remains an uncertainty of how to treat the trading relations; what shall be attributed to the source and what shall be attributed to the filter function (Gobl, 1988). The choice of source parameters thus imposes important constraints on the realization of a corresponding vocal tract filter function and vice versa.

Frequency-domain matching of original and synthesis ensures a proper base for correcting for the difference between the inverse filtering transfer function and a specific synthesizer transfer function. A time-domain derivation of glottal parameters does not necessarily ensure a proper frequency-domain fit. To pay attention to the frequency-domain consequences of a certain time-domain decision, e.g., of the important return phase parameter T_a, is a recommended technique in our present routines. A final check of the spectral match is also needed. A major motivation for in part incorporating a frequency-domain processing in the voice source analysis is

to optimize the choice of T_a. There is also an apparent interest in studying how far one can get with frequency-domain processing alone.

THE LF-MODEL

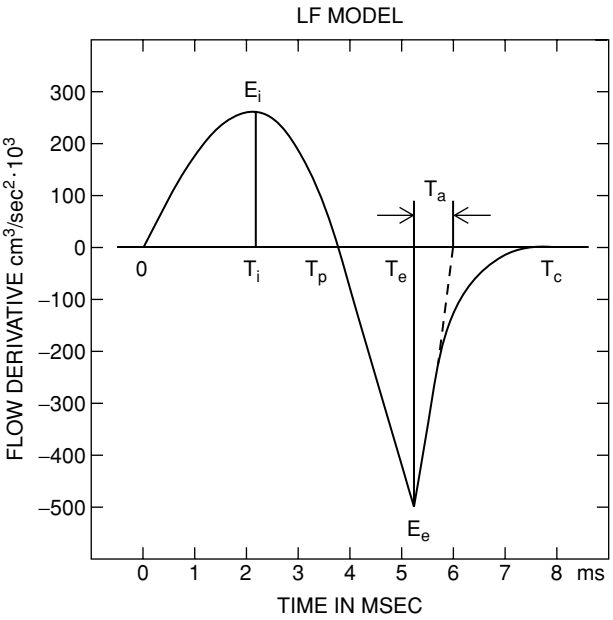
Any linear source-filter model of speech production is an approximation only. Even though the vocal tract transfer from glottal flow to radiated pressure in all essentials is a linear process, there remains a highly nonlinear transformation from a time-varying glottal area function $A_g(t)$ to a corresponding glottal flow $U_g(t)$. The origin of this nonlinearity is the flow-dependent glottal impedence, the instantaneous value of which is determined by trans-glottal pressure fluctuations within a pitch period including formant frequency oscilla-tions. Detailed studies of this acoustic interaction are presented in Ananthapadmanabha & Fant (1982), Fant & Ananthapadmanabha (1982), Fant & Lin (1987), and Fant, Lin, & Gobl (1985). The spectral consequences of the interaction, pulse skewing, truncations, and pulse ripple appear as increased formant excitation levels and bandwidths and local spectral distortions such as multiple zeros which cause spectral dips and the tendency of spectral energy being dispersed towards the high-frequency side of a formant peak region. However, perceptual tests performed by Nord, Ananthapadmanabha, & Fant (1984) and, more recent, informal tests indicate that the acoustic interaction whilst adding somewhat to naturalness is not a decisive quality factor.

A linear model should apparently be capable of generating representative overall glottal flow pulse shapes, ignoring ripple effects. The associated filter functions should be tailored towards representative effective bandwidths and formant frequencies that ensure a best overall match to natural speech.

The LF-model (Fant, Liljencrants, & Lin, 1985), see Fig. 1, is defined by the following glottal flow derivative wave shape:

$$\begin{aligned} &E(t) = E_0 e^{\alpha t} \sin \omega_g t \\ &(t < T_e) \\ &E(t) = -(E_e/\varepsilon T_a) \cdot [e^{-\varepsilon(t-T_e)} - e^{-\varepsilon(T_c-T_e)}] \\ &(T_e < t < T_c) \end{aligned} \tag{3.3.1}$$

It differs from the previous Fant (1979) model in three respects. First, the object of the LF-parameterization is to specify the glottal flow derivative whilst the flow has to be deduced from the integral of the LF time function. Secondly, and this is the main difference, there is included in the LF-model a return phase in the form of an exponential starting at the negative flow derivative peak at time T_e and connected to the onset of the next pulse at time T_0, which is standard practise, or to a fixed time T_c within the "closed phase". The parameter εT_a is a unique function of the other parameters. For small T_a, εT_a is close to 1. The effective duration of the return time T_a is defined by the projection on the time axis of its derivative at time T_e. It can be shown that the essential frequency-domain correspondence of the retum phase is a low-pass filter of the first order with cutoff frequency $F_a = 1/(2\pi T_a)$, see Fig. 2. This is the main parameter for change of spectral slope. A third feature is that



$$E(t) = E_0\, e^{\alpha t} \sin \omega_g t$$

$$(t < T_e)$$

$$E(t) = \frac{-E_e}{\varepsilon T_a} \cdot \left[e^{-\varepsilon (t - T_e)} - e^{-\varepsilon (T_c - T_e)} \right]$$

$$(T_e < t < T_c)$$

$$\omega_g = 2\pi F_g \qquad F_g = \frac{1}{2T_p} \qquad T_c = TO = \frac{1}{FO}$$

$$R_g = \frac{F_g}{F0} \qquad R_k = \frac{T_e}{T_p} - 1 \qquad R_a = \frac{T_a}{TO}$$

$$O_q = \frac{T_e + T_a}{TO} \qquad O_q^1 = \frac{T_e}{TO} \qquad F_a = \frac{1}{2\pi T_a}$$

Figure 1. The LF-model of differentiated glottal flow.

the main body of the LF-source function at $t < T_e$ is represented by a continuous sinusoid with a positive growth factor α, i.e., negative damping. There is accordingly no discontinuity at the flow peak as in the F-model (Fant, 1979) and the spectral slope is more continuous than in the F-model.

The LF-model is inherently a five-parameter model. There ex-ists a large variety of different sets of five parameters that u-niquely define the function. In addition to the direct synthesis parameters E_0, α, ω_g, T_a, and F_0, one can refer to the negative peak E_e and four critical time locations, T_p of max-imal flow, T_e of maximal

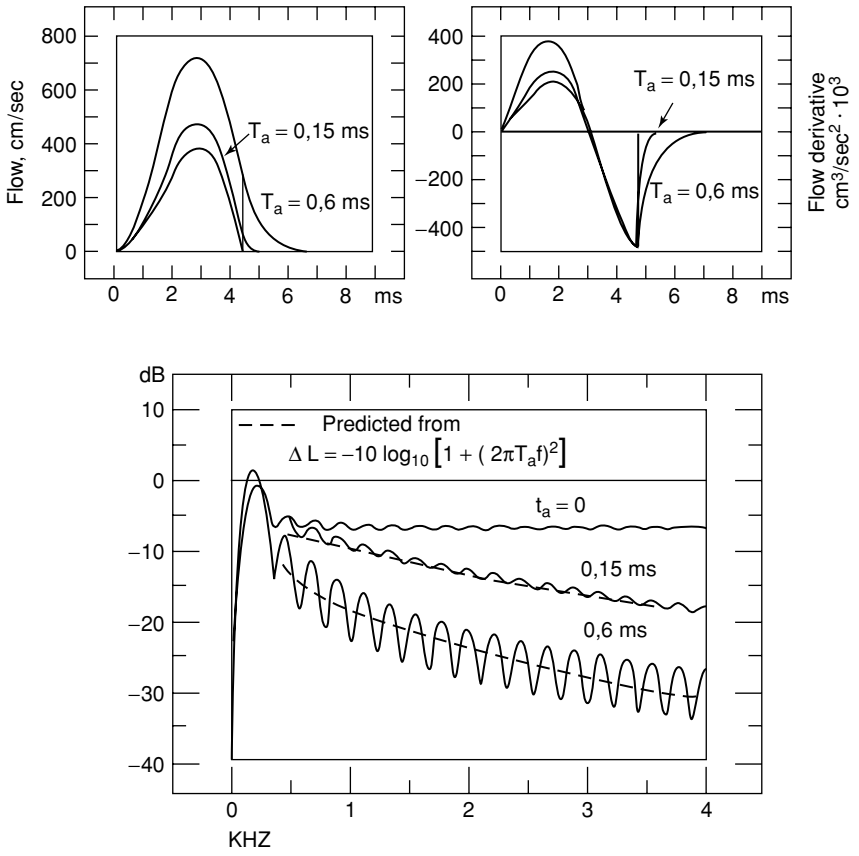


Figure 2. Wave shapes and spectral changes associated with an increase of the return time parameter T_a.

discontinuity in the flow derivative, the return time constant T_a, and the total period length T_0.

These are the set of parameters that usually evolve from inverse filtering. They directly relate to a set of normalized parameters, as indicated in Fig. 1. Thus, the glottal flow rise time T_p is converted to a "glottal frequency" $F_g = 1/2T_p$ which in turn may be related to F_0 as the quotient $R_g = F_g/F_0$. R_g is of the order of 1 and usually varies between 0.7 and 1.6.

A steepness factor is defined by how close T_e comes to T_p. This is expressed by $R_k = (T_e/T_p) - 1$: A dependent parameter is the open quotient which we may define either as $Q_0 = (T_e + T_a)/T_0$ or as $Q_0' = T_e/T_0$. In addition we can normalize T_a by the parameter $R_a = T_a/T_0$.

Much of our data collection (Gobl, 1988) has been concerned with the set of parameters E_e, R_k, R_g, R_a, and F_0. However, for developing source rules we are now looking into an alternative system that is closer related to visual aspects of the

glottal flow and which has a closer relation to general constraints of production and a closer tie to perceptual dimensions. Instead of R_k, we would thus prefer to refer to one of the alternatives E_e/E_i or to U_0/E_e or to its inverse E_e/U_0 where U_0 is the peak glottal flow. Instead of R_a, we find $F_a = 1/2\pi T_a$ to better suite frequency-domain matching aspects.

The spectral consequences of a systematic variation of LF-parameters are brought out in Fig. 3 in which F_g is held constant and in Fig. 4 which illustrates covariation of F_g and R_k whilst both U_0 and E_e are constants. In both figures $T_a = 0$.

Fig. 3 illustrates the constancy of the spectral slope, -6 dB/oct, of the differentiated flow spectrum, whilst the spectrum level increases with E_e/E_i (or inversely with the underlying R_k). After a second differentiation we are in a better position to discuss the balance between a low frequency level in the vicinity of $F_g = 125$ Hz and the spectrum level at higher frequencies. At constant U_0, the spectrum level above $2F_g$ increases about 10 dB when E_e/E_i varies from 1.5 to 4 whilst the level at F_g is constant.

It follows that at constant E_e and varying glottal flow peak U_0, the spectral level varies over the same 10 dB range in the F_g area and is constant at higher frequencies. Thus, U_0 determines the low-frequency level and E_e the high-frequency spectral level. With $R_g = 1$, the level of the source fundamental equals that of upper source harmonics providing E_e/E_i is close to 3 or more precisely, $E_e/E_i = \pi$, as will be derived in Eq. (14). The second harmonic here is a few dB above the fundamental. If we choose a lower F_0 than F_g, i.e., a greater R_g, the second harmonic would be even more dominating over the fundamental. This is typical of a low-frequency pressed voice.

The spectral consequences of these variations can be conceived of as a very selective reinforcement/reduction of low-frequency energy. The perceptual effects are not dramatic.

The specific variations brought out in Fig. 4 are even less apparent. Here, since both U_0 and E_e are fixed, the remaining variation in F_g provides a shift in the frequency range of the second and third harmonic only. In an [a]-vowel context, the effect is barely audible. A much more dramatic effect is the change of T_a as in Fig. 2. In a breathy voice, e.g., in a voiced [h] or in a vowel assimilating glottal abduction, we have measured F_a as low as 100 Hz, which represents a severe low-pass filtering. For average non-breathy female vowels, F_a is in the range of 500–1500 Hz and for males 1000–3000 Hz.

TIME FREQUENCY-DOMAIN RELATIONS

We shall now take a more general view of time frequency-domain relations in voice production. The first step will be to quantify a more rigid relation between peak glottal flow U_0 and the amplitude of the voice fundamental A_0 in a harmonic representation and then proceed to the relation of the excitation parameter E_e to various aspects of formant amplitudes and source harmonics. Consider a source-filter radiation representation

$$P(s) = G(s) \cdot H(s) \cdot R(s) \tag{3.3.2}$$

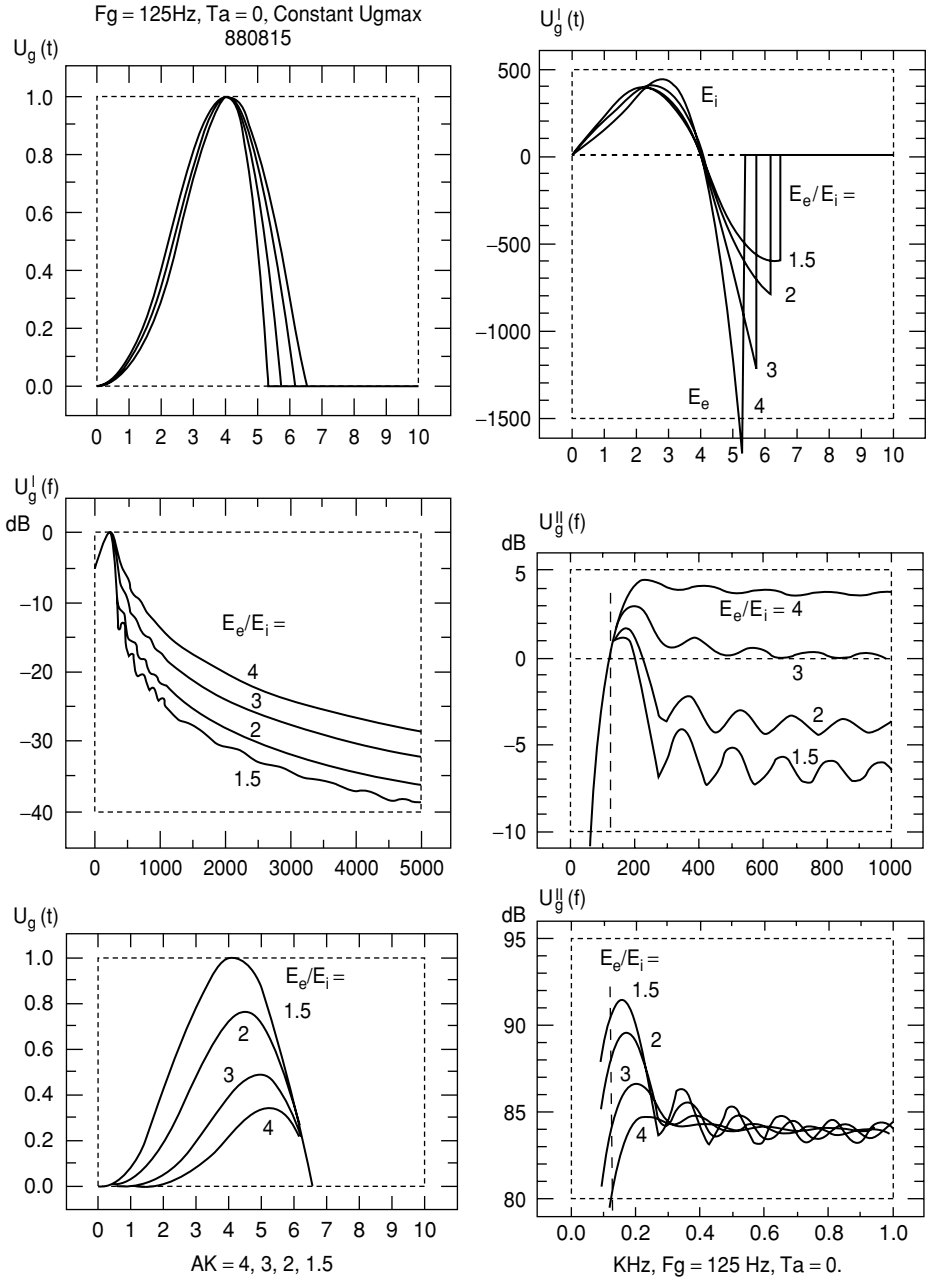


Figure 3. In the top two rows: LF-flow, flow derivative, flow derivative spectrum, and the second derivative spectrum at varying E_e/E_i and constant U_0. In the bottom row: LF-flow and its second derivative spectrum when maintaining constant E_e.

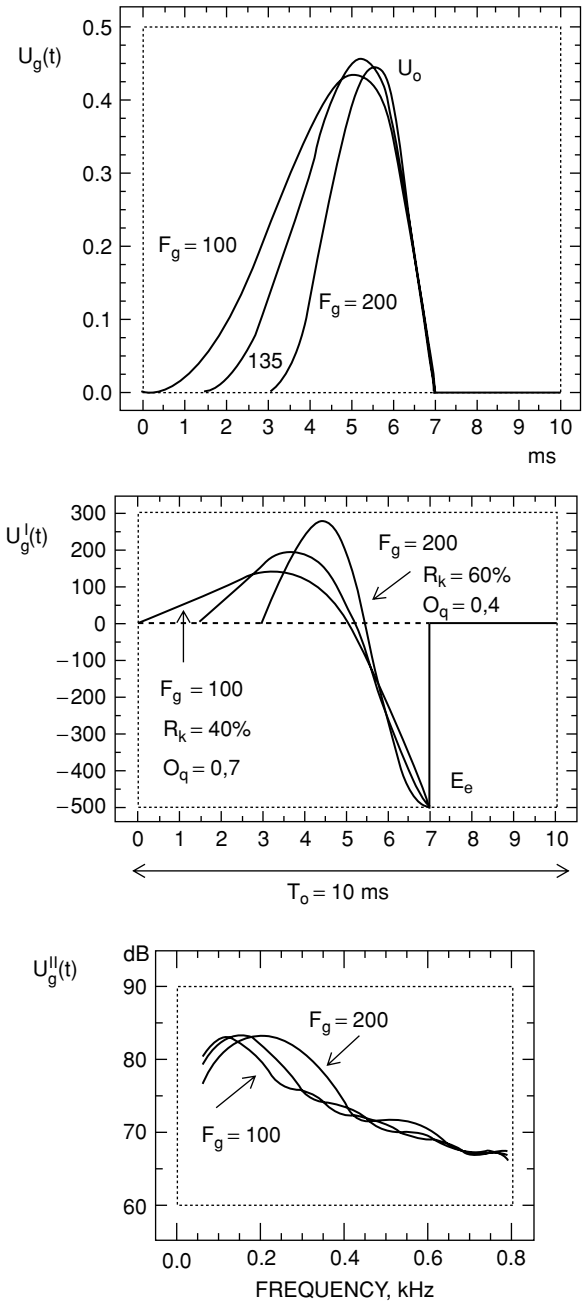


Figure 4. Flow, flow derivative, and second derivative spectrum when maintaining both U_0 and E_e constant and varying R_g (F_g) and dependently R_k.

with absolute values

$$|P(\omega)| = |G(\omega)| \cdot |H(\omega)| \cdot |R(\omega)| \tag{3.3.3}$$

the radiation transfer is

$$|R(\omega)| = (\rho\omega/4\pi a) \cdot k_T(\omega) \tag{3.3.4}$$

see Fant (1979). Here ρ is density of air $1.14 \cdot 10^{-3}$ g/cm^3 and a is the lip-microphone distance in cm. The correction factor $k_T(\omega)$ represents the combined baffle effect of the head and the increase of radiation resistance in excess of ω^2. It is usually neglected in production theory but accounts for a total high-frequency emphasis of about 5 dB from 300 to 4000 Hz. It could be included in a very detailed modelling (Fant, 1979). In the following we shall set $k_T(\omega) = 1$.

A basic step is to convert a glottal flow pulse train into a Fourier series. Given the flow peak amplitude U_0 and an open quotient, Q_0, close to 0.5, it can be shown that the flow source fundamental comes close to $U_0/2$. This is exactly so for two conditions. One is to model the glottal flow as a rectified sine wave, omitting the negative parts. The other is to model the glottal flow as a continuous raised sinusoidal wave with the "closed phase" undefined. With an open quotient of 0.4, the half sine wave model produces a 1 dB lower fundamental than $U_0/2$. For an open quotient of 0.3, the correction is -3 dB. On the other side we have a correction of approximately $+0.6$ dB for open quotients of the order of 0.6–0.8. Denoting the correction factor k and assuming a filter function $H(s) = 1$, i.e., eliminated by inverse filtering or being negligible in the F_0 domain. Eqs. (3) and (4) combined predict a fundamental amplitude of

$$A_0 = U_0 \cdot k \cdot \pi \cdot F_0 \cdot (\rho/4\pi a) \tag{3.3.5}$$

in the radiated wave at a distance of a cm.

For approximate calculations we may usually set $k = 1$. As also indicated in Fig. 5, the voice fundamental is, apart from the influence of the transfer function, proportional to the peak glottal flow and to the voice fundamental frequency.

The derivation of the relation of formant amplitude measures to E_e is more tricky. Assuming $T_a = 0$, i.e., the excitation E_e being a step function in glottal flow derivative or a ramp function in glottal flow, we end up with the following relation between E_e and the initial amplitude A_i of a damped oscillation from a single resonance vocal tract load.

$$A_i = E_e \cdot (\rho/4\pi a) \tag{3.3.6}$$

and

$$A_i/A_0 = (E_e/U_0) \cdot (1/k\pi F_0) \tag{3.3.7}$$

In Eq. (6), the frequency factor of the radiation transfer, Eq. (4), has been traded for the $1/\omega$ factor in the glottal flow derivative spectrum. As illustrated in Fig. 5, A_i is a time-domain entity which needs to be related to a frequency-domain correspondence. As a first step, we choose the peak amplitude of the corresponding resonance curve

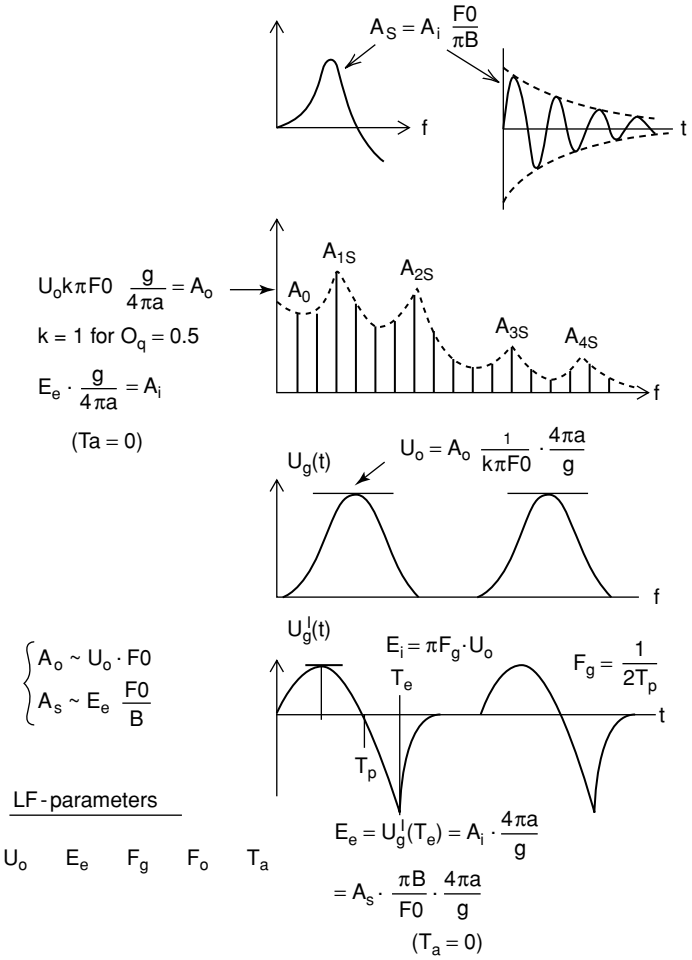


Figure 5. Some basic analytical relations between glottal flow wave shape and formant excitation and spectrum.

A_s in a harmonic spectrum. By appropriately handling transform equations for a conjugate complex pair of poles, we arrive at

$$A_s = A_i \cdot (F_0/\pi B_n) \tag{3.3.8}$$

Here, πB_n is the distance in the s-plane from the formant frequency $j\omega_n$ to the pole $\sigma_n + j\omega_n$ and F_0 enters to ensure a Fourier series property of A_s.

The next frequency-domain operation is to divide A_s by an estimated $Q = F_n/B_{ne}$. This inverse filtering is performed without an exact knowledge of the underlying bandwidth B_n. Add a differentiation in the spectrum with respect to F_0. The result is an "undressed" and differentiated harmonic source component:

$$A_k' = (F_n/F_0) \cdot (B_{ne}/F_n) \cdot A_s \tag{3.3.9}$$

which combined with Eq. (8) provides

$$A_i = \pi \cdot (B_n/B_{ne}) \cdot A'_k \tag{3.3.10}$$

In the following we shall drop the correction term B_n/B_{ne} and also assume that the inverse filtering has been made with a correct $F_{ne} = F_n$. Eq. (7) may now be rewritten:

$$E_e/U_0 = (A'_k(f)/A_0) \cdot \pi^2 \cdot k \cdot F_0 \tag{3.3.11}$$

$A'_k(f)/A_0$ is apparently the spectral level at a frequency f versus that of the fundamental in a +12 dB/oct compensated glottal flow spectrum. It is also assumed that the low-pass effects of a finite T_a have been compensated.

The general expressions for a harmonic A'_k in the inverse filtered and f/F_0 differentiated sound spectrum originating from a glottal flow ramp termination with the slope E_e is simply

$$A'_k = (2/T_0) \cdot (\omega/\omega_0) \cdot (E_e/\omega^2) \cdot (\rho\omega/4\pi a) = (E_e/\pi) \cdot (\rho/\pi a) \tag{3.3.12}$$

Note the similarity to Eq. (6). We thus note a simple time-frequency domain relation

$$A_i = \pi \cdot A'_k \tag{3.3.13}$$

Accordingly, Eq. (11) is valid not only for a hypothetical harmonic at a formant frequency but also for any source partial A'_k within the spectrum well above F_0. An estimate of E_e/U_0 from measured A'_k/A_0 can thus be performed from any harmonic in a range, say above $2F_g$, where $F_g = 1/2T_p$ as in the F- and LF-models and T_p is the glottal flow rise time. Dealing with harmonics outside formant peaks, we are no longer concerned with the error factor B_n/B_{ne} in Eq. (10). A number of independent estimates of E_e/U_0 may accordingly be made which also adds to certify proper bandwidth and formant frequency estimates.

Instead of U_0 we could refer to the maximum derivative E_i in the rising branch. Because of the quasi-sinusoidal shape of the flow derivative, we may with good accuracy express its integral up to T_p by

$$U_0 = (2/\pi)T_p \cdot E_i = E_i/\pi F_g = E_i/(\pi F_0 \cdot R_g) \tag{3.3.14}$$

(The error in U_0 is +2% at $E_e/E_i = 2$ and +11% at $E_e/E_i = 4$.) Eq. (11) may now be rewritten as

$$E_e/E_i = \pi \cdot (A'_k(f)/A_0) \cdot k/R_g \tag{3.3.15}$$

For the specific case of $A'_k = A_0$, which means that the voice fundamental amplitude is the same as that of higher harmonics in the differentiated flow derivative spectrum, we have $E_e/E_i = \pi \cdot k/R_g$ which agrees with the statement made in connection with Fig. 3.

However, as already stated, the derivation above assumes that $A'_k(f)$ values have been corrected for a finite T_a. Let A''_k denote the uncorrected A'_k.

$$A'_k = A''_k \cdot (1 + f^2/F_a^2)^{1/2} \tag{3.3.16}$$

F_a is selected so as to equalize the $A'_k(f)$ contour. From the selective inverse filtering, removing all formants but one, we are used to seeing that the damped oscillation

starts at an amplitude equal to that of the negative spike E_e of the flow derivative residue. Here we have a method for time-domain estimates of T_a from the relation of A_i to E_e which will be further discussed in connection with Fig. 13.

An important conclusion of this section is that a frequency-domain inverse filtering has the potential of deriving the main LF-parameters. For this purpose, it is convenient to choose the following set: U_0, E_0/U_0, R_g, F_a, F_0 which except for R_g now have been discussed R_g is estimated by matching in the domain of the first three harmonics.

It remains to extend this theory to other spectral representations, e.g., broad-band sections or band-pass filtered formant data.

APPLICATIONS TO FREQUENCY-TIME DOMAIN CONVERSION

The frequency-domain inverse filtering and differentiation, outlined in the previous section, amount to the following operations on a spectral level basis.

$$\begin{aligned} L'_k(f) &= L(f) + 20\log_{10}(f/F_o) \\ &\quad - 20\log_{10}[(K_{rr}\cdot(f)\cdot\prod_{n=1}^{r}\cdot|H_n(f)|)] \\ |H_n(f)| &= [(1-x_n^2)^2 + x_n^2/Q_n^2]^{-1/2} \\ x_n &= f/F_n \end{aligned} \tag{3.3.17}$$

where $L(f)$ is the input spectrum level in decibels with possible preemphasis removed and $K_{rr}(f)$ the correction for poles higher than no. r. For $r = 5$. we have

$$\begin{aligned} 20\log_{10} k_{rr} &= 0{,}433x_1^2 + 0{,}000712x_1^4 \quad \text{(dB)} \\ x_1 &= f/f_{ref} = c/4\ell \end{aligned} \tag{3.3.18}$$

The higher pole correction should be scaled to the particular vocal tract length, ℓ_t, Fant (1959; 1960), as derived, e.g., from F4. $\ell_e = (7/4)\cdot(c/F4)$.

We are now in a position to process any harmonic spectrum. We shall start with an attempt to reconstruct in absolute physical scales glottal flow parameters from a study of Swedish vowels undertaken at the Ericsson Telephone company in 1946–1947 reported by Fant (1948) and published by Fant (1959). This may seem a rather "archaeological" undertaking but is motivated by the fact that great care was taken in preserving absolute sound pressure values of spectrum components. However, this is one of the very few studies in the history of speech analysis in which both frequency and amplitude of formants and individual harmonics have been reported and it is the only one I know of with absolute calibration.

Seven males and seven females served as subjects phonating steady-state vowels in an anechoic chamber with a distance of 12.5 cm to a Brüel & Kjaer condenser microphone. Harmonic spectra were recorded on line by a sweep-frequency method. The subjects had to sustain the vowels for about 5 sec. Most of them were amateur singers.

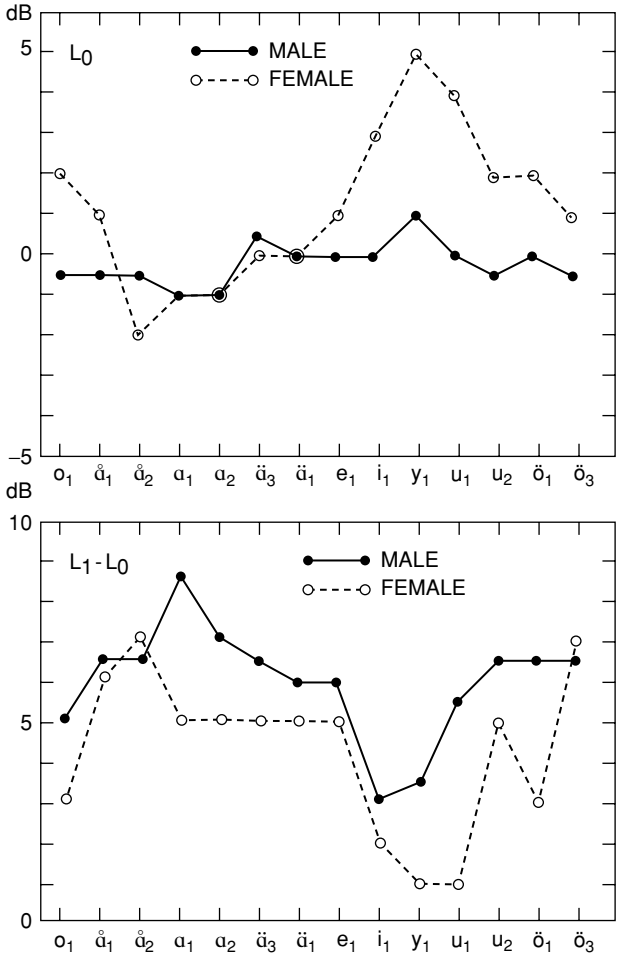


Figure 6. Original data from Fant (1959) on the spectrum level of the fundamental L_0 and its relation to the first formant spectrum level L_1-L_0 within a sequence of vowels.

Figure 6 shows the variation of the voice fundamental amplitude L_o within a sequence of vowels ordered in terms of increasing F_1 in back vowels followed by decreasing F_1 of front vowels and ending with an increasing F_1 series of rounded front and mid vowels. Subscript 1 stands for phonemically long vowels and subscript 2 for short vowels and 3 for long pre-r variants. The u_1 is a maximally rounded front vowel [ʉ:] and u_2 is a mid vowel [ɵ] (Fant, 1983).

Females tend to have higher L_o than men, especially in low F_1 vowels, which may be explained by the relative proximity of F_o to F_1 enhancing L_o of females. As seen in Fig. 7, the situation is reversed after undressing the transfer function, the males showing about 1 dB higher L_o than females. A typical male-female difference is the higher first formant amplitude L_1 versus L_o, as seen in the bottom part of Fig. 6.

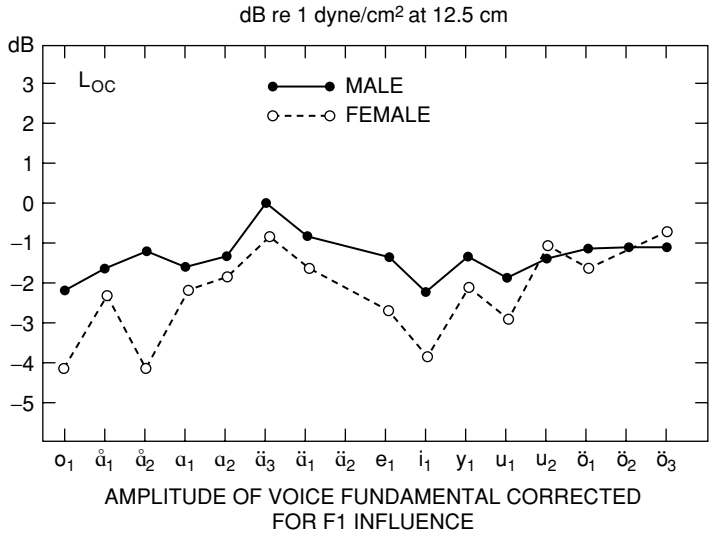


Figure 7. Male and female average data on the voice fundamental amplitude corrected for influence of F_1 and higher formants.

There are systematic trends in Fig. 7 that deserve some comments. The amplitude of the corrected source fundamental L_{oc} is somewhat lower in vowels with low F_1 compared to vowels of higher F_1. This is typically so for o_1 = [u:] and i_1 = [i:] which may be anticipated from production theory. Both the loss of transglottal driving pressure and the pulse skewing caused by a rather extreme vocal tract constriction would reduce the magnitude of glottal flow. This would also be expected for the vowel $å_2$ = [ɔ] which is produced with a narrow pharyngeal constriction.

All vowels were sustained at one and the same subject's preferred F_0 which averaged 125 Hz for males and 217 Hz for females. Inherent F_0-variations were thus eliminated and cannot explain the L_{oc} variations.

The L_{oc} of two reference subjects are shown in Fig. 8. Subject G was a well known Swedish phonetician Olof Gjerdman and subject M a well known Swedish soprano singer and voice specialist, Marianne Mörner. Here, we have evidence for a special aspect of acoustic-aerodynamic interaction modelled by Fant & et al. (1985) and initially observed by Rothenberg (1985).

When F_1 is very close to F_0, as in the subject M's vowel o_1 with $F_0 = 256$ Hz and $F_1 = 270$ Hz or in u_1 with $F_0 = 256$ Hz and $F_1 = 300$ Hz, there exists economy not only by optimal filtering but also a minimum of glottal flow since an F_1 oscillation component opposes the transglottal pressure drop and minimizes the air consumption whilst a relative high E_e excitation is retained. The large step in subject M's L_{oc} from o_1 to $å_1$ of 4 dB and from u_1 to u_2 of 5 dB is explained by the about 1.5 F_1/F_0 ratio in $å_1$ and u_2, which according to Fant & et al. (1985) conditions the opposite effect, i.e., an increased air consumption. Since this effect is much more pronounced in the female voices than in the male voices, it seems to be a plausible explanation in addition to variations of supraglottal constriction.

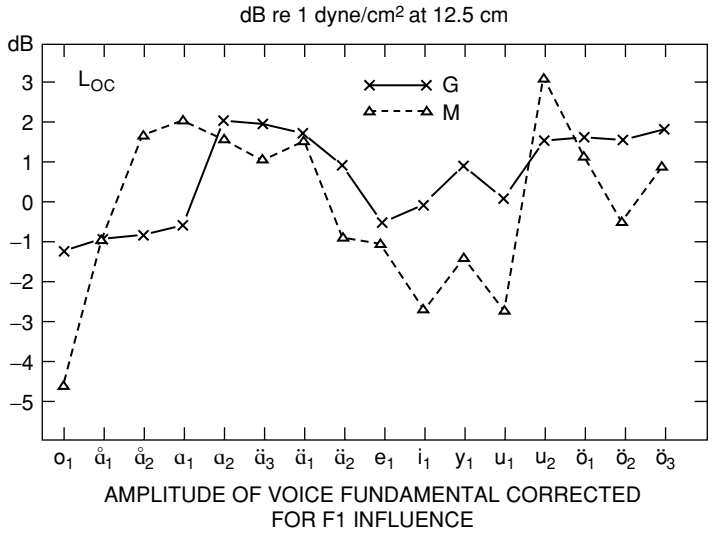


Figure 8. The same as Fig. 7 for a male subject G and a female subject M.

What about glottal flow parameters and absolute scale values? We selected the vowels $a_1 = [\alpha:]$, $a_2 = [a]$, $\ddot{a}_3 = [æ:]$, and $\ddot{a}_1 = [\varepsilon]$ and processed the tabulated F_0, L_0F_1, L_1F_2, L_2F_3, L_3, and F_4, L_4 data from Fant (1959) in accordance with the inverse filtering equations (17) and (18). Glottal peak flow U_0 was then calculated from Eq. (5) with $k = 1$ and the appropriate F_0. Next a representative value of F_a was estimated from Eq. (16) to account for the spectral drop off in regenerated source levels L'_k from F_1 over F_2 to F_3. After this spectral correction, a value of A'_k/A_0 could be estimated to be inserted in Eq. (11) to provide the E_e/U_0 estimate.

This procedure has one snag. The formant amplitude levels initially measured by Fant (1959) were attained by root mean square summations of partials within the formant domain. In order to translate these A_e measures to most likely spectrum envelope peak values A_s, we performed a correction by a factor

$$A_s/A_e = (1 - e^{-y})/(1 - e^{-2y})^{1/2} \tag{3.3.19}$$

where $y = \pi B_n/F_0$, derived by Fant (1959). We found $A_s/A_e = 0.83$ for male average and 0.72 for female average.

Average glottal parameter data for the four vowels have been documented in Table 1 together with data from other surveys of interest. Because of lack of data on the second harmonic, we could not estimate R_g.

There are remarkable similarities across studies. None of the studies, except possibly Holmberg, Hillman, & Perkell (1988), reveals a significant female/male difference in U_0/E_e and the same is true of the related parameter R_k. The U_0/E_e is of the order of 1.1 ms for the Fant (1959) and the Holmberg & et al. (1988) data and of the order of 0.7 ms in the Fant, Gobl, & Karlsson (1987) data. The higher U_0/E_e in the first two studies implies according to Eq. (11) a greater dominance of the voice fundamental amplitude. This is plausible considering the sustained almost

TABLE 1
Glottal flow data

	U_0 cm^3/sec	E_e $10^3\ cm^3/sec^2$	U_0/E_e ms	E_e/E_i	R_g %	R_k Hz	F_a
1. Fant (1959), Present study							
Males, $F_0 = 125$	420	360	1.15				2100
Females, $F_0 = 217$	210	190	1.10				1200
2. Holmberg et al. (1988)							
Males, $F_0 = 116$	230	240	0.95				
Females, $F_0 = 213$	140	120	1.15				
3. Fant et al. (1987)							
Males, $F_0 = 125$			0.65	3.2	1.1	30	1000
Females, $F_0 = 217$ (from 1).			0.70	2.2	0.9	30	430
4. F-domain analysis, Present study							
JS (male), $F_0 = 133$			0.65	3	1.2	28	3000
MS (female), $F_0 = 188$			0.70	2.4	1	30	2000

singing mode in the Fant (1959) phonations which would enhance the fundamental (Sundberg & Gauffin, 1979). In the Holmberg & et al. study (1988), the relative high U_0/E_e might be explained as an abduction assimilation from the unvoiced [p] to the following vowel [æ] in their test words or else originating from a low-pass smoothing of glottal flow termination inherent in the mask technique.

Why do we in spite of the relative constant U_0/E_e observe a typically higher level of the voice fundamental amplitude versus the level of the first formant in females compared to men, see Fig. 10, pertaining to a vowel [ɑ:]? For subject JS, we may note $L_1 - L_0 = 14$ dB and for subject MS, $L_1 - L_0 = 7$ dB. The answer is that

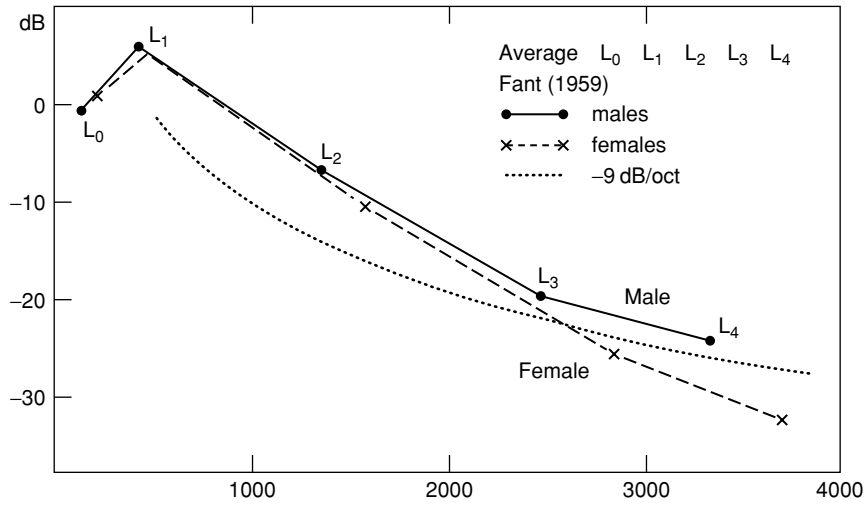


Figure 9. Fant (1959) vowel data averaged over all vowels.

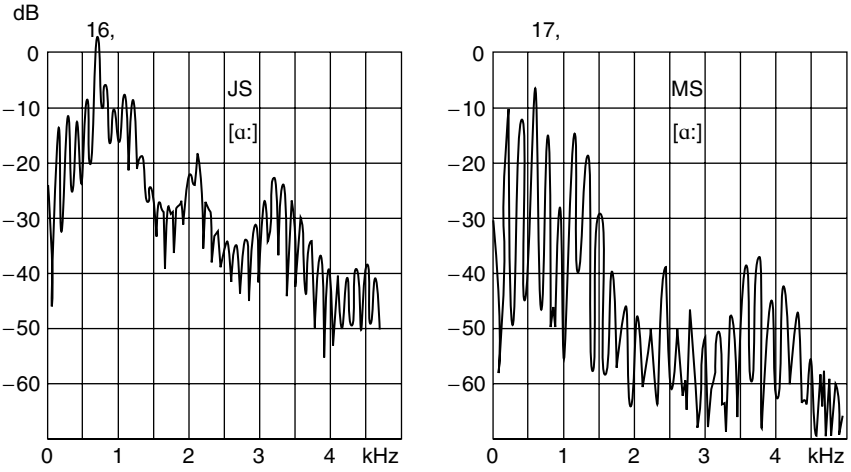


Figure 10. Spectra of the vowel [ɑ:] from the word "ja" produced by a male JS and a female MS.

the F_0 factor in Eq. (11) contributes with 5 dB and that the difference in the L_0 reinforcement from F_1 and higher formants of the vowel [ɑ:] is 2 dB in favor of the female voice. Instead of referring to Eq. (11) we may refer to Eq. (15) in which E_i/E_e takes the place of U_0/E_e and the R_g factor enters. However, it seems to be more direct to refer to the 5 dB difference in F_0 entering Eq. (11) than to refer to a 3 dB difference in E_i/E_e and a 2 dB difference in R_g in Eq. (14).

The absolute magnitude of the peak flow derived from the spectrum data of Fant (1959) is reasonable. The 0.42 liters/sec noted for males and 0.21 liters/sec for females may be compared to the Holmberg & et al. (1988) 0.23 liters/sec for males and 0.14 for females. However, the experimental conditions were quite different and one may note that the latter study reported an average of 0.1 liters/sec additional steady air leakage. Our values range well within typical data reported in the literature, e.g., Rothenberg (1973); Sundberg & Gauffin (1979).

Both U_0 and E_e are typically 5 dB higher for males than for females. This constant E_e/U_0 ratio does not imply that the shape of the female glottal flow pulse is a proportional down scaling of the male pulse. Proportionality would have implied linear scaling in both amplitude and time, and that the amplitude reduction would be the same as the reduction of the time base in terms of $T_p = 1/2F_g$. Under such conditions, the female E_e would equal that of the male E_e, and the E_e/U_0 would have increased by the inverse of the time scale proportionality factor. In reality, the 3 dB higher F_g of females is compensated by a 3 dB lowering of E_e at constant U_0 associated with the wave form change induced by a relatively longer return phase.

This ratio is approximately the gain of the F_a filter, Eq. (16), at the frequency of the formant. An example is brought out in Fig. 13 which refers to an [h]-vowel sequence (Fant, 1980). We may here follow the temporal variation of F_a from 90 Hz in the [h] to 1200 Hz in the following vowel. The associated bandwidth as determined from envelope fitting varies in this context from something larger than 300 Hz to 52 Hz in the vowel. Such bandwidth changes are important to preserve in the

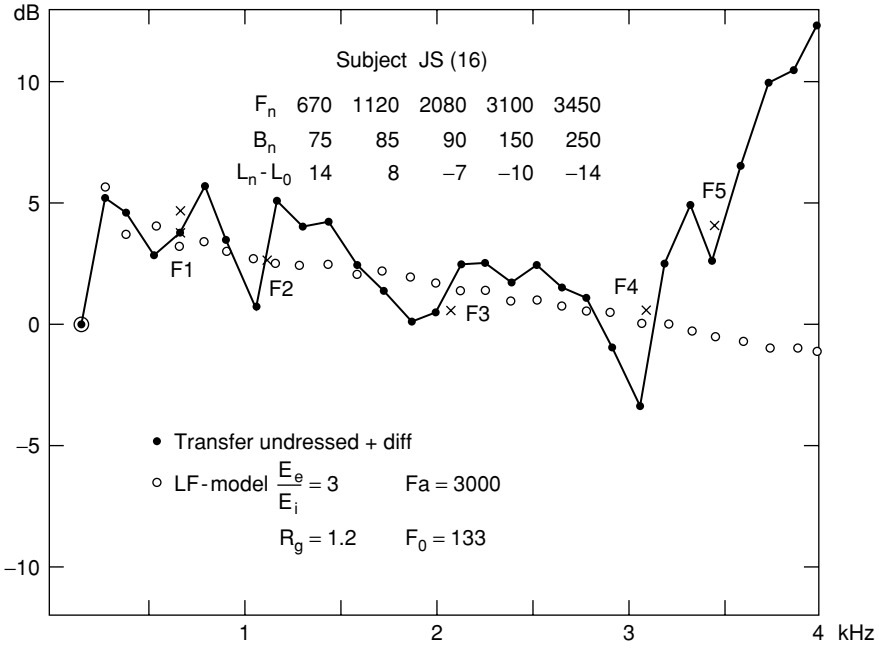


Figure 11. Frequency domain inverse filtering and differentiation of the male [ɑ:] vowel from Fig. 10. This technique is especially useful for determining F_a.

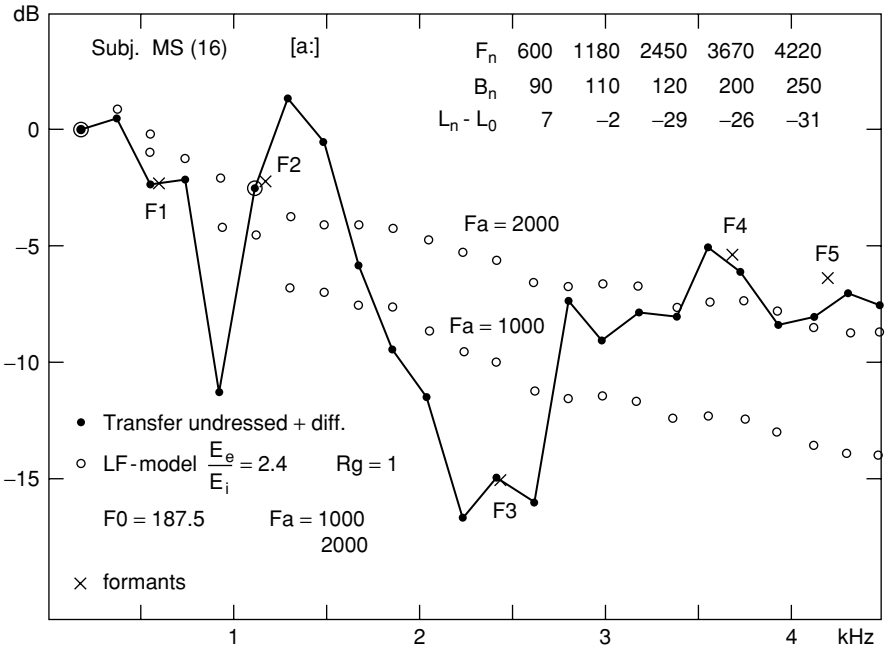


Figure 12. The same for the female subject MS. Two different values of F_a are suggested.

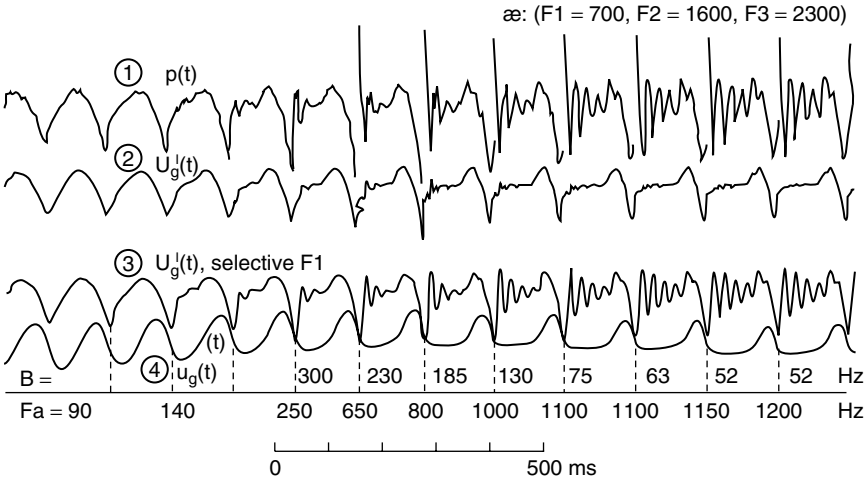


Figure 13. Selective F_1 inverse filtering; curve 2 from the bottom, provides means of determining both B_1 and the spectrum parameter F_a.

resynthesis. A quantitative analysis of glottal damping (Fant, 1960; Badin & Fant, 1984) can be extended by specific rules for specific vowel categories. Because of the glottal inductance and reactive components of the rest of the vocal tract, the glottaly induced bandwidth increase is minimized for frequencies above 1000–1500 Hz. With this in mind, the data of Fant (1960, p. 136), here labelled B_{ref}, may be generalized for any particular glottal flow by the relation

$$\Delta B(t) = B_{ref} \cdot U_g(t)/55 \tag{3.3.20}$$

where $U_g(t)$ is the glottal flow in cm^3/sec, see also Fant (1979).

One aspect of increasing glottal abduction or leakage is that the subglottal system no longer can be neglected in vowel production. A system function distortion enters. It can be specified by the appearance of extra poles and zeros and some finite detuning of poles of the original uncoupled stage. The effect is exemplified in Fig. 14 which pertains to a simulation of the Russian vowels [e] and [a] with the subglottal system of Badin & Fant (1984). The transfer function was obtained with a pressure source introduced in the glottis. The coupled state corresponds to a glottal opening of 0.2 cm^2 and a lung pressure of 2 cm H_2O, while the uncoupled state was simulated with a very high glottal impedance. The [e] vowel displays an extra peak at 1500 Hz, which can be identified with the 1400 Hz extra formant in the female [ɛ] spectrogram at the right in Fig. 14. More extra peaks appear in the [a] transfer function. These effects are very much the same as measured by Fujimura & Lindquist-Gauffin (1971) on a live subject with the sweep-frequency method, see also Fant, Ishizaka, Lindquist-Gauffin, & Sundberg, 1971). It is important to note that overall shifts in spectrum levels occur. The relative attenuation of F_2 of [e] could shed light on the relative weak F_3 of subject MS in Fig. 12.

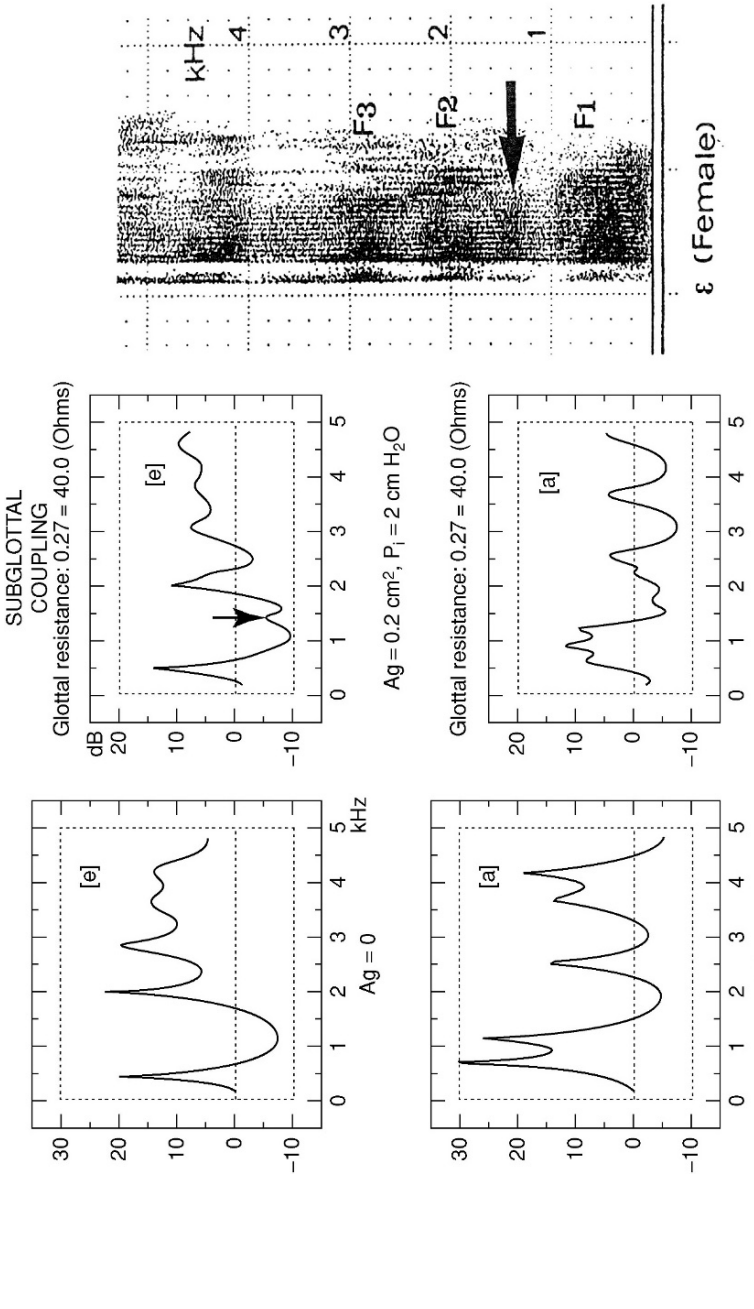


Figure 14. Finite and large subglottal coupling creates spectral distortion of vowels. Additional formants and changes in overall spectrum are apparent. Note for comparison the extra formant in the spectrogram of the female post [d] and pre [t] breathy vowel [ɛ].

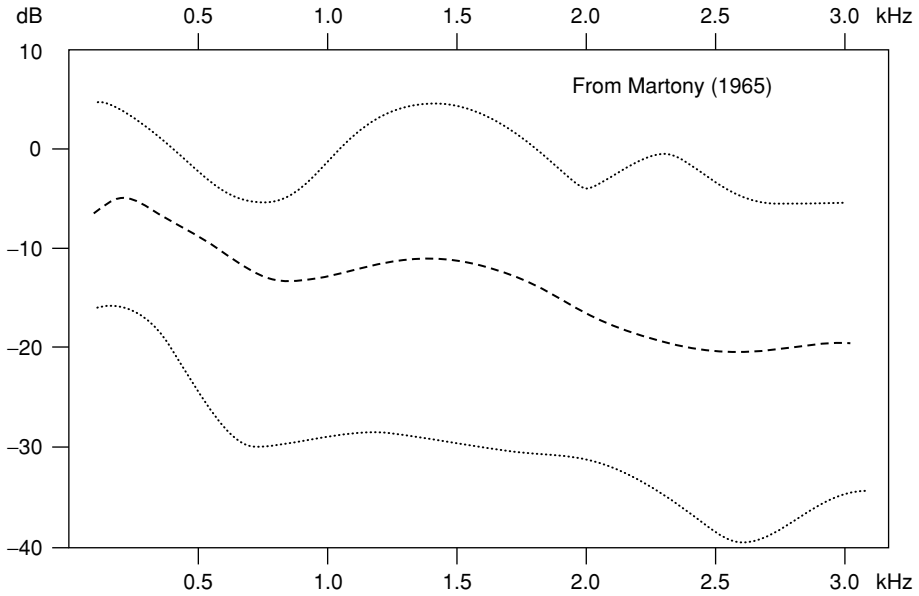


Figure 15. Extreme and mean voice source spectra, from Mártony (1965).

The underlying cause appears to be a zero associated with the third subglottal resonance which appears in this region. We have definite examples of relative prominent extra subglottal formants occasionally seen in spectrograms of female voices. The weak spectral peak at 2800 Hz is the MS [a] spectrum, Fig. 10, could derive from subglottal coupling. However, extra peaks of a few dB are also a common effect of acoustic interaction (Fant & Lin, 1987). The main indication of a subglottal coupling is thus the relatively low F_3 compared to the rest of the spectrum.

We shall end this survey by reference to the frequency-domain inverse filtering performed by Mártony (1965). Fig. 15 shows the mean and extremes of his subjects' source spectra. Before we have more evidence from analysis and modelling, we should execute some caution when attempting to explain the details. The main point is that a typical voice source spectrum may deviate systematically from a uniformaly sloping line.

CONCLUSIONS

We have here touched upon some of the potentialities and difficulties in inverse filtering and voice source parameterization. A frequency domain approach as we have outlined avoids the need of high fidelity low-frequency phase correct recordings and allows a direct processing of harmonic spectra. The technique could be extended to broad-band spectral analysis. The basic advantage of the frequency-domain processin is a closer tie with resynthesis needs. However, it remains to perform a more systematic analysis of the unavoidable differences between true human speech with all interaction effects and linear synthesis.

A non-interactive source-filter system based on best spectral match with the LF-model will probably stand up to rather high quality standards.

Gunnar Fant & Qiguang Lin

REFERENCES

Ananthapadmanabha, T.V. & Fant, G. (1982): "Calculation of True Glottal Flow and its Components". Speech Communication 1, pp. 167–184.

Badin, P. & Fant, G. (1984): "Notes on Vocal Tract Computation", STL-QPSR 2–3/1984, pp. 53–108.

Carré, R. (1987): "Review of French Work on Vocal Source—Vocal Tract Interactions, pp. 371–375 in *Proc. XI ICPhs, Vol. 3*, Academy of Sciences of the Estonian SSR, Tallinn.

Fant, G. (1948): "Analys av de svenska vokalljuden", L M Ericsson protokoll H/P 1035.

Fant, G. (1959): "Acoustic Analysis and Synthesis of Speech with Applications to Swedish", Ericsson Technics No. 1.

Fant, G. (1960): *Acoustic Theory of Speech Production*, Mouton, The Hague.

Fant, G. (1979): "Glottal Source and Excitation Analysis", STL-QPSR 1/1979, pp. 85–107.

Fant, G. (1980): "Voice Source Dynamics", STL-QPSR 2–3/1980, pp. 17–37.

Fant, G. (1982): "Preliminaries to Analysis of the Human Voice Source", STL-QPSR 4/1982, pp. 1–27.

Fant, G. (1983): "Feature Analysis of Swedish Vowels—A Revisit", STL-QPSR 2–3/1983, pp. 1–19.

Fant, G. & Ananthapadmanabha, T.V. (1982): "Truncation and Superposition", STL-QPSR 2–3/1982, pp. 1–17.

Fant, G. & Lin, Q. (1987): Glottal Source—Vocal Tract Acoustic Interaction", STL-QPSR 1/1987, pp. 13–27.

Fant, G., Gobl, C., & Karlsson, I. (1987): "The Female Voice—Experiments and Overviews", J. Acoust. Soc. Am. 82, p. S92(A).

Fant, G., Ishizaka, K., Lindquist-Gauffin, J., & Sundberg, J. (1972): "Subglottal Formants", STL-QPSR 1/1972, pp. 1–12.

Fant, G., Liljencrants, J., & Lin, Q. (1985): "A Four-parameter Model of Glottal Flow", STL-QPSR 4/1985, pp. 1–13.

Fant, G., Lin, Q. & Gobl, C. (1985): "Notes on Glottal Flow Interaction", STL-QPSR 2–3/1985, pp. 21–45.

Fujimura, O. & Lindquist-Gauffin, J. (1971): "Sweep-tone Measurements of Vocal-tract Characteristics", J. Acoust. Soc. Am. 49, pp. 541–558.

Gobl, C. (1988): "Voice Source Dynamics in Connected Speech", STL-QPSR 1/1988, pp. 123–159.

Holmberg, E.B., Hillman, R.E., & Perkell, J.S. (1988): "Glottal Air Flow and Transglottal Pressure Measurements for Male and Female Speakers in Soft, Normal, and Loud Voice", J. Acoust. Soc. Am. 84, pp. 511–529.

Karlsson, I. (1988): "Glottal Wave Form Parameters for Different Speaker Types", pp. 225–231 in *Proc. of SPEECH 88*, 7th FASE Symp., Edinburgh.

Mártony, J. (1965): "Studies of the Voice Source", STL-QPSR 1/1965, pp. 4–9.

Monsen, R.B. & Engebretson, A.M. (1977): "Study of Variations in the Male and Female Glottal Wave", J. Acoust. Soc. Am. 62, pp. 981–993.

Nord, L., Ananthapadmanabha, T.V. & Fant, G. (1984): "Signal Analysis and Perceptual Tests of Vowel Responses with an Interactive Source Filter Model", STL-QPSR 2–3/1984, pp. 25–52.

Rothenberg, M. (1973): "A New Inverse Filtering Technique for Deriving the Glottal Air Flow Wave Form During Voicing", J. Acoust. Soc. Am. 53, pp. 1632–1645.

Rothenberg, M. (1985): "Cosi Fan Tutte and What It Means". draft for discussion. Fourth Int. Vocal Fold Physiology Conf., New Haven, CT.

Sundberg, J. & Gauffin, J. (1979): "Wave Form and Spectrum of the Glottal Voice Source", pp. 301–322 in (B. Lindblom & S. Öhman, eds.) *Frontiers of Speech Communication*, Academic Press, London.

CHAPTER 4

SPEECH ANALYSIS AND FEATURES

The first article (Fant, 1962) is an early attempt to provide a structured interpretation of speech spectrograms in terms of acoustic criteria for manner and place of articulation. These still remain a general source of information for phonetic classification. The presentation of discrete segments is supplemented by a graph of connected speech as a succession of overlapping articulatory events, which underlies principles of coarticulation, and supports the general rule that a single phoneme may be related to several successive acoustic segments, and conversely, that properties of a single acoustic segment may be influenced by several successive phonemes intended by the speaker. These many-to-one relations are by now well established in phonetics.

The second article (Fant, 1997) was written for a handbook on Acoustics. The main emphasis is on speech analysis and processing techniques. An aspect sparsely covered in the literature but outlined here, is how to optimize time averaging of sampled data for spectrum analysis of unvoiced sounds in order to maintain a sufficient temporal and spectral resolution at a tolerable level of superimposed random noise. In addition, the article outlines the source-filter aspect of speech production and exemplifies inverse filtering and speech processing for prosodic studies.

The third article (Fant, 1986) was written for an MIT symposium on Invariance and Variability of Speech Processes. The title "Features—fiction and facts" reflects some scepticism to current attitudes, which tend to involve abstractions with too little support of data. I contributed with an overview of distinctive feature analysis of Swedish, in more detail treated in Fant (1975). Another question brought out in the symposium was to what extent feature recognition is guided by auditory or speech motor functions. This became one of the main issues in discussions of my paper which are retained in the article together with references. These are of a historical interest, reflecting the engagement of phonetics, psychology and neurophysiology at that time. A prominent example is Alvin Liberman's motor theory of speech perception.

A main point at the symposium was whether there exists an absolute invariance. My position is in the negative. Even if we can define procedures for feature recognition in specified context, we have to accept the principle of relational invariance as coined by Roman Jakobson (Jakobson, Fant and Halle, 1952). The pragmatic consequence is that in order to approach the nucleus of the speech code we need to structure the variability. Discussions pro and contra absolute invariance have lost their significance.

The topic of the speech code is outlined in more detail in the next article, number 4. The general notion of a code implies relations between message units and signal units. In speech communication we are concerned with the relation of language units to units and properties of the speech act. These can be defined on all parts of the speech chain, from production to perception and with maximum detail retained in

the speech wave. However, in a spectrogram the information bearing elements are often hidden in the detail structure, and thus difficult to decipher without resort to an articulatory interpretation.

The relative success of speech synthesis has created an illusion that we have a profound insight in the speech code. The illusion becomes especially apparent when we try to operate in the reverse direction, that is given a record of the speech wave, attempting to decipher what was said.

Advanced techniques of automatic speech recognition require a more extensive insight in the speech code than what is available today. An apparent difficulty is the complex encoding of prosodic elements, and in general the great range of variability with respect to speaker type and speaking style. However, even our present insights are difficult to handle. The documentation is fractionalized. We lack concise summaries and there is a knowledge gap between phonetics and technology to consider.

Individual variations with respect to distinct and reduced articulation are exemplified. In connected speech a voice stop may degenerate to a voiced continuant, and a nasal consonant may be realized without oral closure, i.e. by nasalization only. Temporal patterns of fully developed voiced and unvoiced stops are shown in one of the figures.

SELECTED ARTICLES

[4.1] Fant, G. (1962). Descriptive analysis of the acoustic aspects of speech. *LOGOS,* Vol **5**, No. 1, 3–17.

[4.2] Fant, G (1997). Acoustical Analysis of Speech. In M.J. Crocker (ed.) *Encyclopedia of Acoustics,* John Wiley, Vol**. 4**, 1589–1597.

[4.3] Fant, G. (1986). Features—fiction and facts. In J. Perkell and D. Klatt (eds.) *Invariance and Variability of Speech Processes*, Lawrence Erlbaum Ass. Publ.1986, 481–491.

[4.4] Fant, G. (2001). On the Speech Code. *TMH-QPSR* 2–3 2001, 61–67. (Revised and updated version of an article, The Speech Code, in C. von Euler, I. Lundberg and G. Lennerstrand (eds.) *Brain and Reading*. MacMillan, London, 1982, 171–182.

ADDITIONAL READING

Jakobson, R., Fant, G. and Halle, M. (1952). *Preliminaries to speech analysis. The distinctive features and their correlates.* Acoustics Laboratory, Massachusetts Inst. of Technology, Technical Report No. 13 (58 pages). Published by MIT press, seventh edition, 1967.

Fant, G. (1973). *Speech Sounds and Features.* The MIT Press. Cambridge, MA, USA, (contains a selected number of articles).

Fant, G. (1991). What can basic research contribute to speech synthesis? *Journal of Phonetics*, **19**, 75–90.

CHAPTER 4.1

DESCRIPTIVE ANALYSIS OF THE ACOUSTIC ASPECTS OF SPEECH[1],[2]

SPEECH RESEARCH OBJECTIVES

The scientific study of speech is at present in a transitional stage of development. The classical articulatory or rather physiological phonetics dealing mainly with a description of the speech mechanism and with articulatory correlates of phonetic symbols is still the basic source of knowledge in phonetics courses at linguistic faculties, although acoustic phonetics is gaining ground. Acoustic phonetics, dealing with the structure of speech as sound waves and the relations of this structure to any other aspects of the speech communication act, does not lack traditions either. This field is developing rapidly as a result of the last few years' intensified investments in speech research from communication engineering quarters.

One aim of the technical speech research is to lay a foundation for techniques of producing artificial speech and of machine identification of spoken words. Applications such as more efficient speech communication systems, book-reading aids for the blind, and means of visual and tactile recording of speech for communication with the deaf, as well as specific voice controlled automata, are within reach of present technology or may be expected to be so in a not too distant future.

The perfect speech typewriter, representing the engineering criterion of a profound knowledge of the acoustic nature of speech and of dialectal, individual, and contextual variations, is a more distant object—it is rather a symbol of combined efforts in speech research. This profound knowledge does not exist yet.

SPEECH ANALYSIS IN THEORY AND PRACTICE

The techniques of synthesizing speech are already quite advanced. Analysis techniques have not been developed to the same extent, and this is especially true of the analysis directed towards teaching a machine to recognize spoken items. This is not due to lack of research efforts. On the contrary, there is a considerable amount of work undertaken on the use of large digital computers for machine identification of speech, but this work is still in an initial instrumental phase of methodological studies[1]. Phoneme recognizing machines of a simpler analog type have been constructed but their performance has not been very advanced. The possible vocabulary or phoneme inventory has been restricted, and the machines have not responded very well to any one else than "his master's voice"[2,11].

What we really lack is a descriptive study of the visible sound patterns of speech providing an acoustic mapping of the spectrographic correlates to phonetic signs and categories with due regard to particular language, dialectal, individual, and contextual variations. A speech researcher may be well acquainted with the art of synthesizing speech by general rules but the same man is probably not

able to decipher the text of a spectrographic record of which he has no a priori information[3].

The difficulties may in part be due to technical shortcomings of commercially available spectrographs, but there are other reasons, such as the lack of a rationale for going through the necessary learning process. Small deviations of the visual pattern may be highly significant for phonemic discriminations whereas quite apparent pattern features may be primarily related to accidental voice characteristics of the speaker. A spectrogram provides an overdetailed reference for the formal contents of a speech message. A basic problem in speech analysis is to formulate the complex transforms whereby the phonetically significant aspects may be extracted from the mass of data available. The pioneering work on the establishment of the metalanguage of Visible Speech is that of Potter, Kopp, and Green[31]. This is a valuable reference but it has shortcomings such as the restriction of the frequency range to that of telephony, i.e., to approximately 3200 c/s upper limit.

The lack of quantitative data on acoustic correlates of phonetic units is especially great for consonants and the more extensive vowel studies available refer to single stressed test words[2] or to sustained forms[8]. What about the distinctive feature approach by Jakobson, Fant and Halle[18]? Is it not possible to learn to read Visible Speech simply by reference to a maximum of 12 distinctive pattern aspects within any sound? The answer is no. Not without the addition of a considerable amount of linguistically redundant information. The particular choice of features[4] is supported by the main systematizing principles of classical phonetics. The distinctive features[7,18,19] are described in terms of the articulatory and the corresponding acoustic and perceptual correlates of linguistically relevant spectrographic studies of speech.

The limitations of the preliminary study of Jakobson, Fant and Halle[18] are that the formulations are made for the benefit of linguistic theory rather than for engineering or phonetic applications. Statements of the acoustic correlates to distinctive features have been condensed to an extent where they retain merely a generalized abstraction insufficient as a basis for the quantitative operations needed for practical applications. It should also be remembered that most of the features are relational in character and thus imply comparisons rather than absolute identifications. The absolute references vary with the speaker, his dialect, the context, the stress-pattern, etc., according to normalization principles which have not been fully investigated.

It should be noted that a specification of speech wave data may be translated to any of an infinite number of alternative forms, each based on a different choice of variables. This is true of instrumental techniques as well as of the technical and conceptual operations performed on the raw material from analysis. A linguist may radically change a specificational system in order to gain a small saving in specificational costs. The minimum redundancy of the system becomes the holy principle and a purpose in itself. The engineer is more interested in the application of the system and will generally accept some redundancy in order to facilitate automatic recognition procedures or to clarify the nature of a distinction. However, in several respects the linguistic and engineering systems should be identical. The more rigidly and unambiguously a linguistic distinction can be correlated to quantitative speech wave data, the more useful it will be for engineering applications. Investigations into the

quantitative aspects of formulating distinctive features are much needed. Some experimental work in this direction has been undertaken by Halle and associates[12,15,17].

Speech synthesis is an important tool for testing the relative importance of various aspects of the sound patterns contributing to a distinction. Valuable empirical information on these "cues" and on the general rules for synthesizing sppech stems from the well-known work at the Haskins Laboratories[4,24,25]. Similar work is now also under progress at various other places[5].

One of the achievements of acoustic speech research is the study of the analytical ties between the physiological and the acoustic aspects[7,33,34]. Given the evidence of the dimensions of the vocal cavities, it is possible to calculate the essentials of the spectral properties of the corresponding speech sound. There is also a reverse predictability, though to a lesser extent due to the fact that compensatory forms of articulation can provide rather similar speech wave patterns.

The rules relating speech waves to speech production are in general complex since one articulatory parameter, e.g., tongue height, affects several of the parameters of the spectrogram. Conversely, each of the parameters of the spectrogram is generally influenced by several articulatory variables. However, to establish and learn these analytical ties is by no means a hopeless undertaking. Some elementary knowledge in acoustics is valuable, but the main requirement is a sound knowledge of articulatory phonetics.

TRANSCRIPTION OF SPEECH SPECTROGRAMS

A common observation when spectrograms of ordinary *connected* speech are studied is that modifications and omissions of speech sounds are frequent. Carefully pronounced single testwords and phrases may differ considerably from ordinary speech. These effects may cause transcription difficulties. Shall the investigator transcribe the spectrograms according to the phonemic structure or shall he, according to phonetical principles, write the phonetic symbols of what he hears? A third possibility might be to infer from the spectrogram how the speech has been produced and adapt the transcription thereafter. The latter method is quite feasible in view of the apparent articulatory significance of phonetic symbols, but the technique will have to rely on the use of an extended set of phonetic signs just as the phonetic transcription by ear utilizes a greater inventory of signs than the phonemic transcription. The choice of system, phonemic, perceptual, or articualtory, is primarily a matter of the purpose of the investigation. The articulatory transcription is a powerful method of checking the perceptual transcription and can be utilized once the investigator has become sufficiently accustomed to reading spectrographic patterns.

Fig. 1 exemplifies Visible Speech spectrograms produced with a Sona-Graph analyzer. The text was "Santa Claus" spoken by an American subject[6]. On the top of the figure there appears the normal broad-band (300 c/s) and the narrowband (45 c/s) spectrogram. The middle and the bottom spectrograms were made after a speed reduction by a factor of 2 which implies an effective doubling of the filter bandwidths, i.e., 600 c/s and 90 c/s respectively. Normally a broad-band spectrogram shows the formant structure whereas the narrow-band analysis displays a harmonic spectrum. In case of high-pitched voices, however, the 600 c/s-analysis is needed in

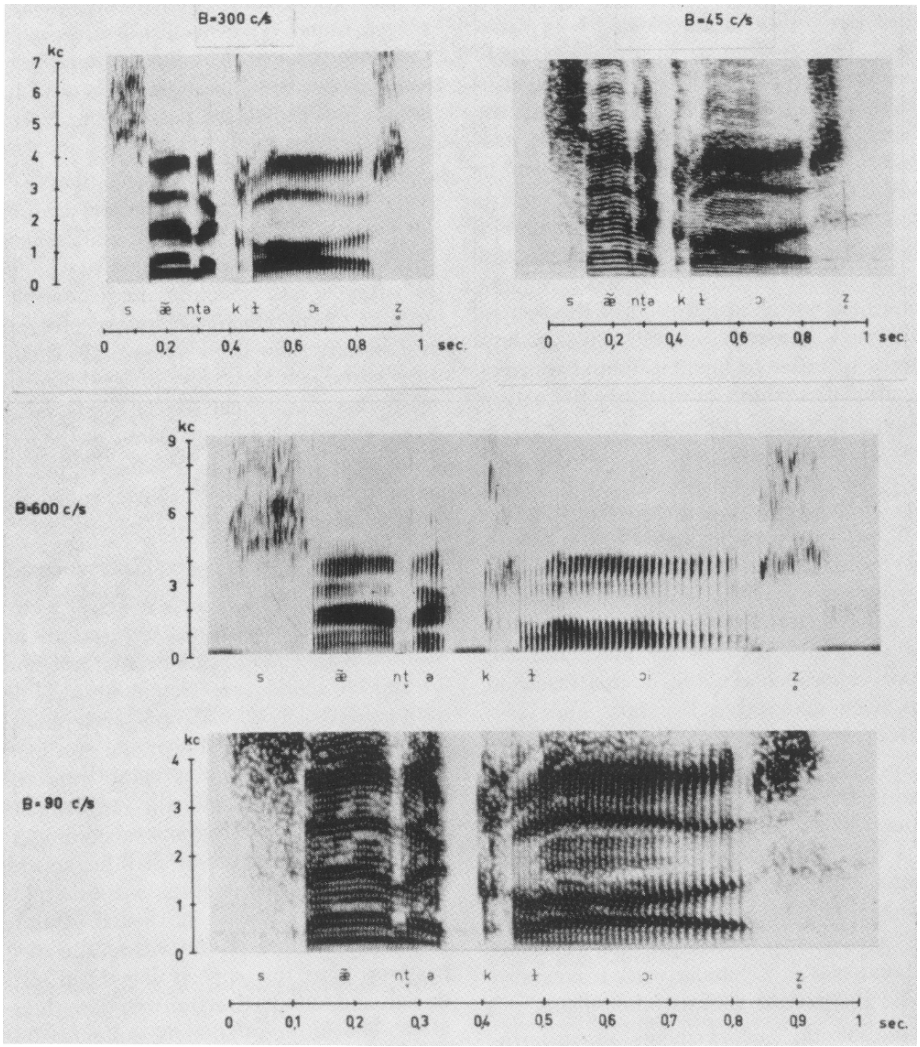


Figure 1. Spectrograms illustrating the effects of varying the bandwidth of the spectrum analyzer. In "narrow"-band (B = 45 c/s) analysis (upper right) the harmonics are resolved in the vowel [æ̃]. In the "broad"-band (B = 300 c/s) analysis (upper left) the formants are resolved. When the speech material is played into the analyzer at half speed, the time-scale is stretched by a factor of 2 and all frequencies are divided by the same factor. The apparent bandwidths of the analyzer then becomes B = 90 c/s and B = 600 c/s respectively. All spectrograms pertain to one and the same utterance, "Santa Claus".

order to avoid harmonic analysis and retain the formant structure. At a low-pitched interval of speech, on the other hand, the 90 c/s-filter provides an optimal frequency resolution of the formant structure.

The auditive transcription of the utterance was [sæ̃ntə klɔ:z]. A segmentation of the spectrum in terms of successive sound intervals, or in other words sound segments,

should be performed from the broad-band display and not from the narrow-band display since the latter will tend to smooth out rapid shifts of the spectral composition. The very distinct boundary between the [s] and the [æ̃] in the form of a shift of the spectral energy distribution and the shift from a voiced to an unvoiced sound is typical. Most of the other sound boundaries are also distinct.

As seen from the split first formant the speaker has apparently nasalized the entire [æ̃] in anticipation of the /n/ and there is no separate [t]-segment except for a weak high frequency burst in the latter part of the [n]-segment. Alternatively, it might be argued that there is no separate [n]-segment, the intended /n/ being signaled by the nasalization of the [æ̃] and of the following voiced nasalized dental stop [đ̃]. Another observation of some interest is that the [z]-segment is devoiced, i.e., no traces of vocal cord vibrations appear within the fricative.

The stop sound [k] of Claus has first a period of silence, the occlusion. Then comes the explosion in the form of a transient and then a continuant noise structure, the latter part of which is merely an unvoiced beginning of the [l]. It has been shown by Truby[35] that the [l] of a cluster [kl] is often fully articulated even before the explosion is released.

THE DISCRETE VERSUS THE CONTINUOUS VIEW OF SPEECH

Divergent opinions have been expressed on the nature of speech. The concept of speech as a sequence of discrete units with distinct boundaries joined together as beads on a string is contrasted to the view of speech as a continuous succession of gradually varying and overlapping patterns. This divergency has been discussed by Joos[20], Hockett[16], Halle[13,14], Pike[30], and others. What evidence do we have in favor of one or the other view? Fig. 2 illustrates various concepts. These are from the top:

a) A sequence of ideal non-overlapping phonemes.
b) A sequence of minimal sound segments, the boundaries of which are defined by relative distinct changes in the speech wave structure.
c) One or more of the sound features characterizing a sound segment may extend over several segments.
d) A continuously varying importance function for each phoneme describing the extent of its dependency of particular events within the speech wave. Overlapping curves without sharp boundaries.

The models above may appear to represent quite different views of the nature of speech. They are, however, not contradictory in any way. The overlap in the time domain according to d) does not invalidate the concept of the phonemes as discrete and successive in a)[13]. The representation in a) relates to the message aspect of the speech communication whereas representation b) and c) pertain to the speech wave and d) more to the perception of speech.

It is of interest to note that spectrographic pictures of speech often display quite distinct boundaries between successive parts along the time axis. These boundaries are related to switching events in the speech production mechanism such as a shift in the primary sound source, e.g., from voice to noise, or the opening or closing

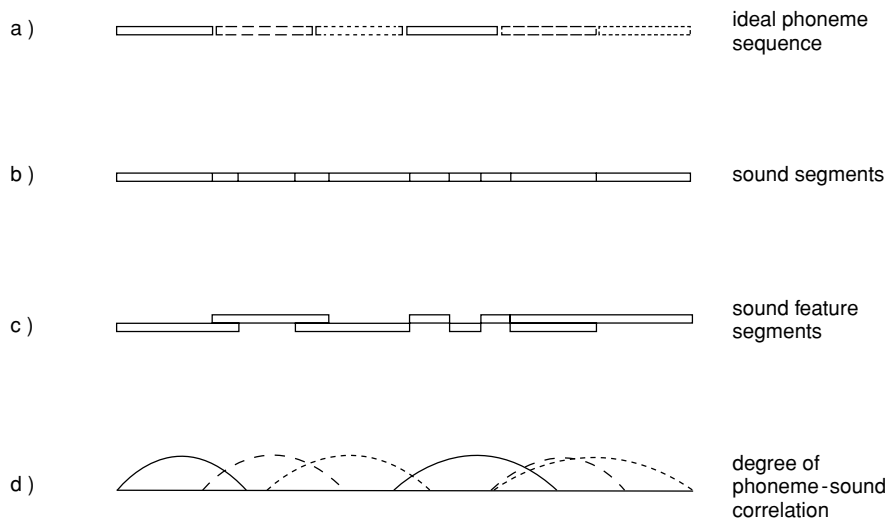


Figure 2. Schematic representation of sequential elements of speech. a) is the phonemic aspect, b) and c) represent acoustic aspects, and d) shows the degree of phoneme-sound correlation.

off of a passage within the vocal cavities, the lateral and nasal pathways included. Less distinct sound boundaries may be defined from typical changes in the pattern of formant frequencies. A common aspect of spectrographic records not shown in Fig. 2 is the more or less continuous variation of some of the formants with respect to their frequency locations. Formant frequency patterns may vary within and across sound segment boundaries.

The number of successive sound segments within an utterance is greater than the number of phonemes. Fully developed unvoiced stops, for instance, contain at least two sound segments, the occlusion and the burst, and the latter may be subdivided into an explosion transient and a short fricative. The first part of a vowel following the burst generally assimilates the voicelessness of the preceding sound. It is a matter of convention whether this sound segment is to be assigned to the vowel, or to the preceding "aspirated" consonant.

Sound segments defined from the procedure above may be decomposed into a number of simultaneously present sound features. Boundaries between sound segments are due to the beginning or end of at least one of the sound features but one and the same sound feature may extend over several successive sound segments. One example seen in the spectrogram of Fig. 1 is the nasalization of a vowel adjacent to a nasal consonant. The most common example would be the continuity of vocal cord vibrations over a series of voiced sounds.

Sound segment boundaries should not be confused with phoneme boundaries. Several adjacent sounds of connected speech may carry information on one and the same phoneme, and there is overlapping in so far as one and the same sound segment carries information on several adjacent phonemes. The typical example is the influence exerted by a consonant on a following vowel. The extent to which a phoneme of the message has influenced the physical structure of the speech wave

often varies continuously along the sound substance as indicated by Fig. 2 d. One practical method of investigating these dependencies is by means of tape-cutting techniques whereby the removal of a part of the sound substance is correlated with the phonemic discrimination loss[7].

POLE-ZERO DESCRIPTIONS OF SPEECH SPECTRA

The engineer's concept of speech is very much influenced by an analytical methodology which has been called analysis-by-synthesis[32]. Any short segment or sample of natural speech may be described in terms of the parameters of a synthesis procedure providing a piece of artificial speech approximating the natural sample with an accuracy which depends on the complexity of the specification.

In one specificational system the spectral energy of a sample is quantitized in terms of the frequency, intensity, and bandwidths of the major energy peaks, the formants. This is the "parallel synthesizer" system in which each formant is fabricated separately and fed in parallel to a mixer. The second system is referred to as the "series synthesizer" scheme in which the output from an electrical analog to the primary sound source is led through a number of consecutive resonance and antiresonance circuits, the combined filtering effect of which is a good approximation to the filtering of the vocal cavities. Providing the bandwidths of these spectral determinants, in mathematical terminology poles and zeros, are made a unique function of their frequency locations, it follows that formant intensities as well as the intensities at any part of the spectrum will be predictable from the frequency locations of the resonances and the antiresonances supplemented by the additional information on the intensity and spectral composition of the source[8].

A complete specification thus comprises a statement of the frequencies and bandwidths of each of the poles and zeros of the vocal tract, and the frequencies and bandwidths of each of the poles and zeros of the source. In addition a scale factor representing source-intensity is needed and a statement concerning the nature of the source, whether of voice or noise character and, if voiced, the frequency of the voice fundamental, F_0. Sounds comprising both a noise source and a voice source are regarded as the superposition of two sounds, one voiced and one of noise character.

The filter function of an ideal nonnasalized vowel does not contain any zeros, i.e., anti-resonance effects. The resonance frequencies of the vocal tract, i.e., the pole frequencies, are labeled F_1, F_2, F_3, F_4, etc. The term F-pattern has been suggested as the compound term for a specification of these frequencies[6,7,8].

As viewed from X-ray moving film, articulation is a continuity of movements. The resulting continuous variations in the dimensions of the vocal cavities determine uniquely the variations of the vocal tract resonance frequencies, the F-pattern. There is thus a continuity of F-pattern within any length of utterance and across any sound segment boundary. However, some boundaries are set by a rapid shift of the F-pattern.

The transitional cues whereby a consonant may in part be identified by its influence on an adjacent vowel may thus be described in terms of F-pattern variations. The term "hub" from the book *Visible Speech*[31] is thus identical with F_2.

Only in non-nasalized, non-lateral sounds produced from a source located at a vibrating or a narrow glottis can the F-pattern up to F_3 be seen with optimal clarity.

Under these circumstances the vocal tract filter function does not possess any zeros. When the source is located higher up in the vocal tract there will appear zeros at approximately the same frequencies as the poles representing the resonances of the cavities behind the consonantal constriction. The spectral contribution of a pole and a zero of the same complex frequency amounts to nothing, i.e., the pole-zero pair may be removed from the specification without any effect on the spectrum of the sound to be synthesized. In these instances the pole and the zero are "bound". Those poles which represent the resonance frequencies of the cavities in front of the source, on the other hand, are "free" from adjacent zeros and thus appear as formants in the spectrum of the sound. There are also "free zeros" which depend on the geometry of the back cavities including those parts of the constriction which lie behind the source. The free zeros may under favorable conditions be seen as spectral minima in an amplitude-frequency display.

The alternative condition for the appearance of zeros in the specification of the vocal tract filter function, is that the sound propagated from the vocal cord source to the lips is shunted by the nasal cavities. Similarly, the nasal output of a sound segment of nasal murmur is submitted to the shunting effects of the mouth cavity as a side chamber. A third possibility is in the production of an [l]-sound. The laterally propagated sound is submitted to some degree of frequency selective shunting by the mouth cavity behind the tongue.

The addition of a shunting cavity system introduces not only zeros but also extra poles. The first nasal resonance in [æ̃] in the spectrogram of Fig. 1 is thus an extra formant. The associated zero, probably located between FN1 and F1 causes a marked decrease of the intensities of both these formants. The other pole-zero pairs of the nasalization do not radically change the phonetic value of the vowel.

If the mouth and nose outlets are closed, there is still some sound propagated through the vibrating walls of the vocal cavities. This is the case of the voiced occlusion of stop sounds. The second and third formants of the voiced occlusion are generally very weak and thus below the reproduction threshold of the spectrograph. If the mouth passage is gradually opened from the closed state to a merely constricted state as in voiced fricatives, there is a gradual rise in the intensities of the higher formants. This rise continues with an increasing mouth-opening and is followed by a shift up in frequency of the first formant. An octave increase in F_1 is correlated with +12 dB increase in the whole spectrum level above F1.

Spectral descriptions in terms of poles and zeros are the results of a processing of the primary data from spectrum analysis which is performed either by means of a digital computer or by means of a series connected speech synthesizer. A third method would be to perform the matching by paper and pencil and an inventory of resonance and anti-resonance curves.

It should be noted that the analysis-by-synthesis approach even without a detailed matching procedure allows a reader of spectrograms to avoid errors in the identification of the particular type and order number of a formant and increases the accuracy in the estimation of formant frequencies. The mere knowledge of the interrelations between frequencies of formants and the relative levels to be expected within a spectrum is an insurance against errors[8].

In most studies for phonetic descriptive purposes, vowels and other zero-free sounds can be described by an F-pattern alone. The source characteristics are generally of very small interest. An exception is F_0, the frequency of the voice fundamental. When zero-functions are to be avoided, the spectra of other sounds are specified in terms of frequencies and intensities of major formants or by some other approximation. The F-pattern should be stated in addition. If the F-pattern formants are not directly observable in the spectrum of the particular sound segment it might be possible to interpolate these frequencies from the F-pattern variations in adjacent sound segments.

A TENTATIVE SYSTEM OF SPEECH SEGMENT CLASSIFICATION

Speech can be divided into a sequence of sound segments the acoustic boundaries of which are definable either from specific articulatory events or from the corresponding time-selective changes in the spectral composition of the speech wave. The following is an attempt to describe the possible structure of these elementary constituents of the speech wave as a basis for phonetic descriptive work and automatic recognition schemes. The classification of sound segments and their sound features should be detailed enough to provide correlates to any category of interest, thus not only to phonemic units. This is the difference of the present approach to that of the earlier work by Jakobson, Fant and Halle[18].

As discussed above, a sound segment is of the dimension of a speech sound or smaller and there may occur several successive sound segments within the time interval of the speech wave traditionally assigned to the phoneme. The number of successive sound segments within an utterance is therefore generally larger than the number of successive phonemes, as conceptually indicated in Fig. 2.

When sound segments are decomposed into bundles of simultaneous sound features it is often seen that a single sound feature carrying a minimal distinction may extend over all sound segments of importance for a phoneme, including sound segments which essentially belong to adjacent phonemes. A typical example of this is the GA /r/ phoneme, the retroflexion (acoustically low frequency F_3) of which generally modifies neighboring sounds. In other instances, such as the voiced/voiceless distinction, it can be the sound segment of adjacent phonemes that carry the major part of the relevant sound feature (the lengthening of a preceding vowel is a voicing cue of intervocalic consonants).

The acoustic basis for identification of sound features and for the establishment of fine gradations and subdivisions within a sound segment of arbitrary composition can be stated in terms of the following parameters, most of which are time variable within a sound segment.

SPEECH PARAMETERS

1. Segment duration.
2. Source intensity (short-time sample of a specified time location within the segment).

3. Source energy (product of segment duration and the time average of source intensity within the segment).
4. Source spectrum (either a short-time sample or the time average of the source intensity-frequency distribution within the segment).
5. Voice fundamental frequency, F_0.
6. F-pattern (= F_1, F_2, F_3, etc.).
7. Sound intensity (short-time sample of a specified location within the segment).
8. Sound energy (product of segment duration and the time average of sound intensity within the segment).
9. Sound spectrum (either a short-time sample or the time average of sound intensity-frequency distribution within the segment).

It has not been attempted to select a set of independent measures in this list, but rather to exemplify what sort of basic data is made use of in acoustic specifications.

The source characteristics have to be determined by removal of the formant structure from the sound. As indicated in the previous section this is done by spectrum-matching or inverse-filtering techniques, in more general terminology by means of analysis-by-synthesis techniques.

The first step in the analysis of a sound segment according to the proposed scheme would be a classification in terms of a set of primary features which will be called the *segment type features*. For convenience, these features are referred to by speech production terminology and are to be considered as binary in nature, i.e., expressing presence versus absence of a specific quality. In this respect they reflect the constraints of the human speaking mechanism and correspond to what is commonly referred to as "manner of production". The second step in the analysis of a sound segment is essentially a classification in terms of the "place of articulation". The term *segment pattern features* is adopted here in order to form a more general concept applicable to both articulatory and speech wave phenomena.

LIST OF SEGMENT TYPE FEATURES

Feature number	*Feature*
	Source features
1	voice
2	noise
3	transient
	Resonator features
4	occlusive
5	fricative
6	lateral
7	nasal
8	vowellike
9	transitional
10	glide[9]

As indicated in the list of segment type features above, there are three possible sound sources supplying the primary acoustic energy of a speech sound segment. These are *voice* (vocal cord vibrations), *noise* (random noise from turbulent airflow through narrow passages and past sharp obstacles), and *transient* (single shock excitation of the vocal cavities). The transient is due to the sudden release of an over-pressure or a sudden checking of an airflow at any obstruction in the vocal cavities, the vocal cords included. In this sense, voice is identical to quasi-periodically repeated transients. In a broad-band spectrogram of a voiced stop the transient is seen as an additional vertical striation which is non-synchronous with the pitch pulses. Additional noise may also be found in this explosion segment, though of less duration than in an unvoiced stop. In an unvoiced stop the transient precedes a noise interval of fricative or aspirative (vowellike) type. The duration of the interval is not the duration of the transient source, which is very small, but the duration of the damped oscillations excited by the transient. When these extend into the following noise interval there is overlap, i.e., co-occurrence of noise and transient. The typical example of co-occurrence of voice and noise is in voiced fricatives. There is a tendency of the voicing to dominate in the early part and the noise in the later part of the fricative. The extreme case of separate sound segments was pointed out in connection with the [z] of Fig. 1.

The resonator features are on the whole independent of the source features. In one and the same sound segment it is possible to find almost any combination of the segment type features. The possible cooccurrences and their statistics have not been studied in detail yet. The resonator features may be described at the level of speech production as follows.

SPEECH PRODUCTION CORRELATES

4. *Occlusice*: Complete closure in the mouth or in the pharynx.
5. *Fricative*: Very narrow passage for the air stream at an obstructed region of the mouth or the pharynx.
6. *Lateral*: Central closure combined with lateral opening in the mouth cavity.
7. *Nasal*: Nasal passages connected to the rest of the vocal system owing to a lowered velum.
8. *Vowellike*: Free passage for the air stream through the pharynx and the mouth cavities.
9. *Transitional*: The articulators moving at a high speed within the segment.
10. *Glide*: The articulators moving at a moderate speed within the segment.

The speech wave correlates of the resonator features may be described as follows:

4. *Occlusive*: The spectrum of a voiced non-nasal occlusive is dominated by a formant Fl of a very low frequency F_1 (the voice bar). However, with considerable high-frequency pre-emphasis it may be possible to detect F2 and F3.

5. *Fricative*: Spectra of voiced fricatives can display the whole F-pattern up to F4 but with less intensity and a lower frequency F_1 than vowellike sounds. A fricative produced with a supraglottal noise source is recognized by a high-frequency noise area in the spectrum. Compared with an unvoiced vowellike sound of a similar articulation, the fricative spectrum displays a larger high-frequency emphasis. The typical fricative is a noise sound, the spectral energy of which is largely contained in formants from cavities in front of the articulatory narrowing.

6. *Lateral*: Sound segments of lateral articulation produced with a voice source possess the vowellike feature except for a reduction of either second, third, or fourth formant intensity due to the first zero of the shunting mouth cavity behind the tongue. An additional high-frequency formant is generally seen. The oral break provides a typical discontinuity in the connection to a following vowel. The lateral sound segment is generally, but not always, of lower frequency F_1 than a following or preceding vowel.

7. *Nasal*: A voiced occlusive nasal (nasal murmur) is characterized by a spectrum in which F2 is weak or absent A formant at approximately 250 c/s dominates the spectrum, but several weaker high-frequency formants (not always seen in spectrograms) occur, one typically at 2200 c/s. These higher formants are generally weaker than for laterals. The bandwidths of nasal formants are generally larger than in vowellike sounds. Voiced vowellike nasal sounds (nasalized vowels) possess the nasal characteristics as a distortion superimposed on the vowel spectrum. Typical nasalization cues are addition of the first nasal resonance in the region below the first formant of the vowellike sound and simultaneous weakening and shift up in frequency of the first formant, F_1.

8. *Vowellike*: The F-pattern formants are clearly visible in the spectrogram. In the case of voiced or unvoiced vowellike sounds produced with a glottal source it is required that at least F1 and F2 be detectable. F3 should also be seen providing F_1 and F_2 are not located at their extreme low frequency limits. A specific feature of sounds produced with a glottal source is that the relative formant levels are highly predictable from the particular F-pattern, i.e., from the formant frequency locations[10]. Vowellike noise sounds produced from a supraglottal source possess a rather weak first formant, F_1. This is especially the case with [h]-segments produced with a tongue articulation of a high front vowel. Unless the fricative feature is superimposed there should not occur a prominent high-frequency noise area in the sound spectrum.

9. *Transitional*: The spectrum changes at a relatively fast rate in the segment. The first part of a vowel following a voiced stop or nasal is characterized by a rapid change in at least one formant frequency, e.g., F_1. The transitional sound segment ends where the major part of the formant transition is completed.

10. *Glide*: The spectrum changes at a relatively slow rate but faster than for a mere combination of two vowels. Variants of [r] [1] [j] [w] sounds occur as glides.

SEGMENT PATTERN FEATURES

Articulation	Speech Wave
11. Tongue fronted	$F_2 - F_1$ large.
a) Prepalatal position	F_2 high, F_3 maximally high.
b) Midpalatal position	F_2 maximally high and close to F_3.
12. Tongue retracted	$F_2 - F_1$ small F_1 comparatively high.
13. Mouth-opening (including tongue section and lips) narrow	F_1 low.
14. Lips relatively close and protruded (small lip-opening area)	$F_1 + F_2 + F_3$ lower than with a larger lip-opening and the same tongue articulation. A progressing lip closure alone causes a decrease in each of F_1, F_2, and F_3 but with varying amounts depending on the particular tongue position. The effect on F_3 is pronounced in case of prepalatal tongue positions.
15. Retroflex modification	
a) Alveolar articulation	F_4 low and close to F_3.
b) Palatal articulation	F_3 low and close to F_2.
16. Bilabial or labiodental closure	F_2 in the region of approximately 500–1500 c/s depending on the tongue location of the associated vowel or vowellike segment. A palatal tongue position favors high F_2. The noise spectrum of the fricative [f] is essentially flat and of low intensity.
17. Interdental articulation	F_2 1400–1800 c/s. Fricative noise of [θ] much weaker than for [s] and with a more continuous spectrum. Center of gravity is higher than for the labiodental fricative [f].
18. Dental or prealveolar articulation	F_2 in the region of 1400–1800 c/s, F_3 high. Fricative noise strong. The main part of the [s]-energy is above 4000 c/s. This cutoff frequency is lower for alveolar than for dentals.
19. a) Palatal retroflex articulation	F_3 low. The fricative noise of [ʂ] is of high intensity and is carried by F_3 and F_4.
b) Palatal articulation with tip of tongue down	F_2 and F_3 high. Strong fricative noise centered on F_3 and F_4 and also on F_2 providing the tongue pass is sufficiently wide. The lower frequency limit of [ç] noise is higher than for retroflex sounds.
20. Velar and pharyngeal articulation	F_2 medium or low. A large part of the fricative noise is carried by F2. The F-pattern except F1 is clearly visible.
21. Glottal source	The entire F-pattern including F1 is visible.

The existence of *complex articulations* should be kept in mind. The most apparent example referred to above is the freedom of the tongue to take any position during lip closure which makes the F-pattern of labials variable. In dentals the back of the tongue is partially free to approach the back wall of the pharynx which lowers F_2 and increases F_1. This is the case of the "dark l". The articulatory contrast between a wide unobstructed and a narrow divided pharynx, resulting in a high versus a low F_2, is the counterpart of the hard/soft distinction in Russian consonants.

THE NORMALIZATION PROBLEM

The phonetic identity of a speech sound is to some extent dependent on the sound context, that is, formant frequencies within a sound segment have to be judged by reference to the average formant frequencies of the speaker and to have his voice fundamental frequency. Variational features are in some respects more essential than the absolute characteristics of the speech wave. This fact is confirmed by experiments with synthetic speech[23].

An international standard of phonetic pronunciation norms could be established by reference to a few selected speakers. For the Cardinal vowels the pronunciation of Daniel Jones has been considered authoritative[21,22]. Another alternative is synthetic speech[3]. The quality of synthetic speech[8] can be made sufficiently high to fulfill the minimum requirements of naturalness. The advantage compared with real speech would be that the acoustic specification could be made more exact.

A more difficult task is to establish a unique code between the measurable parameters of any sample of live speech and its absolute phonetic quality[28]. The analysis-by-synthesis approach[32] would be to specify the sample by the parameters of synthesis providing equal phonetic quality.

We would also like to be able to predict these settings from the available data inherent in the speech wave. However, normalization techniques have not developed far enough yet even for the simpler task of machine identification of the phonemic structure of a spoken message.

One of the most important factors involved in normalization is to take into account the influence of the size of the speaker's vocal tract. The F-pattern frequencies are to a first approximation inversely proportional to the length of the speaker's vocal tract from the glottis to the lips. Children have smaller heads than adults and their formant frequencies are thus on the average higher. The average female-male difference is of the order of 20%. However, normalization is not merely a question of a constant scale factor.

SUMMARY

This article aims at summarizing the present status of speech analysis techniques, specifically spectrographic analysis. Special attention is given to the problems of segmenting speech into successive phonetic elements and to the categorization of such minimal sound segments in terms of segment type (manner of production) and segment pattern (place of articulation). For this purpose the relations between

the physiological parameters of speech and corresponding acoustic speech wave characteristics have been summarized in the form of a dictionary.

It is pointed out that segment boundaries are associated with changes in the manner of production (voiced/voiceless, fricative/non-fricative, nasal/non-nasal, etc.) whereas the place of articulation determines acoustic patterns that vary more or less continuously within and across segment boundaries. There are as a rule a larger number of sound segments than phonemes in any utterance. For this reason and because of coarticulation effects any phoneme is generally signalled by several successive sound segments. Conversely, any sound segment is generally influenced by several adjacent phonemes of the speech message transcription.

NOTES

[1] Paper presented at the Wenner-Gren Foundation for Anthropological Research Symposium on Comparative Aspects of Human Communication at Burg Wartenstein/Austria, September 1960.

[2] The research reported in this article has in part been carried out under contract USAF 61 (052)-342, and with support by the Swedish State Council of Technical Research.

[3] As a matter of fact I have not met one single speech researcher who has claimed he could read speech spectrograms fluently, and I am no exception myself. I only know of the group of subjects at Bell Telephone Laboratories who participated in a Visible Speech learning experiment in 1945. Speech researchers would, however, benefit from going through this learning process. It would, aid them in teaching machines to do the same job.

[4] A few words may be apropos here to explain the nature of distinctive features. If a minimal difference is found between two phonemes, it is highly probable that the same distinction will recur in several other phoneme pairs. Thus, the difference between /s/ and /f/ is the same as between /z/ and /v/, and between /t/ and /p/, and /d/ and /b/, and between /n/ and /m/. This is the' acute/grave distinction according to the terminology of Jakobson, Fant, and Halle. It is similar to and stands in complementary distribution with the distinction between /i/ and /u/, /e/ and /o/, and /ae/ and /a/, which motivates the usage of the same term acute/grave also for vowels. Within the consonants referred to above it is apparent that the relation of /z/ to /s/ is the same as of /v/ to /f/, /b/ to /p/, and /d/ to /t/. The main advantage of the distinctive features approach is that the number of basic signs is minimized. Maximally 12 distinctive features are sufficient for defining any phoneme of most languages.

[5] Phonetics Department, University of Edinburgh; Massachusetts Institute of Technology, Cambridge/Mass., U.S.A.; Royal Institute of Technology, Stockholm, Sweden.

[6] General American. The subject was born in Texas.

[7] See for instance Truby[35] and Öhman[36,37]

[8] I know of several vowel studies providing data on formant frequencies which are invalidated by an inability of the investigator to keep track of one and the same formant within a series of vowel sounds. Nasal formants are often confused with the F-pattern formants. Similar difficulties exist in automatic formant-tracking schemes.

[9] In a recent publication it has been considered desirable to omit feature 10 since it may be included in feature 9. See Fant, Lindblom[10].

[10] An alternative to the *vowellike* feature would accordingly be *zero-free* which on the speech production level implies non-nasalized, non-lateral, glottis source sound, acoustically correlated to the predictability of formant levels from the F-pattern. An apparent F1 would be one necessary condition. A second alternative would be to retain the term vowellike but restrict the feature to glottal sources, in which case the first formant F1 must be present. However, in case of both whispered vowels and [h]-sounds articulated with the tongue in an [i]-position it is highly probable that supraglottal sources exist alone or in addition to glottal sources. This

would lead to the classification of some whispered vowels and [h]-sounds as vowel-like and other as non-vowellike. The best choice among these alternatives has to be determined from experience.

REFERENCES

1. David, E. E., Jr.: Artificial Auditory Recognition in Telephony. IBM J. 2, 1958, 294–309.
2. Davis, K. H., Biddulph, R. and Balashek, S.: Automatic Recognition of Spoken Digits, in Communication Theory, ed. W. Jackson, London, 1953, 433–441.
3. Delattre, P., Liberman, A. M. and Cooper F. S.: Voyelles synthétiques a deux formantes et voyelles cardinales. Maitre Phonétique, 96, 1951, 30–36.
4. Delattre, P., Liberman, A. M. and Cooper, F. S.: Acoustic Loci and Transitional Cues for Consonants. J. Acoust. Soc. Am. 27, 1955, 769–773.
5. Fant, C. G. M.: On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies. For Roman Jakobson, 's-Gravenhage, 1956, 109–120.
6. Fant, C. G. M.: Modern Instruments and Methods for Acoustic Studies of-Speech. Proc. of VIII Internat. Congr. of Linguistics, Oslo. Oslo, 1958, 282–358; and Acta Poly-technica Scandinavica Ph 1 (246/1958), pp. 1–91.
7. Fant, C. G. M.: Acoustic Theory of Speech Production. 's-Gravenhage, 1960.
8. Fant, C. G. M.: Acoustic Analysis and Synthesis of Speech with Applications to Swedish. Ericsson Technics. 15, No. 1, 1959, 3–108.
9. Fant, C. G. M.: Phonetics and Speech Research, invited paper presented at the International Conference on Research Potentials in Voice Physiology, Syracuse, N. Y., May 29-June 2, 1961; to be publ. in the Proc. from this conference.
10. Fant, C. G. M. and Lindblom, B.: Studies of Minimal Speech Sound Units. Speech Transmission Laboratory, Quarterly Progress and Status Report No. 2/1961, 1–11.
11. Fry, D. B. and Denes, P.: The Solution of Some Fundamental Problems in Mechanical Speech Recognition. Language and Speech, 1, 1958. 35–38.
12. Halle, M.: The Sound Pattern of Russian. 's-Gravenhage, 1959.
13. Halle, M.: The Strategy of Phonemics. Word. 10, 1954, 197–209.
14. Halle, M.: Review of Manual of Phonology by C. F. Hockett. J. Acoust Soc. Am. 28, 1956. 509–511.
15. Halle, M., Hughes, C. W. and Radley, J. P.: Acoustic Properties of Stop Consonants J. Acoust. Soc. Am. 29, 1957, 107–116.
16. Hockett, C. F.: Manual of Phonology. Indiana Univ. Publications in Anthropology and Linguistics, No. 11, Bloomington, 1955.
17. Hughes, C. W. and Halle, M.: Spectral Properties of Fricative Consonants. J. Acoust. Soc. Am. 28, 1956, 303–310.
18. Jakobson, R., Fant, C. G. M. and Halle, M.: Preliminaries to Speech Analysis. M.I.T. Acoustics Lab. Tech. Rep. No. 13. 1952; 3rd printing.
19. Jakobson, R. and. Halle, M.: Fundamentals of Language. 's-Gravenhage, 1956.
20. Joos, M.: Acoustic Phonetics. Language, 24, 1948, 1–136.
21. Ladefoged, P.: The Classification of Vowels, Lingua, 5, 1956, 113.
22. Ladefoged, P.: The Perception of Vowel Sounds. Edinburgh, 1959, Ph.D. Thesis, Univ. of Edinburgh.
23. Ladefoged, P. and Broadbent, D. E.: Information Conveyed by Vowels. J. Acoust. Soc. Am. 29, 1957, 98–104.
24. Liberman, A. M.: Some Results of Research on Speech Perception. J. Acoust Soc. Am. 29, 1957, 117–123.
25. Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P. and Cooper, F. S.: Minimal Rules for Synthesizing Speech. J. Acoust. Soc. Am. 31, 1959, 1490–1499.
26. Miller, R. L.: Auditory Tests with Synthetic Vowels. J. Acoust. Soc. Am. 25, 1953. 114–121.
27. Olson, H. F. and Belar. H.: Phonetic Type writer. J.'Acoust. Soc. Am. 28, 1956, 1072–1081.

28. Peterson, G. E.: The Information Bearing Elements of Speech. J. Acoust. Soc. Am. 24, 1952, 629–637.
29. Peterson, G. E. and Barney, H. L.: Control Methods Used in a Study of the Vowels. J. Acoust. Soc. Am. 24, 1952, 175–184.
30. Pike, K. L.: Language as Particle, Wave and Field. The Texas Quarterly II, 1959, 37–54.
31. Potter, R. K., Kopp, A. G. and Green, H. C.: Visible Speech. New York, 1947.
32. Stevens, K. N.: Toward a Model for Speech Recognition. J. Acoust. Soc. Am. 32, 1960, 47–55.
33. Stevens, K. N. and House, A. S.: Development of a Quantitative Description of Vowel Articulation. J. Acoust. Soc. Am. 27, 1955, 484–493.
34. Stevens, K. N. and House, A. S.: Studies of Formant Transitions Using a Vocal Tract Analog. J. Acoust. Soc. Am. 28, 1956, 578–585.
35. Truby, H. M.: Acoustico-Cineradiographic Analysis Considerations with especial reference to certain consonantal complexes. Acta Radiologica, Suppl. 182, Stockholm, 1959, Ph.D. Thesis, Univ. of Lund.
36. Öhman, S.: On the Contribution of Speech Segments to the Identification of Swedish Consonant Phonemes. Speech Transmission Laboratory, Quarterly Progress and Status Report No. 2, 1961, pp. 12–15.
37. Öhman, S.: Relative Importance of Sound Segments for the Identification of Swedish Stops in VC and CV Syllables. Speech Transmission Laboratory, Quarterly Progress and Status Report No. 3, 1961, pp. 6–14.

CHAPTER 4.2

ACOUSTICAL ANALYSIS OF SPEECH*

1. INTRODUCTION

The object of acoustical analysis of speech is the sound field emitted by a human speaker, usually picked up from a pressure-sensitive microphone. The purpose of the analysis is to perform a signal processing appropriate for sampling, storage, and visualization of relevant data. The most common procedure is spectrographic analysis, that is, to display the speech wave as a pattern in a time-frequency-intensity domain. Supplementary analyses of the temporal variation of the voice fundamental frequency F0, formant frequencies, intensity, and voice source parameters may be synchronized with a spectrographic display for general phonetic descriptive purposes and for extraction of information bearing elements with applications in speech synthesis and recognition.

Acoustical analysis of speech also involves the acoustical aspects of the speech production mechanism [11]. The time-frequency-intensity patterns of speech waves are intimately related to the sound generating and sound-shaping mechanisms. With this wider view of the acoustics of speech, we are in a better position to relate speech wave patterns to the framework of phonetics and linguistics, in other words, to derive acoustical correlates of vowels and consonants and of prosodic categories such as stress and intonation patterns, thus paving the way for a better understanding of the speech code. A more complete treatment of this subject is found in [1–3].

2. SPECTRUM ANALYSIS

Basically two types of spectral representations are used in speech analysis, amplitude versus frequency and frequency-intensity versus time. The amplitude versus frequency analysis usually pertains to a frame of short duration, of the order of 3–30 ms, and is associated with a specific location in time. This is referred to as a spectral section. If the frame duration is extended to cover a larger portion of speech of the order of a sentence or more, the result is a long-time average spectrum, which mainly reveals speaker-specific characteristics. The short-time spectral section can provide quantitative data on specific parts of a speech sound. In order to visualize an utterance in all its details, we need a running short-time analysis with updating of spectrum calculations at very short intervals of the order of 1 ms. This is the basis of the time-frequency-intensity spectrogram commonly used in acoustic-phonetic studies of speech.

In a spectrogram, see Figure 1, spectral intensity is represented by the density or by the color of the marking within areas of minimal time-frequency resolution. These are conceptually defined by an extension T in time and B in frequency. By the law of reciprocal spread, the product BT is a constant of the order of 1. Thus a fine frequency resolution, that is, a small B, implies a large T and thus a low temporal resolution. This is needed for portraying individual harmonics, which require that

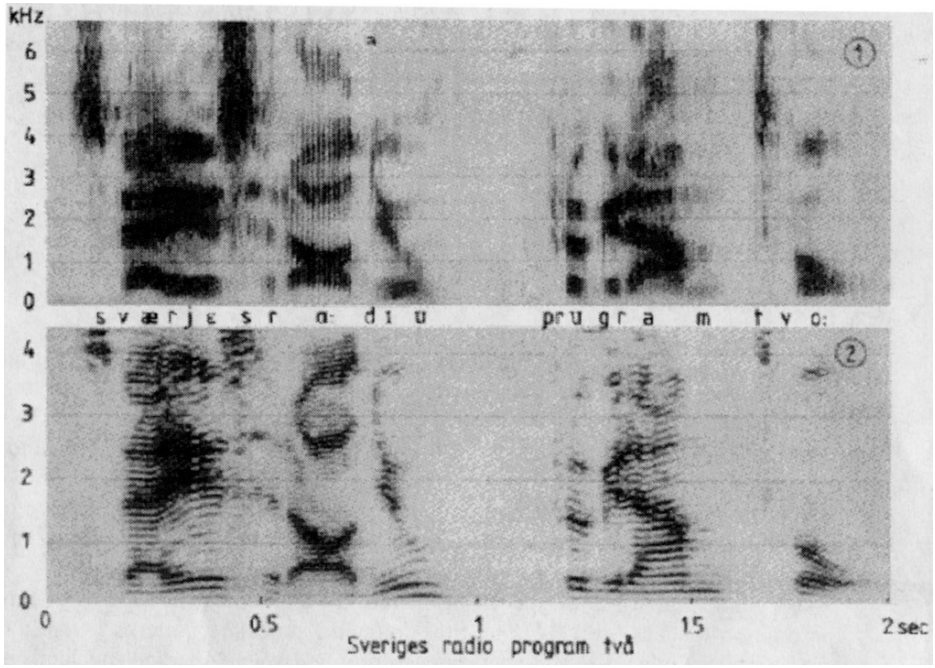


Figure 1. Frequency-intensity-time spectrograms. Broadband analysis above and narrowband analysis below.

B is smaller than F0, the fundamental frequency of voiced sounds. F0 is defined as the inverse of the time interval T_0 between consecutive air pulses emitted through the vibrating vocal cords, $F0 = 1/T_0$. The typical range for a male voice is 60–200 Hz and somewhat less than an octave higher for an average female voice. If, on the other hand, the analysis bandwidth B is larger than F0 the harmonics are no longer resolved, and the spectral pattern is dominated by individual formants, that is, peaks induced by vocal tract resonances.

The increased temporal resolution associated with the larger B makes it possible to follow rapid spectral variations such as the onset and decay of spectral intensity within a voice fundamental period. Thus, the narrowband analysis $B < F0$, suits a frequency domain harmonic representation, while the broadband analysis, $B > F0$, brings out a time-domain view of voiced sounds, as a sum and sequence of damped sinusoids representing the response of vocal tract resonant modes to each successive excitation impulse from the glottal source. One such damped oscillation is thus the time domain equivalent of a formant. The implication is that the bandwidth should be chosen to satisfy one of these two conditions. In practice, narrowband analysis is mainly used for amplitude-frequency sectioning of vowels and other voiced sounds, while broadband analysis is commonly used in time-frequency-intensity spectrograms and for the sectioning of unvoiced fricatives and stops, as illustrated in connection with Figure 2. An intermediate choice of B of the order of F0 creates

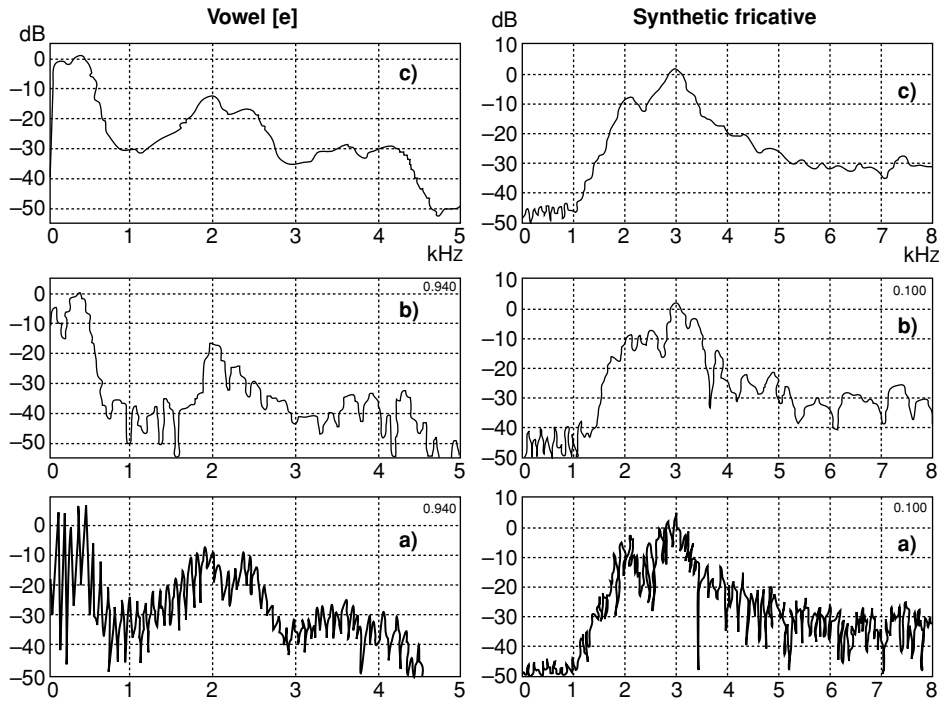


Figure 2. Spectral sections of a vowel [e] (left) and of a synthetic fricative [ʃ] (right). Sampling frequency 16 kHz.
a) simple FFT with N = 640 sample points, B = 25 Hz, df = 25 Hz frequency spacing, T = 40 ms.
b) N = 64, B = 250 Hz, 576 zero samples added, df = 25 Hz, T = 4 ms.
c) same as b) but average is taken over 10 successive 4 ms spectrum samples, T = 40 ms.

an ambiguity of what is a formant and what is a harmonic. The analysis bandwidth as well as the frequency spacing of data points may also be adapted to an auditory scale, such as the technical mel scale [4].

$$\text{Mel} = 1000[\ln(1 + \text{f}/1000)/\ln 2] \tag{4.2.1}$$

or an approximation to the Bark scale [5]

$$\text{Bark} = 7\ln[\text{x} + (\text{x} + 1)^{0,5}] \tag{4.2.2}$$

where $\text{x} = \text{f}/650$.

3. Technical Methods of Spectrum Analysis

Early attempts to perform spectrum analysis of speech employed a Fourier series analysis of speech waveforms recorded on an oscillograph. The breakthrough came in the middle of the 1940s with the Visible Speech sound spectrograph developed at the Bell Telephone Laboratories [6]. This became the prototype of the Key Electric

Sonagraph, a spectrum analyzer that produced printouts by passing a spectral intensity modulated current through an electro-sensitive paper. Today, spectrographic analysis is generally performed digitally with a general-purpose computer and a laser printout. A definition equal to that of a special purpose analog spectrograph sets demands both on the printer and on the rate of overlapping frames of spectrum analysis.

Spectrum analysis may be performed either as a filtering process or by a digital Fourier analysis of the waveform based on the discrete Fourier transform (DFT) [7–9], which usually is implemented by the Fast Fourier Transform (FFT) [10]. The most common filtering method is by means of a filter bank analysis. Another method is to employ a single bandpass filter, narrow-band or broad-band, the effective center of which is varied continuously as in a wave analyzer. This method was used in the Sonagraph.

The conceptual analogy between the output of a bandpass filter and a component of a Fourier series is basically related to the temporal weighting function of the envelope of the impulse response of the bandpass filter.

A mathematical identity exists [4], when we consider the speech signal to be convolved with the impulse response envelope as a weighting function before being submitted to a formal Fourier analysis. The effective duration of the impulse response, which is of the order of 1/B serves as a functional equivalent to the frame duration T of a DFT defined by

$$V_n = \sum_{i=0}^{N-1} s_i \exp(-j2\pi ni/N) \tag{4.2.3}$$

where *N* is the number of samples, s_i is the sample nr i, and V_n the spectral component number n. For an approximate rectangular window function, the frame duration is $T = N/F_s$ where F_s is the sampling frequency. A typical combination for narrow-band analysis would be $N = 512$ waveform samples of $F_s = 16$ kHz, which provides $N/2 = 256$ spectrum samples up to the Nyquist frequency 8 kHz. These are spaced at intervals of $df = B = 1/T = 31.25$ Hz, each representing a bandwidth of $B = 1/T = 31.25$ Hz. For broadband applications, say for an equivalent bandwidth of $B = 1/T = 250$ Hz, there is a need to calculate spectrum samples at smaller intervals than B.

An improved definition of the spectrum envelope is attained by adding a frame of zero amplitude samples prior to the computation. Thus, with $N = 64$ waveform samples at $F_s = 16$ kHz appropriate for $B = 250$ Hz, we need to add $64(8 - 1) = 448$ zero amplitude samples in order to achieve a spacing of spectral data points of 31.25 Hz. Similarly, for a temporal definition higher than that implied by successive frames at T second intervals, one must repeat the calculation at shorter intervals than T. For a broadband spectrographic display, this is achieved by updating spectral calculations at intervals of the order of 1–2 ms, or as densely as motivated by the definition of the computer screen or the recording medium.

An additional averaging of spectral data is needed for removing statistical uncertainties in amplitude-frequency sections of unvoiced sounds. The randomly occurring striations, which are typical of time-frequency-intensity spectrograms of

fricatives, are the natural consequences of the filtering of random noise. These striations occur at random time intervals that average 1/B. For producing clean amplitude-frequency sections with an acceptable random error, a spectral section needs to be averaged over a time interval T_a, substantially greater than the frame duration $T = 1/B$.

From statistical considerations it follows that the uncertainty in the estimate of the amplitude A_e of a spectral component is proportional to $(T/T_a)^{0,5}$. On a decibel scale, and assuming T_a greater than T, the following approximate expression for the random ripple σ_e. superimposed on a spectral component A_e has been empirically validated [1]:

$$20\log_{10}(\sigma_e/A_e) = 4(BT_a)^{-0,5} \qquad \text{d B} \tag{4.2.4}$$

A convenient method of averaging, especially suited for filter bank spectrum analysis, is by means of a lowpass filtering of the rectified signal in each channel in which case the substitution $T_a = 1/2W$, where W is the lowpass cutoff frequency may be used.

The practical consequences of the choice of analysis parameters are illustrated in Figure 2 pertaining to a human vowel [e] and a synthetic unvoiced fricative [$\int$]. In both cases the sampling frequency was 16 kHz. A direct FFT spectrum calculation with $N = 640$ waveform samples covering an interval of $T = 40$ ms, produces 320 spectrum samples up to 8 kHz with a bandwidth $B = 25$Hz and a spacing of $df = 25$ Hz. This direct approach is adequate for portraying the harmonic structure of the vowel spectrum. Here, we may identify the voice fundamental frequency $F0 = 90$ Hz and formant peaks at $F_1 = 400$ Hz, $F_2 = 1900$ Hz and $F_3 = 2400$ Hz. The weakly apparent peak at 3250 Hz is associated with F4 and the peaks at 3700 Hz and 4200 Hz with F5 and F6 respectively.

The same direct approach applied to the fricative produces a random fine structure that in accordance with Eq. (4) is of the order of ± 4 dB. Although the two formants at 2000 Hz and 3000 Hz appear as expected, there remains an uncertainty about the possible occurrence of other spectral peaks. A more appropriate processing, of a 40-ms sample of the fricative involves the following two steps. First, an analysis bandwidth of $B = 250$ Hz, requiring $N = 64$ waveform samples, is selected. In order to retain a spectral definition of $df = 25$ Hz, a number of $9(64) = 576$ empty samples are added before the FFT is executed. However, the random component in the spectrum remains the same, as illustrated by the middle graph of Figure 2. It is effectively reduced by taking the average of 10 successive FFT spectra covering an effective frame width of $T_e = 40$ ms, as illustrated by the top graph of Figure 2. In agreement with Eq. (4), the ripple is now reduced to the order of $+/-$ 1 dB.

The same operation applied to the vowel also enhances the outline of the spectrum envelope by effectively removing the harmonic fine structure. In these examples, the analysis embraced a 40-ms speech sample. For analysis of stop bursts, a 20-ms sample is recommended. It should also be kept in mind that the effective duration of a Hanning or Hamming window is about 40% less than the nominal value of a rectangular window. In other words, the particular weighting function emphazises the middle part of the window.

An alternative and more common method of deriving a smoothed spectrum envelope, as in Figure 2a, is by means of a cepstrum de-convolution [7–9], the first step of which involves calculating the log spectrum of a sufficiently long frame of the speech wave, usually of the order of 40 ms. This is submitted to an inverse Fourier transform that is windowed to retain a region up to a few milliseconds after which follows a new Fourier transform which is the cepstrum envelope. One may also apply a linear prediction (LPC) analysis to model the spectrum envelope as an all-pole system [7–8], see the section of speech parameter extraction.

Another point often overlooked in digital spectrum analysis of speech is the difference between a conventional DFT or FFT and a proper Fourier series harmonic extraction. The relation is

$$A_n = 2V_n/T_0 \tag{4.2.5}$$

where V_n is the Fourier integral component, Eq. (3). T_0 is the voice fundamental period $T_0 = 1/F0$, and A_n is the amplitude of the corresponding harmonic. The factor 2 derives from the transform from a double-sided to a single-sided spectrum. Scale factors are often lost in digital processing of speech. In this case, if lost, the F0 scale factor would cause errors in the tracking of the amplitude of the voice fundamental or its harmonics within an utterance.

4. SPEECH PARAMETER EXTRACTION

One object of speech analysis is to extract essential parameters of the acoustical structure, which may be regarded as a process of data reduction and an enhancement of information-bearing elements. In order to attain an effective and reasonable complete specification, one has to rely on synthesis models. An important subsystem is that of voiced sounds with negligible coupling to the nasal and subglottal systems.

In a source-filter decomposition [11], the filter function is ideally described by the so called F-pattern. These vocal tract resonance modes F1, F2, F3, F4 . . . are quantified by their formant frequencies F_1, F_2, F_3, F_4, the corresponding formant bandwidths B_1, B_2, B_3, B_4, formant amplitude levels L_1, L_2, L_3, L_4. The labels F1, F2, F3, F4 may be used to refer to formant peaks as well as to formant frequencies. Also, an alternative notation for F_0 is F0, which is used here. A present trend is to avoid subscripts if they are not needed.

For male speech the normal range of variation is F_1 = 180–800 Hz, F_2 = 600–2500 Hz, F_3 = 1200–3500 Hz, and F_4 = 2300–4000 Hz. The average distance between formants is 1000 Hz. Females have on the average 20% higher formant frequencies than males, but the relation between male and female formant frequencies is non-uniform and deviates from a simple scale factor [2, 30]. Formant bandwidths may vary considerably. They are of the order of $B_n = 50(1 + f/1000)$ Hz.

Formant amplitudes vary systematically with the overall pattern of formant frequencies and the spectral properties of the voice source. The basic parameter of the voice source is the fundamental frequency F0 which determines the intonation.

Some of these parameters may be inferred from visual inspection and processing of a spectrogram or an oscillogram. Thus, in a narrowband spectrogram, F0 may be

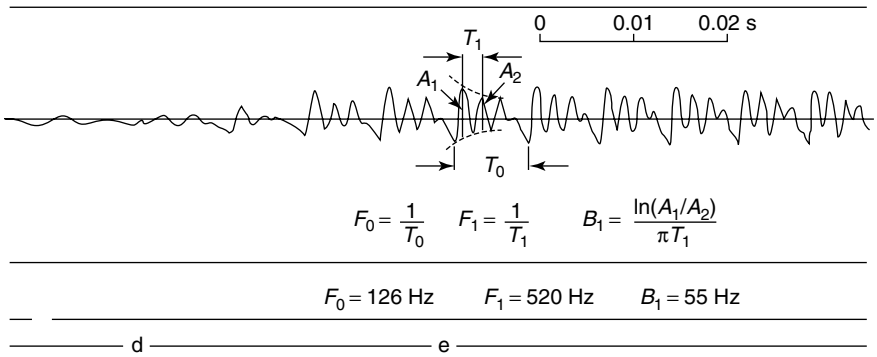


Figure 3. Oscillogram of the syllable [de] with extended time scale and removal of energy above F_1, illustrating the time-domain extraction of voice fundamental frequency, F_0, the first formant frequency F_1, and first formant bandwidth B_1. From [4].

traced from one of the harmonics. Crude information on F0 is also available in a broadband spectrogram from the distance $T_0 = 1/F0$ between successive voice pulse striations. The time domain extraction of F0 is performed with greater accuracy in an oscillogram with expanded time scale. With suitable preprocessing to isolate a single formant, the oscillographic trace may also be used to determine formant frequencies and bandwidths, as is illustrated in Figure 3. The bandwidth is here determined from the exponential decay factor , $\exp(-\pi B_n t)$, of the formant oscillation envelope within the closed glottis interval of the voice cycle.

Fully automatic methods of continuous tracking of speech parameters, such as F0 and formant frequencies, still need to be established. The problem lies in the reliability of performance [9]. Any existing system of F0 extraction is prone to fail at some instances depending on individual voice types and registers. Typical errors are jumps to the second harmonic and the indeterminacy of F0 at aperiodicities such as voice creaks. A system may perform quite well on one type of voice but not on other voicees.

An even more difficult task is the automatic extraction of formant frequencies and bandwidths. Linear prediction coding, which models voiced sounds as the response of an all-pole function to a simple sequence of excitation impulses [13] has become a standard routine for formant tracking. However, the human voice source departs from the all-pole model in several ways and may show local spectral peaks and dips.

Also, the vocal tract transfer function generally contains both poles and zeros due to nasal and subglottal coupling. Even with more realistic and complex models of speech production, an automatic extraction of formant frequencies frequently breaks down. This is especially the case at high fundamental frequencies where the tracking system tends to pick up harmonics rather than formants. For simple applications in vowel analysis, LPC techniques may provide accurate estimates of formant frequencies, providing the results are carefully checked against spectrograms. Typical errors [14] are temporal jitter from one frame to the next, tendencies at high F0 to synchronize on harmonics, and occasional jumps to a higher

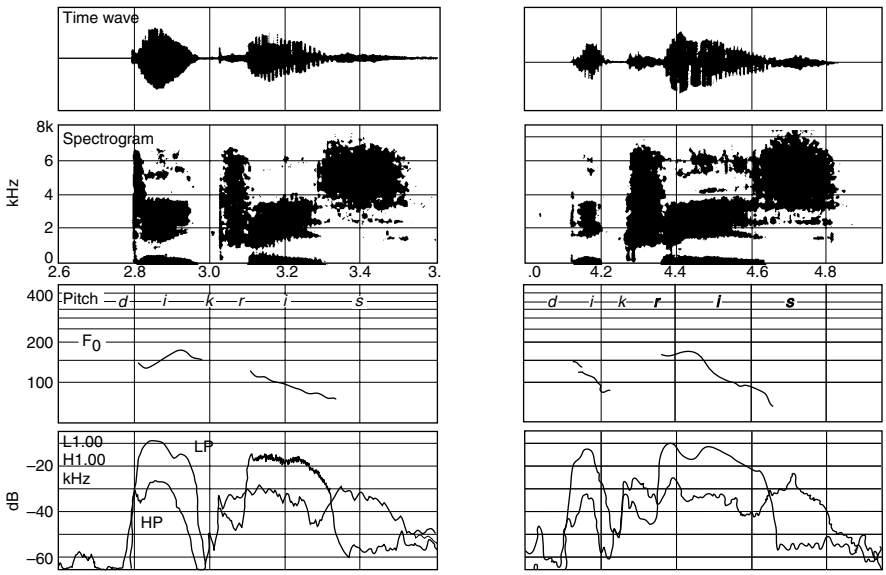


Figure 4. Oscillogram, spectrogram, F0 contour, intensity (lowpass 1 kHz and highpass 1 kHz) illustrating contrasting stress patterns of the word "decrease", pronounced as noun (left) and as verb (right).

or a lower formant. Linear prediction coding determination of bandwidths is less reliable.

For many laboratory applications, the computer derived spectrogram is combined with a synchronous oscillogram and an extracted F0 curve, preferable on a log frequency scale. In addition, one may display the time course of overall intensity with or without special pre-emphasis or other filtering. This is exemplified in Figure 4, which illustrates two words with contrasting stress patterns: the noun *decrease* with stress on the first syllable and the verb *decrease* with stress on the second syllable. The two intensity curves at the bottom pertain to lowpass and highpass filtering respectively, both with cut off frequencies at 1000 Hz. The relative dominance of the first syllable of the noun and of the second syllable of the verb is reflected in a larger duration and in higher F0 and intensity. The contrast in intensity comparing the first and the second syllable is especially apparent in the highpass intensity contour, which reflects a relative greater dominance of higher partials of the voice source at a higher stress level.

5. THE VOICE SOURCE

The temporal and spectral shape of any speech sound is a function of both source and transfer function characteristics, which is illustrated in Figure 5. The source of voiced sounds may be regenerated by submitting the speech wave to an inverse filtering, which in effect cancels the poles and zeros of the transfer function [12, 15, 16, 28, 29]. Since the transfer from volume velocity at the lips to the sound pressure

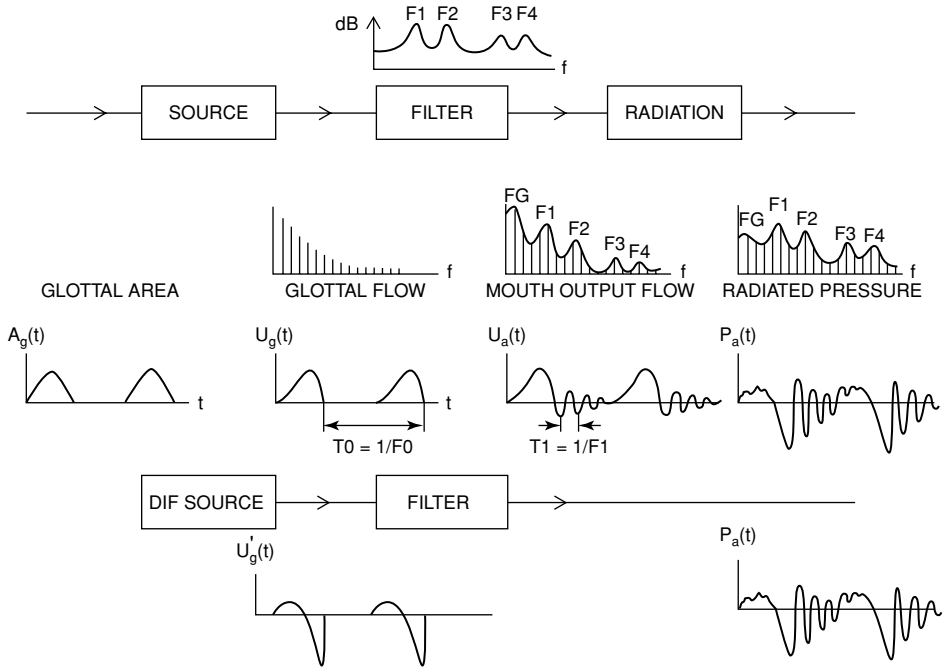


Figure 5. Frequency- and time-domain view of source-filter decomposition of voiced sounds. The negative-going peak of glottal flow derivate source at the bottom is a scale factor of formant amplitude.

in front of the speaker involves a differentiation, the net result of an inverse filtering without integration is to regenerate the time derivative of glottal flow, which appears in the lower left part of the figure. The negative-going spikes of this pulse train are the derivatives of glottal flow at instances of glottal closure. These are measures of excitation strength.

A continuous inverse filtering is illustrated in Figure 6, which pertains to the second syllable of the word *adjö,* [ajø:] In the lower graph a sample of the [ø:] vowel has been submitted to a routine analysis of voice source parameters, which shows an overall fair match with our reference LF model [16]. Exceptions in the form of time-domain and frequency domain irregularities occur, for example, the double peaked positive part of the glottal flow derivative and the associated spectral minimum just above 1 kHz. Such perturbations are a natural consequence of superposition and aerodynamic non-linearities adding to speech naturalness [27–29].

A further complication in voice source analysis is the covariation of source and filter characteristics associated with glottal adduction-abduction gestures [17–19]. Another is the sensitivity of the source parameters to vocal tract constrictions [28, 29].

The inhibitory influence of a supra-glottal narrowing on the voice excitation amplitude explains the relative low amplitude and smoothed out waveform of the [j]-source function in Figure 6. Similar interaction effects are also encountered in semi-constricted vocalic intervals. [15, 28].

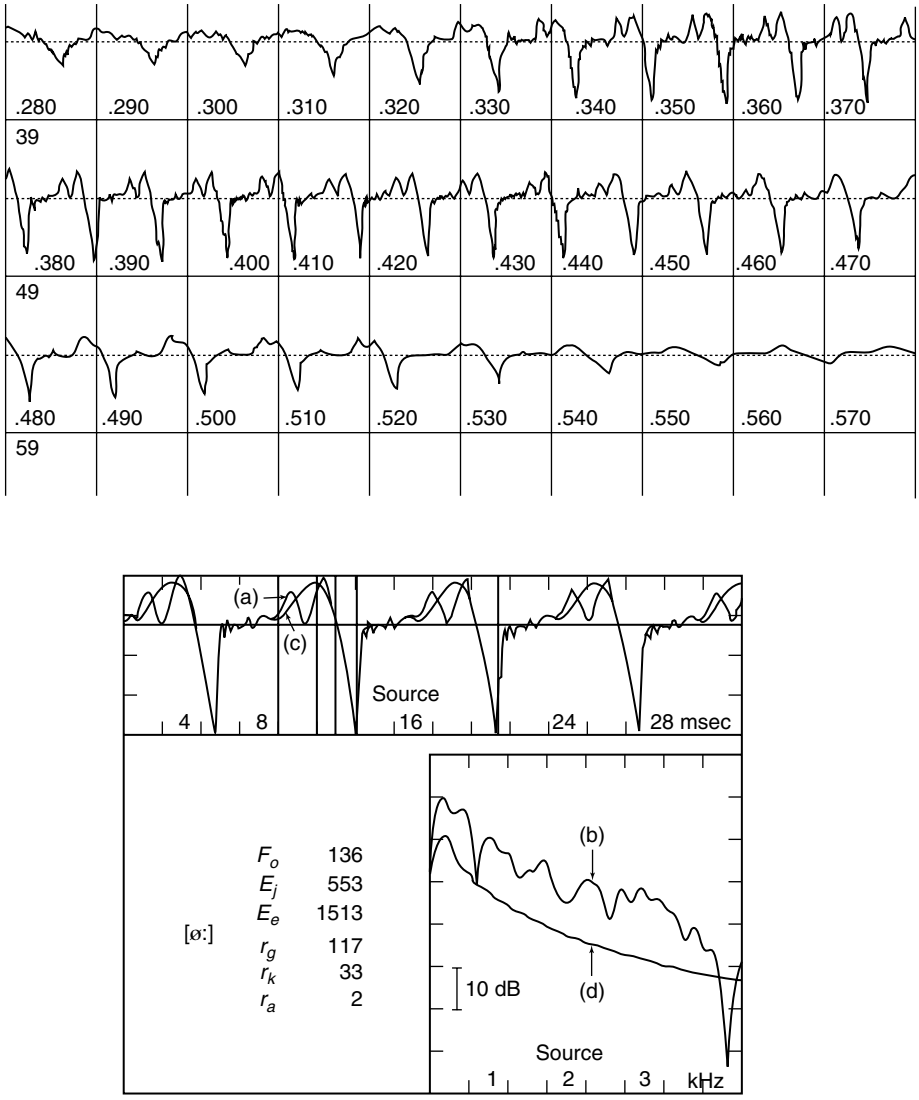


Figure 6. Continuous inverse filtering of the [jø:] part of the word "adjö", spoken by a male subject. The [j] covers the first half of the top line. A routine LF source parameter analysis of a sample of the [[] is illustrated in the lower graph, where the smooth curves pertain to the LF model. From [19].

The inherent intensity of a vowel is not only dependent on its formant pattern but also on its source. An emphatic stress on a syllable containing a vowel produced with a relatively constricted vocal tract, for example, [i:]. [y:]. [ʉ:], and [u:] in Swedish, may cause an additional decrease of the constriction area and, accordingly, a total decrease of source strength [28]. The relation between stress and intensity thus becomes rather complex.

6. READING SPECTROGRAMS. THE SPEECH CODE

A spectrogram, examples of which were given in Figure 1 and Figure 4, provides the potential of a more complex insight into a specific utterance than what can be achieved from any direct set of observations of the production process. With appropriate knowledge, the spectrogram may serve as a window for inferring articulatory activity and organizational principles extending the view all the way up to the brain. It is our basic means of developing speech synthesis strategies and, more generally, of developing knowledge of the speech code [3], which is the relation between messages and sound waves. A speech message is not only confined to a sequence of words, syllables, and phonemes, usually referred to as the segmental structure. A most important part is the prosodic structure, essentially signaled by patterns of F0, relative duration, and intensity. In general, the prosodic pattern may be regarded as superimposed on the segmental structure. Besides grammatical distinctions, as already exemplified in Figure 4, the prosodic pattern carries information about relative emphasis and semantic grouping and about speaker-specific denotations, attitudes and speaking style [2]. A language representative prosody is essential for comprehension and for securing naturalness in speech synthesis. Pathological speech also displays recurrent patterns with their specific codes.

Human speech, as viewed from a spectrogram, is a mixture of continuous and discrete elements. The continuous elements reflect a system of overlaid simultaneous movements of speech articulators with often independent starting and ending points [20]. Coarticulation and reduction account for contextual adjustments of speech patterns and undershoot of targets [3]. The continuous elements are interrupted by discrete breaks in gross pattern type, for example, the switching between voiced and unvoiced patterns and between sound and silence. Both the discrete and the continuous elements are used in segmenting utterances into smaller units, phones, corresponding to phonemes or parts of phonemes. Thus, a stop consonant may be assigned an interval from the offset of sound at the initiation of an occlusion up to the onset of voicing after the release. This total interval may further be subdivided into separate parts, such as occlusion, release transient, frication, and a possible aspiration [20]. In other instances, for example, when two voiced sounds occur in succession (e.g. two vowels or a vowel and a voiced consonant), the segmentation has to rely on continuous formant pattern variations in order to locate a boundary between two presumed target points. This is a problem in establishing routines for segmentation and labeling of speech [21].

The relation between acoustic segments and phonemes is indeed complex. Thus, one phoneme on the message level may influence several successive acoustic segments. Conversely, one acoustic segment is usually influenced by a domain of a message that includes more than one phoneme as well as by specific prosodic attributes, such as stress and phrase junctures. However, an oversimplified but fairly potent rule is that a speech sound may be identified by a specific segment type and by the pattern of formant transitions in and out of the segment. For a fuller account, see [1, 20].

Our knowledge of speech as a code, relating message contents and speaker characteristics to specific sound patterns, is still in a developing phase. Spectrogram

reading exercises serve a most useful purpose of confronting models with reality, a continuous process of learning, revision, and updating of knowledge.

The ultimate level of performance of speech synthesis and recognition will depend on how deeply we can penetrate the speech code including basic mechanisms, human behavior, language structure, and specific variations [3]. Synthesis architecture [22, 23] is fairly well established with respect to basic acoustic theory [4, 11, 24] and principles of digital implementation. [8, 25]. An important trend is to employ articulatory encoding of speech processes [3, 26, 31] instead of formant coding or as an organizational principle. However, to improve the quality of speech synthesis we need a deeper insight into the speech code. Ideally, the computer should have some understanding of what it is going to say.

NOTE

* *From Encyclopedia of Acoustics*, edited by Malcolm J. Crocker ISBN 0-471-8O465-7 c 1997 John Wiley & Sons. Inc. Vol. 4, 1589–1598 (revised version).

REFERENCES

[1] Fant, G. (1968). Analysis and synthesis processes. In, Manual of Phonetics, Chapt. 8, 173–276 (B. Malmberg, ed.). Amsterdam, North-Holland Publ. Co.
[2] Fant, G., Kruckenberg, A. and Nord, L. (1991). Prosodic and segmental speaker variations. Speech Communication 10, 1991, 521–531.
[3] Fant, G. (1991). What can basic research contribute to speech synthesis? Journal of Phonetics, 19, 1991, 75–90.
[4] Fant, G. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. Ericsson Technics 1–1959, 1–106.
[5] Fant, G. (1983). Feature analysis of Swedish vowels—a revisit. STL-QPSR 2–3/1983, 1–19.
[6] Potter, R. K., Kopp, A. G. and Green, H.C. (1947). Visible Speech. D. van Nostrand, New York.
[7] Rabiber, R. L. and Schaefer, R. W. (1978). Digital Processing of Speech Signals. Prentice Hall, Engelwood Cliffs, NJ.
[8] Furui, S. (1989). Digital Speech Processing, Synthesis and Recognition. Marcell Dekker. New York.
[9] Hess, W. (1983). Pitch Determinations of Speech Signals. Algorithms and Devices. Springer Verlag, Berlin.
[10] Markel, J. D. (1971). FFT Pruning. IEEE Trans. Audio Electro-acoust., Vol. AU-19.
[11] Fant, G. (1960). Acoustic theory of speech production. The Hague, Netherlands: Mouton, 2nd edition. 1970, (Translated into Russian, Nauka, Moskva, 1964).
[12] Fant, G. (1961). The acoustics of speech. In, Proc. of the Third Intl. Congress on Acoustics, Stuttgart 1959. (L. Cremer, ed.). Amsterdam 1961, 188–201.
[13] Markel, J. D. and Gray, A. H. (1975). Linear Prediction of Speech. Springer Verlag, Berlin.
[14] Wakita, H. (1982). Linear Prediction of Speech and its Application to Speech Processing. In J.P Haton (ed.) Automatic Speech Analysis and Recognition. Reidel, Boston, 1–20.
[15] Fant, G. (1993). Some problems in voice source analysis. Speech Communication 13, 7–22.
[16] Fant, G., Liljencrants, J., and Lin, Q. (1985). A four-parameter model of glottal flow. STL-QPSR 4/1985, 1–13.
[17] Klatt, D. H. and Klatt, L.C. (1990). Analysis, Synthesis and Perception of Voice Quality Variations among Female and Male Talkers. J. Acoust. Soc. Am, Vol 87, No 2, 820–857.
[18] Fant, G. and Lin, Q. (1988). Frequency domain interpretation and derivation of glottal flow parameters. STL-QPSR 2–3/1988, 1–21.

[19] Gobl, C. (1988). Voice Source Dynamics in Connected Speech. STL-QPSR 1/1988, 123–159.
[20] Fant, G. (1962). Descriptive analysis of the acoustic aspects of speech. LOGOS, Vol 5, No. 1, 3–17.
[21] Carlson, R. and Granström, R. (1986). A Search for Durational Rules in a Real Speech Data Base. Phonetica, Vol. 43, 140–154.
[22] Carlson, R., Granström, B. and Hunnicutt, S. (1991). Multilingual Text-to-Speech Development and Applications. In A.W. Ainsworth (ed.), Advances in Speech Hearing and Language Processing. JAI Press, London.
[23] Klatt, D. H. (1980). Software for a Cascade/Parallel Formant Synthesizer. J. Acoust. Soc. Am. Vol 67. 971–995.
[24] Flanagan, J. L. (1972). Speech Analysis Synthesis and Perception. Springer Verlag, New York.
[25] Gold, B. and Rabiner, L.R. (1968). Analysis of Digital and Analog Formant Synthesizers. IEEE Trans. Audio Electro-acoustics, Vol Au-16, No. 1, 81–94.
[26] Lin, Q. (1990). Speech Production Theory and Articulatory Speech Synthesis. Dr. Sc. Thesis, Dept. of Speech Communication and Music Acoustics, KTH, Stockholm.
[27] Båvegård, M. and Fant, G. (1994). Notes on glottal source interaction ripple. STL-QPSR 4/1994, 63–78.
[28] Fant G. (1997). The voice source in connected speech. Speech Communication 22, 125–139.
[29] Fant G. (1995). The LF-model revisited. Transformations and frequency domain analysis. STL-QPSR 2–3/1995 119–155.
[30] Fant, G. (1975). Non-uniform vowel normalization. STL-QPSR 2–3/1975, 1–19.
[31] Stevens, K. N. and Bickley, C. A. (1991).Constraints Among Parameters Simplify the Control of Klatt Formant Synthesizer. J. Phonetics, Vol 19, No. 1, 161–174.

CHAPTER 4.3

FEATURES: FICTION AND FACTS

INTRODUCTION

These are some personal views on distinctive feature analysis acquired during the course of more than thirty years since my early cooperation with Roman Jakobson and Morris Halle. Most of what I have to say is neither new nor radical and has been expressed in more detail in earlier publications (e.g., Fant, 1973), containing several articles on the subject (e.g., Fant, 1969, 1970). A recent article (Fant, 1983) on Swedish vowels provides additional perspectives. The following is a summary and an attempt to relate feature theory to the topic of this volume.

Distinctive-feature theory has two main purposes. One is to develop a language-universal system of phonetic categories selected to serve phonological classificatory functions. The other is to describe essentials of the speech code, i.e., distinctive dimensionalities and mechanisms of encoding within the speech chain.

It was the great undertaking of Roman Jakobson to attempt to unify these two objectives within the same theoretical frame. Jakobson's influence has been immense. He insisted on absorbing all possible background information from spoken language and research in speech production, acoustics, and perception for this purpose. Still, he was aware of the difficulties of the task and it was he who coined the term "preliminaries" for our joint work. He was always open to new suggestions and could accept new classificatory solutions given new evidence.

Yet there remains in my view an inherent difficulty in reaching a language-universal solution optimal for both purposes. The mechanisms and codes are a unique product of human physiology and, thus, universal while the specific choice of classificatory features for describing a specific language becomes highly dependent on the investigator's phonetic background, the terminology he or she is accustomed to, and to attempts to optimize the feature inventory to simplify descriptions of language structure and usage. One difficulty is that one and the same physical fact may be described in so many different ways. From vowel theory we know that a parametric description in terms of F_1 and F_2 by rotation of coordinates is equivalent to a description in terms of $F_2 - F_1$ and $F_1 + F_2$ which calls for a change in feature labels.

The supposedly happy marriage between phonology and phonetics has its inherent shortcomings and some of us like Peter Ladefoged might argue for a respectful divorce. I am skeptical about the task of formulating a unique ultimate set of classificatory features appropriate for all languages of the world but I do believe in a continued search for insight into universal speech and hearing mechanisms and dimensions and their distinctive role within specific languages. This should at least be our immediate goal.

Jakobson's work has paved the way for the development of the field. His idea of applying tonality features and compactness to consonants as well as to vowels (Jakobson et. al., 1952) relates to basic acoustic and auditory dimensions, such as

whether a sound is dominated by high- or low-frequency components and if the energy is concentrated or distributed in the frequency domain. That the auditory system preserves such dimensions is apparent from recent neurophysiological findings (e.g., Sachs, Young & Miller, 1982). This parallelism between consonant and vowel dimensionalities focuses on the output of the speech chain and, thus, on properties essential to the code. The Chomsky and Halle (1968b) approach, on the one hand, is to define features from the production level and not to worry too much about their acoustic and perceptual integrity.

Both systems have their drawbacks. The parallelism within the Jakobson system is not a sufficient basis for all consonant categories and the articulatory base of the Chomsky and Halle system does not explain the code.

FEATURES IN THEORY AND PRACTICE

The stable parts of theory are our real insight in language structure and usage and in the production, acoustics, and perception of speech. This is by no means complete but our knowledge can be tested, updated, and expanded from a solid basis. To me, phonetics is the stable partner of the marriage, while phonology is promiscuous in its experimenting with widely different frameworks and choice of features for describing one and the same inherent phenomenon. I have myself contributed to the phonological diversity by proposing several alternative solutions to the classification of Swedish vowels.

In a pragmatic sense, it does not matter which framework of classificatory features you adopt as long as you know how to handle them phonetically. A column within a feature matrix must first of all serve as an appropriate address for finding a specific phoneme or class of phonemes. In addition, there is the requirement of a maximally simple and direct relation to phonetic events and their ordering in parameter space and time. As an example of pragmatism, Carlson and Granström adopted parts of the Chomsky and Halle feature system for our text-to-speech synthesis program. Labials are accordingly classified as [+anterior], [−coronal] but the program does not have an independent mode of realization of each of these categories. It is the combined presence which triggers labiality. I am, for reasons of principle, against such encoding of phonetically autonomous dimensions by a combination of totally unrelated dimensions. Another example within the Chomsky and Halle system is the minimal distinction of /r/ being [−anterior] and /l/ being [+anterior]. Within the Jakobson, Fant & Halle (1952) system the phoneme /h/ is encoded as [−consonantal], [−vocalic] which I have supported by the roundabout argument that /h/ lacks or has rather weak consonantal attributes (no significant zeros, lack of F-pattern contrast) and is less vowellike than adjacent vowels.

In the Chomsky and Halle (1968b) system, the phoneme /h/ in addition to the [−consonantal], [−vocalic] base is encoded as [+back], [−front] in distinction to the glide /y/ which is labeled [−back], [+front]. This solution obscures the role of the tongue body features which relates to the place of articulation for the /y/ but to

the place of the source for /h/ which, of course, can have the same or similar tongue articulation as /y/.

Thus, the hunt for maximum economy often leads to solutions that impair the phonetic reality of features. As a reaction, we could employ traditional phonetic terminology in a feature matrix and abolish the use of minus classifications. Still, we maintain a minimum-average redundancy in the outcome by, on the average, 3 plus-sections needed per phoneme (Fant, 1969). A second stage adding contextual constraints will reduce possible combinations, which is an important consideration in automatic speech recognition.

A simple one-to-one relation between phonetic events and phonological entities is exceptional. Speech segmentation is frequently ambiguous and expected events may be missing. A string of several successive phonetic events or segments may carry information about a single phonological segment or distinctive feature, and the converse, a many-to-one relation, introduces a complex set of conditional factors affecting a single phonetic segment or a parameter or cues.

One way out of this dilemma is to develop improved models of speech production which make us more aware of temporal continuities of speech parameters and their constraints induced by physiological, linguistic, situational, and personal factors. Such a view of speech, as a continuous complex vector in space and time anchored in production, must be followed up on the level of the speech wave and speech perception. It will free us from the bonds of segmenting the speech wave with maximal precision prior to recognition, but it will require a more profound insight in the perceptual relevance of rapidly varying speech-vector states.

The articulatory framework should be more or less the same for vowels and consonants, i.e., contain a specification of both main tongue-body configuration and place of articulation which need not coincide exactly with the place of maximum narrowing. Consonants, thus, have an added main tongue-body feature, as introduced by Chomsky and Halle (1968b), and vowels can employ consonantal elements. An example is the extreme palatal narrowing found in the long (tense) Swedish [i:] or [y:] when stressed, while the Swedish long tense [ʉ:] in addition to its extreme narrow lip opening is produced with an apical elevation which advances the place of minimum cross-sectional area anteriorly to that of the vowel [i:] or [y:]. In R-colored vowels, on the other hand, the consonantal modification is a phonologically distinctive element.

The coarticulation model of Öhman (1967a) with separate inputs for vowel and consonant commands is an appropriate starting point for articulatory analysis. His notion (Öhman, 1967b) of redistribution of physiological energy as a result of varying prosodic patterns could be extended to define a dimension affecting scale values of targets and, furthermore, the duration and intensity of phonetic segments and vowel-consonant contrasts. The position of a syllable within a sentence, lexical stress, emphasis and deemphasis, and citation forms versus connected speech will all condition such an energy or emphasis factor. It is also related to the tense/lax dimension and has to be followed up by intrasyllabic relations.

This is a suggestion for how to organize a search for rules that will predict speech variability.

ARTICULATORY VERSUS ACOUSTIC PERCEPTUAL ANCHORING: VOWELS

Should distinctive features accordingly be anchored in the articulatory domain? There are examples which point at a single, well-defined, articulatory event conditioning several acoustic cues. Glottal adduction appropriate for voicing not only facilitates the voicing prior to the release of a voiced stop but reduces also the voice onset time (VOT) and lowers the starting point of F_1 in the following vowel. The duration of a previous vowel may also be increased, though marginally only, except when the unvoiced member, in addition, is aspirated (preocclusion aspiration). However, all these cues have one and the same acoustic effect, that of increasing the amount of and continuity of low-frequency energy. The voiced/voiceless, lax/tense, and unaspirated/aspirated distinctions have developed a symbiosis in many languages, like Swedish and English. In my view, it would be against the principle of phonetic reality to consider only one of these features to carry the /g, b, d/ versus the /k, p, t/ distinction.

On the other hand, to back up an output-oriented view, there are examples of a diversity of articulations appropriate for one and the same phoneme, e.g. /r/. In a recent article on Swedish vowels (Fant, 1983), I point out that the short [ɵ] and long [ʉ:], spelled /u/, differ drastically in terms of articulation. The [ɵ] is close to a back vowel [o], while [ʉ:] is not less fronted than [i:]. Acoustically, they are also further apart than any two long/short members of a pair, but they still occupy one and the same subspace (see Figure 22.1). In relation to other rounded front vowels, they are extra "flat," i.e., they have a lower $F_1 + F_2'$ than [ø:] and [y:].

Complex vowels systems as in Swedish are a challenge to the binary principle since vowels eventually attain mutually adjusted positions within an acoustic space of F_1, F_2', and F_3 or F_1 and F_2 with a tendency of regular spacing in terms of auditory measures, as the Bark unit. Still, I feel it is appropriate to analyze the system in terms of binary distinctions. The basic feature of gravity with the correlate of small $F_2 - F_1$ separating back vowels from front or centralized vowels has a perceptual relevance. Thus, starting out from the vowel [o] and introducing a small forward shift of the tongue to bring F_2 to a location more than 3.5 Bark away from F_1, transcending the boundary beyond which F_1 and F_2 are processed centrally as separate formants, produces a vowel which is perceived as [ɵ]. This is in accordance with the experiments of Chistovich and Lublinskaya (1979). When two formants come closer than about 3.5 Bark they are perceived in terms of a weighted mean value which is the case for grave vowels.

These findings add support to Stevens' (1969) quantal theory. I could also add that the acoustics of speech production acts as a stabilizer for the extreme grave and the extreme nongrave vowels. Once F_2 and F_1 or F_3 and F_2 come closer than about 3 Bark, the distance will change but little with a more extreme back or extreme front articulation. Moreover, the gravity measure will be insensitive to moderate changes in the place of articulation.

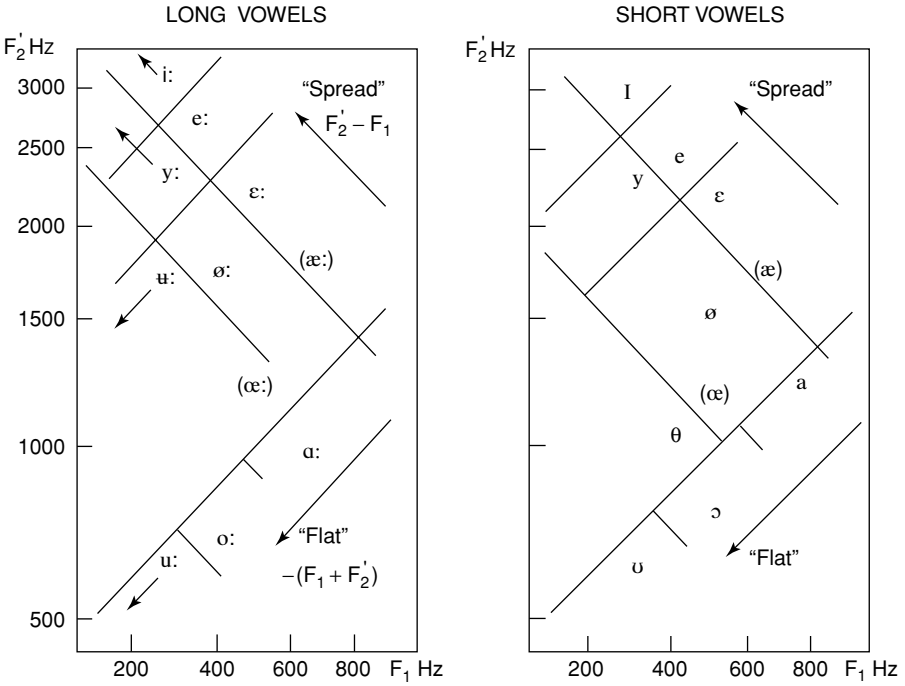


Figure 22.1. Swedish long and short vowels within a F_2' versus F_1 frame with auditory (Bark) adjustment of frequency scales. The [æ] and [œ] are pre-r allephones of /ɛ/ and /ø/. Observe the relational similarity within the two groups of vowels. The +45 degree slanting line above the back vowels corresponds to the critical distance $F_2-F_1 = 3.5$ Bark.

Tranmüller (1983) has found that females and males preserve one and the same F_3-F_2 for front vowels. This is mainly a consequence of linear scaling of vocal-tract dimensions. A nonuniform difference, such as tha of the relatively shorter female pharynx, would not counteract this trend. In grave vowels, the energy within the F_2 and F_1 modes are fairly equally divided between the front and back of the vocal tract. This is a requirement for proximity. The same is true of F_3 and F_2 of front vowels. The vowel [i] is an exception from this rule since F_2 of [i] has a distinct back cavity and F_3 a front-cavity affiliation. However, the F_3-F_2 measure of [i] is greater than that of other front vowels. The vowel [i] produced with a relatively long pharynx, as in male speech, is characterized by a F_3 closer to F_4 than to F_2. An additional stabilizer of F_1 of the vowel [i] is that it can never be lower than the closed tract resonance frequency of about 200 Hz as set by the wall impedance. These stability phenomena have been mentioned by Stevens (1969).

Other discrete effects within the Swedish vowel system is the diphthongal gesture with extreme articulatory narrowing, towards a homorganic consonant, i.e., [i:→j], [u: → β].

ABSOLUTE VERSUS RELATIONAL INVARIANCE

A fundamental concept in distinctive feature theory, as outlined by Jakobson et al. (1952), is that of relational invariance. The critical scale value of a parameter at which a feature shifts from plus to minus is conditioned by the specific context of preceding, present, and following features within an area of the distinctive feature matrix and, of course, by prosodic, individual, and situational factors. The minimum requirement is the presence of a vectorial component along a specific distinctive parameter dimension. A compactness distinction within vowels should therefore, ceteris paribus, always involve a higher F_1 for the compact member. In my 1983 study of Swedish vowels, I have chosen the opposite of compactness, i.e., diffuseness to separate /i/ and /e/ from /ɛ/. The long Swedish vowels [i:] and [e:] may differ rather little in F_1 whilst the higher F_2' is the correlate of the sharpness feature assigned to [i:].

Once the acoustic vowel space is normalized with respect to individual and contextual factors, the operation to select a specific vowel may be broken down into a succession of absolute binary decisions in terms of delimiting boundary lines. The sequence of operations is that implied by a coding tree which in my suggestion for Swedish vowels involves [grave], [flat], [extra flat], [diffuse], [sharp] in addition to the long/short distinction (see Table 22.1 and Figure 22.2).

A few words could be said about the marking of redundancies in distinctive feature matrices. I indicate by parenthesis predictabilities from features located lower down in the matrix or in the associated coding tree. Thus, [+extra flat] implies [+flat], and [+sharp] implies [+diffuse]. I use blanks for features that are bypassed in the coding tree, i.e., which are irrelevant or implied by prior branches. Thus, [−flat] implies [−extra flat] and [diffuse] implies [−sharp].

The search for absolute invariance of feature correlates irrespective of context, as pursued by K. N. Stevens, is a challenge (Blumstein & Stevens, 1979. 1980; Stevens, 1980). I have especially in mind his discussion of the correlates of place of articulation of stops. Even though a single spectral section at the early part of a stop release may retain fairly stable "acute," "grave," or "compact" attributes, the specification becomes more precise if we include contextual elements. However, since Stevens allows' for temporal contrast effects, one could systematically allow the term "absolute" to mean without reference to phonological identity of the context.

TABLE 22.1
Binary Feature Matrix

Parameter	Feature	[u:] [u]	[o:] [ɔ]	[*a*:] [a]	[ɛ:] [ɛ]	[e:] [e]	[i:] [i]	[y:] [y]	[ʉ:] [ɵ]	[ø:] [ø]
$-(F_2-F_1)$	grave	+	+	+	−	−	−	−	−	−
$-(F_2'-F_1)$	flat	(+)	+	−	−	−	−	+	(+)	+
$-(F_2'-F_1)$	extra flat	+	−					−	+	−
$F_2'-F_1$	diffuse				−	+	(+)	(+)		
$F_2'-F_1$	sharp					−	+	+		−

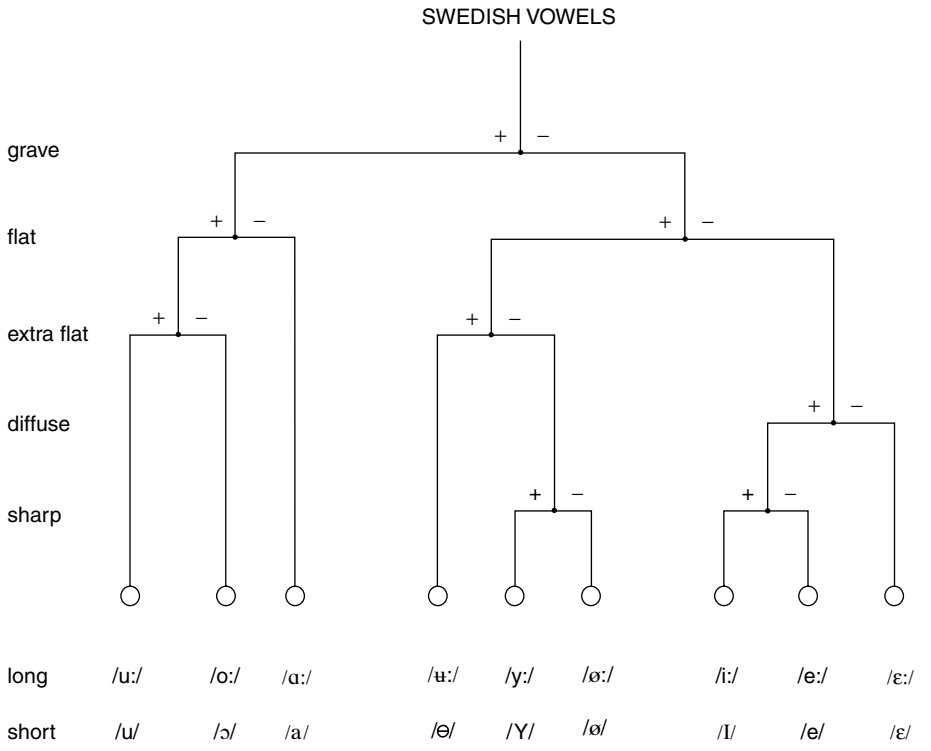


Figure 22.2. Distinctive feature coding tree of Swedish vowels.

This view can then incorporate an integration of cues within a finite window extending sufficiently far in time before and after the stop release to catch transitional phenomena.

The common denominator of an ensemble of velar and palatal stops produced with various degrees of lip rounding is the energy concentration in the F_2' range of the following vowel. In addition to the element of spectral concentration, we thus retain the requirement of a location of the concentration with reference to the F-pattern without having to make a decision about the context. Similarly, the positive going F_2, F_3 versus time transition of labials adds important characteristics to the relative low-frequency dominant spectrum section at the release.

In my view, human speech perception relies on gestalt decoding rather than on isolated short-time spectral patterns or templates. The interdependency of the several cues found within the domain of a single distinction is considerable, as found in many experiments at the Haskins Laboratories. Experience from speech synthesis also indicates that when listeners are confronted with stimuli in a boundary region between phonemes, they may react with different weights to the variation of some of the cues or parameters entering the feature complex. This also speaks for the importance of the complete gestalt or of major cues within the gestalt. The auditory

system probably makes efficient use of the entire evidence available. Why should we limit our descriptive work to less precise specifications or to a diluted specification which can operate in all contexts?

Whether this gestalt is anchored in the listener's internalized concept of what a vocal tract can do (Dorman, Lawrence & Liberman, 1979), or whether it is performed by virtue of an experience about auditory patterns in speech is more a matter of preference of verbal phrasing than a real difference.

Sheila E. Blumstein: Comment

I would like to comment on the last point made in the Fant paper. Fant states, "Whether this gestalt is anchored in the listener's internalized concept of what a vocal tract can do . . . or whether it is formed by virtue of an experience about auditory patterns in speech is more a matter of preference of verbal phrasing than a real difference." One of the primary goals in looking for invariance in speech is to determine how we process speech and language, and to determine what the underlying mechanisms are for such processing. It is not a matter of preference in choosing one anchor over another. It is in effect an empirical issue and a very critical one. If the anchor is expressed in terms of the auditory domain, in terms of the articulatory domain, or in terms of some mental representation mediating the two, then the nature of the mechanisms subserving the use of speech and language will be quite different, as will our models of speech/language processing. For example, the analysis by synthesis model of speech was developed largely because of a failure to find invariant acoustic patterns corresponding to phonetic features. It may turn out that there are indeed invariant auditory patterns as well as invariant vocal tract configurations subserving a common mental representation. However, these are issues which can only be resolved empirically, and which will have critical theoretical consequences.

Nelson Kiang: Comment

In giving us his personal views on distinctive feature analysis, Fant expresses skepticism about basing a complete view of human speech perception on isolated, short-time spectral patterns or templates. He seems to encourage further study of physiological mechanisms as one important direction for future work. As a physiologist, I would like to comment on the outlook for such an approach.

Ultimately, the descriptions of how we process speech or any other auditory stimulus will come from physiology. Unfortunately, we do not know the physiological mechanisms for perceptual constancy which would be fundamental for providing definitive answers to the questions raised in this volume. It is easy to sketch plausible schemes for any well-designed operation but this is different from knowing the actual mechanisms.

After attending this symposium for three days, I am reminded of the debates of the Chinese Naturalists over two thousand years ago about how to account for the

properties of material things. Their solution led to the five element theory, in which all the qualities of physical things could be explained as combinations of water, fire, wood, metal, and earth. This theory appeared to be satisfactory in accounting for the diversity of identifiable things as well as differences between specific samples of a type. Today, most modern scientists believe that the building blocks of things are at the atomic and molecular levels, and that invariances in material things are based on similarities in structure at lower levels.

Similarly, the question of what underlies the concept of "species" raised problems for classical biologists. A dog was a dog even though there could be many different kinds of dogs. All attempts to define individual species in terms of key characteristics tended to have an ad hoc quality. Today, most biologists are convinced that the invariances that define species occur at the level of DNA molecules which constitute the building plans for organisms. Certain differences in these molecules will correspond to variations within a species, whereas others will preclude the development of fertile offspring thereby creating distinct species.

In both these examples, the problem is how to account for invariances that define categories and still recognize permissible variations within a category. In each case, the resolution came from a reductionistic approach. Even though the number of cases for which a full synthetic demonstration is available is small, the presently accepted solutions appear satisfactory.

I submit that the scarch for the bases of invariance in speech perception and production at either acoustic or phonological levels is probably doomed to failure because the invariance must be at neural levels. In taking this view, I do not wish to denigrate the valuable work of the type presented in this volume. Such work will provide factual evidence that will have to be accounted for in any acceptable theory. My present view is that no reasonably realistic neurophysiological formulation to explain speech processing and generation has been proposed at a sufficiently interesting level to warrant detailed analysis. As data from direct studies of the human brain in action become commonplace, the issues raised in this volume will become increasingly relevant to cognitive neuroscientists who will in fact have to account for "gestalt decoding" in terms of specific neural mechanisms. I urge speech scientists to interact with such workers and to insist that their students acquire some familiarity with neurophysiological concepts. By such means, speech studies will receive new revitalizing influences and Fant's cautious optimism will be justified.

Gunnar Fant: Response to Kiang

As speech scientists, we have a tendency to project our preestablished models of language on our models of brain functions, but we are eager to follow the advance of neurophysiology of speech and hearing to learn more and revise our working models. Meanwhile, we resort to available evidence from other stages within the speech chain to fill in the gaps in an overall theory of speech coding and decoding.

Alvin M. Liberman: Comment

My concern is first with invariance only in the conversion from sound to phonetic structure, and then with the facts that such invariance ought, in my view, to take into account.

Because of the way we speak, the acoustic information for a phonetic segment commonly comprises a large number and wide variety of cues, most of them dynamic in form. These cues span a considerable stretch of sound, grossly overlap the cues for other segments, and are subject to a considerable amount of context-conditioned variation.

The phonetic perceiving system is sensitive to all the acoustic cues. None of these cues is truly necessary; all are normally used; and their relative importance bears little relation to their salience as it might be reckoned on a purely auditory basis.

Perception of phonetic structure is immediate in the sense that there is no conscious mediation by, or translation from, an auditory base. Generally, listeners are only aware of the coherent phonetic structure that the cues convey, not of the quite different auditory appearances the cues might be expected to have, given their overlap, context-conditioned variation, number, diversity, and dynamic nature. Thus, taking stop consonants and their dynamic formant-transition cues as a particular example, I note that listeners are not aware of the transitions as pitch glides (or chirps) and also as (support for) a stop consonant; listeners are only aware of the stop. Yet these same formant transitions *are* perceived as pitch glides (or chirps) when—on the nonspeech side of a duplex percept, for example—they do not figure in perception of a phonetic segment.

These facts have two implications. One is that the invariance between sound and phonetic structure should be sought in a general relation between them that is systematic but special, not in particular connections that are occasional and discrete. The relation can be seen to be systematic to the extent it is governed by lawful dependencies among articulatory movements, vocal-tract shapes, and sounds, dependencies that hold for all phonetically relevant behavior, not just for specific and fixed sets of elements. The relation has got to be special because the vocal tract and its organs are special structures that behave, most obviously in coarticulation, in special ways. A second implication is that the special relation between sound and phonetic structure is acted on in perception by a system that is appropriately specialized for the purpose.

If the foregoing assumptions are correct, then the invariance in speech is not unique. Rather it resembles, at least grossly, the kinds of special invariances that are found in many perceptual domains. Accordingly, the system that is specialized for phonetic perception can be seen as one of a class of similarly specialized biological devices. All take advantage of a systematic but special invariance between the "proximal" stimuli and some property of the "distal" object. The result is immediate perception of the properties that make it possible to identify the invariant distal object.

Consider visual perception of depth as determined by the proximal cue of binocular disparity. There is a general and systematic, yet special, relation between the

distal property (relative distance of points in space) and the proximal stimulus (disparity). The relation is general and systematic in that it is governed by the laws of optical geometry and holds for all points (within its range) and for all objects, not just for some. The relation is special because it depends on the special circumstances that we have two eyes, that they are so positioned (and controlled) as to be able to see the same object, and that they are separated by a particular distance. Neurobiological investigation has revealed an anatomically and physiologically coherent system that is specialized to process the proximal disparity and relate it to the distal depth. Given that specialization, perception of depth is automatic and immediate; there is no conscious mediation by, or translation from, the double images we would see if, in fact, we were perceiving the proximal disparity as well as the distal property it specifies.

Other perceptual phenomena have the same general characteristics. Auditory localization and the various constancies come immediately to mind, and, if we put aside questions about phenomenal "immediacy," so too do such processes as those that underlie echolocation in bats and song in birds. These are surely specializations if only because each such process, or module, is as different from every other as is the invariant relation it serves. The phonetic module differs from many of the others in at least two ways.

To make one of the differences clear, I would turn again to binocular disparity and depth perception as representative of a large class. In this case the distal object is "out there," a physical thing in the narrow sense of the word "physical," and the invariant relation between its properties and those of the proximal stimulus is determined, as already indicated, by optical geometry and the separation of the eyes. In speech, however, the distal object—a phonetic structure—is a physiological thing, a neural process in the talker's brain, and the invariant relation between its properties and those of the proximal sound is determined in large part by neuromuscular processes internal to the talker but available also to the listener. Thus, the specialized phonetic module might be expected to incorporate a biologically based link between production and perception. Such a link is not part of the disparity module or of the other perceiving modules it exemplifies, though it may very well characterize the "song center" module of certain birds.

A second important difference in the nature of the invariance (and its module) has to do with the question: What turns the module on? In the case of binocular disparity, the answer is a quite specific characteristic of the proximal stimulus—namely, disparity. Notice, however, that disparity has no other utility for the perceiver but to provide information about the distal property, depth. There are, accordingly, no circumstances in which the perceiver could use the proximal disparity as a specification of, or signal for, some other property. Disparity and the depth it conveys do not compete with other aspects of visual perception such as hue or form but complement them. Not so in phonetic perception. There is, first of all, the fact that the speech frequencies overlap those of nonspeech. More to the point, the formant transitions that we don't want to perceive as chirps when we are listening to speech are very similar to stimuli that we do want to perceive as chirps when we are listening to birds. Thus, almost any single aspect of the proximal stimuli can be used for

perception of radically different distal objects: phonetic structures in a talker's head or acoustic events and objects in the outside world. The module can hardly be turned on by some specific (acoustic) property of the proximal stimulus. Not surprisingly, then, we find in research on speech perception that the module is not turned on that way, but rather by some more global property of the sound. Thus, just as in the perception of phonetic segments all cues are responded to but none is necessary, so too in identifying sound as speech.

How, then, is the module turned on? What invariant property of the sound causes the listener to perceive that the distal object is a phonetic structure and not some nonlinguistic object or event? I offer a suggestion. Suppose that auditory stimuli go everywhere in the nervous system that auditory stimuli can go, including the language center. The language center has to answer the question: Could these sounds, taken quite abstractly, have been produced by linguistically significant articulatory maneuvers, also taken quite abstractly? If the answer is yes, then the module takes over the purely phonetic aspects of the percept, and the auditory appearances are inhibited. (Auditory aspects that are irrelevant to the phonetic, such as loudness or hoarseness, are perceived as attributes of the same distal object.) If the answer is no, then the phonetic module shuts down and the ordinary auditory appearances of the stimuli are perceived. Hence the common experience of those who work with synthetic speech that when the sound includes configurations that the articulatory organs cannot produce as well as those it can, the percept breaks, correspondingly, into nonspeech and speech. Phenomenally, the nonspeech stands entirely apart from, and bears no apparent relation to, the speech, even though the acoustic bases for these wholly distinct percepts were perfectly continuous. The same arrangement for turning the module on (or off) might account for the fact that certain kinds of acoustic patterns—for example, sine waves in place of formants—can be perceived as speech or as nonspeech depending on circumstances that in no way alter the acoustic structure of the stimulus. It also helps to explain how, as in the unnatural procedures of duplex perception, we can disable the mechanism that forces the choice between speech and nonspeech, and so create a situation in which exactly the same proximal formant transition is simultaneously perceived (in the same context and by the same brain) as critical support for a stop consonant and also as a nonspeech chirp. At all events, there is a kind of competition between phonetic perception and other ways of perceiving sound. A consequence is that the phonetic module produces a more or less distinct mode of perception in a way that modules like depth perception do not. This phonetic mode accommodates a class of distal objects that are distinguished, not only by their role in language, but also by the special nature of the invariant relation by which they are connected to sound.

REFERENCES

Blumstein, S.E., & Stevens, K.N. (1970). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustic Society of America*, 66(4), 1001–1017.

Blumstein, S.E., & Stevens, K.N. (1980). Perceptual invariance and onset spectra for stop consonants in various vowel environments. *Journal of the Acoustic Society of America*, 67, 648–662.

Chistovich, L. A., Lublinskaya, V.V., Malinnekova, T. G., Ogorodinikova, E. A., Stoljarova, E. L., & Zhykov, S.J.A. (1982) Temporal Processing of Peripheral Auditory Patterns. In R. Carlson & B. Granström (Eds) *The representation of speech in the peripheral auditory system*. Amsterdam: Elsivier/North-Holland Biomedical Press.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper-Row.

Chomsky, N. & Halle, M. (1968b). The sound pattern of silence in phonetic perception. *Journal of the Acoustic Society of America*. 65(6), 1518–1532.

Fant, G. (1969). Distinctive features and phonetic dimensions, *Speech Transmission Laboratory Quarterly Progress and Status Report*, 2–3/1969, 1–18, Royal Institute of Technology Stockholm.

Fant, G. (1973). *Speech Sounds and Features*. Cambridge, MA: MIT Press.

Fant, G. (1983). Feature analysis of Swedish vowels—a revisit. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 2–3/1983, 1–18, Royal Institute of Technology Stockholm.

Sachs, M.B., Young, E.D. & Miller, M.I. (1982). Encoding of speech features in the auditory nerve. In R. Carlson & B. 'granström (eds), *The representation of speech in the peripheral auditory system*. Amsterdam: Elsevier Biomedical.

Stevens, K.N. (1972) The quantal nature of speech. Evidence from articulatory-acoustic data, In E.E. David, Jr & P.B Denes (Eds), *Human communication: A unified view*: New York: McGraw-Hill.

Stevens, K.N. (1980). Acoustic corrlates of some phonetic categories. *Journal of the Acoustic Society of America*, 68(39), 836–842.

Traunmuller, H. (1983) Articulatory and perceptual factors controlling the age and sex conditioned variability in formant frequencies of vowels. To be published in *Speech Communication*.

Öhman, S. (1967). Word and sentence intonation: A quantitative model. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 2–3/1967, 20–54. Royal Institute of Technology, Stockholm.

CHAPTER 4.4

ON THE SPEECH CODE

1. INTRODUCTION

Speech technology has provided important tools for applications in man-machine communication systems and is growing rapidly. But there is a risk that expansion will be limited by insufficient attention to the potentialities of speech and language research. The symbiosis between technology and basic research that has made possible the advance, now shows a tendency to turn into polarisation. At present, speech technology is highly dependent on statistical tools and large data bases, whilst phonetic research tends to focus on narrowly defined problems or abstract issues with small or no relevance for the overall code of spoken language.

The term speech code has been coined to represent the knowledge of how linguistically defined units are realised in the speech act. In this sense it conforms with general phonetics. A more precise representation of the speech code is in programs for text-to-speech synthesis and the reverse, strategies for speech recognition from a phonetic analysis of speech wave patterns.

At present we are doing fairly well in text-to-speech synthesis, but quality and prosody could certainly be improved. Automatic speech recognition is a much more difficult task. Although there has been a steady advance we have not reached a level of performance for reliable hands-free operation. At present, visual communication is more secure and has gained a priority in mobile telephony. Long time investments in basic research are needed. Around the corner, in a view towards the future, even more demanding and spectacular applications are anticipated, such as language and voice translation over the telephone. Increased investments in basic speech research should also be motivated by the obvious gain in linguistics and phonetics and in a vast field of human speech communication.

2. THE CONCEPT OF THE SPEECH CODE

The notion of a code implies relations between message units and signal units. In speech communication, we are concerned with the relation of language units to units of the speech act. This is the speech code, also referred to as the phonetic code. How are units of spoken messages encoded in the acoustic signal that radiates from a speaker's mouth and reaches our ears to be heard and understood?

We may extend this notion of the speech code to relations within a chain of successive encoding stages within the processes of speech production and speech perception.

Our main reference for the speech signal is the speech wave as observed from oscillographic and spectrographic records. These have the potential advantage of preserving maximally complete specifications of the physical aspect of speech. Of course, such records lack information that the speakers convey through facial

expressions and other aspects of body language which constitute a sub-code of complementary and socially acknowledged and patterns.

Studies of underlying articulatory processes cannot provide as complete and exact descriptions as the speech wave analysis. However, speech wave patterns are inherently complex and need to be interpreted by reference to possible underlying articulatory patterns and perceptual constraints. The encoding rules within the speech chain, e.g., relating articulatory movements and resonator configurations to speech wave patterns, have become important constituents of the overall code (Fant, 1960; 1962, 1980, 1989, 1991). Articulatory interpretation of spectrograms is a key to the understanding of the speech code.

The speech chain model becomes more illusive when applied to brain functions. We loose insight in encoding mechanisms and signal structure. We are left with partial insights that can serve as basis for speculations about general aspects of the code. The same is true of auditory functions and speech perception. The close ties observed between articulation and language units have influenced theories of coordinated speech production and speech perception mechanisms, e.g., the motor theory of speech perception developed by Alvin Liberman (Liberman and Mattingly, 1985). However, we do possess a substantial amount of general knowledge of auditory mechanisms and speech perception, such as various aspects of frequency selectivity, critical bands of hearing, time constants, masking and adaptation that can aid an auditory adapted speech wave analysis (Fant,1967, 1978; Carlson and Granström, 1982). Auditory constraints potentially allow a substantial data reduction of sampled speech wave patterns. In addition, a general knowledge of phonetically relevant auditory patterns can contribute to the development of a less complex and more precise specification of the code.

When discussing the speech code, we are apparently faced with two different communication situations. One employs a human listener. The other employs a human or a computer attempting to read and decipher a spectrographic representation of an unknown text. The strategies we develop for the latter tasks may only in part reflect the true speech code, but they are of great technical importance. The basic strategy, as already mentioned, is to ask ourselves "How has this been produced?" and "How can this be perceived?". In both cases, the listener's or the spectrogram reading person's expectancy becomes an important part of the decoding process. We hear what we expect to hear and sometimes nothing else.

Lindblom (1989) has pointed out the trading relations between distinctiveness and redundancy. Even so-called clear and careful speech often displays omissions and substitutions on the phonological level as well as far-going reductions of speech patterns. We encounter a continuity of reductions from complete forms to a total extinction of information-bearing elements of a phoneme. Elision is ignored in top down expectancy.

To the speech code also belong the specifics of individual voice types and speaking and reading style as well as deviant speech behaviour. *Any consistent regularities that can be subjected to a descriptive analysis may be incorporated in the speech code*.

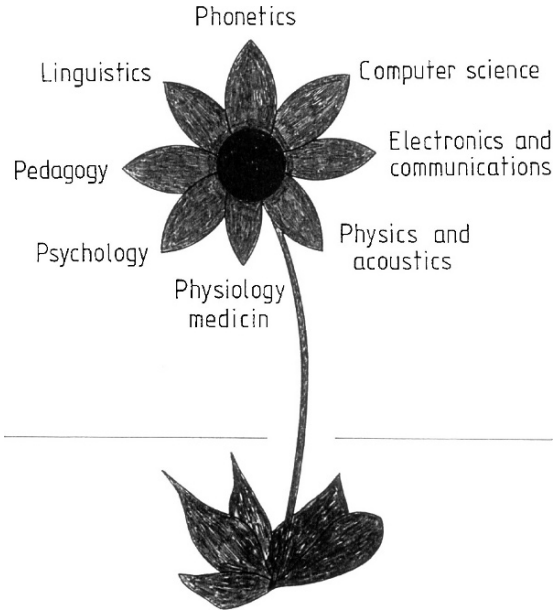


Figure 4.4.1. Areas of speech communication sciences.

3. APPLICATIONS

By now, it is evident that the concept of the speech code is an integral part of general and experimental phonetics and speech research. Figure 1 symbolises the many branches of speech communication that are concerned: Communication engineering, acoustics, linguistics, phonetics, psychology, physiology, speech, hearing research, and in many applications such as, language teaching, handicap aids and rehabilitation.

Speech technology has supplied us with talking computers that are capable of converting any arbitrary text in ordinary spelling to a fairly intelligible synthetic speech. We may also program and train computers in a limited task of recognition of spoken words and sentences and to perform speaker verification. I shall not go into details here but it may suffice to state that computers are good readers but bad listeners. This is the reverse of general dyslexic behaviour.

Computer based text reading has found many applications in aids for the blind and in aids for speech handicapped persons. In these applications the input to the synthesis is a string of alphanumerical signs, i.e., a digital representation of the text which may be retrieved from a computer or from an optical character recognition system scanning a text for the user. Text-to-speech conversion has found use in prothesises for non-speaking persons, tailored to suit the users' specific motor capabilities (Fant, 1984; Galyas, Hunnicutt and Fant, 1992).

Here in Sweden we have been engaged in successful applications of speech synthesis in teaching aids for children with reading and spelling difficulties (Dahl

and Galyas, 1987). The technical development of such training aids and speech prosthesis now includes means of lexical prediction (Hunnicutt, 1996) by letting the computer suggest a possible continuation given the first entries of a word.

However, text-to-speech synthesis still lacks elements of naturalness and intelligibility. Also, we need a more varied individual choice of voice type i.e., child versus adult and female versus male voice. Speech synthesisers are often said to have a foreign accent. This is more or less pronounced and it can be tiring to listen to synthesisers, especially if they have not been programmed to stop now and then and to take a pause for breathing and afterthought.

4. Invariance and Variability

The relative success of speech synthesis has created an illusion that we have a profound insight in the speech code. This illusion becomes especially apparent when we try to operate in the reverse direction, that is, given a record of the speech wave, attempting to decipher what was said.

Although performance is gradually improving, automatic speech recognition is still in a primitive stage of development, generally limited to single-word utterances and speaker-adapted functions which require an initial calibration and training to fit the specific voice.

Why is this so? Why could we not develop rules for identification of phonemes irrespective of talker and context to convert speech to writing? Or even better, why not attempt a decoding of speech in terms of the phonetic inventory of natural sound classes, e.g., the distinctive features of Jakobson, Fant, and Halle (1952)? A limited number of about ten features would suffice to describe most of the phonemic contrasts in the languages of the world. Could one not simply implement their acoustic attributes in a computer program as a simple recipe for speech recognition? This naive thought, confusing abstractions and realities, has been followed up and it has failed for obvious reasons.

Distinctive features are in the first place minimal categories of a phonemic inventory. Their acoustic correlates, on the other hand, are dependent on the particular context. An invariant conceptual denominator can be formulated to fit an overall phonetic frame, but its descriptive power tends to be diluted and is found insufficient for recognition work. The contextual variability imposed by the speech code sets the limit. Still, distinctive feature theory remains a most important tool for gaining an understanding of the basic structure of speech and will find a more concrete base as we learn more about the speech code.

5. In Quest of the Speech Code

Let us go further into the nature of the speech code. There are often great differences between words pronounced in citation form under laboratory conditions and in real everyday speech. This is a matter of both tempo, relative emphasis, distinctiveness and stressed/unstressed contrasts. However, even a normal, clear reading of a text presents problems. Contextual variations are considerable. Temporal spread, fusion,

and overlay of simultaneous articulatory features have to be taken into account (Fant, 1962; Fant, Kruckenberg and Nord, 1991).

Phonological segments at the message level are abstract discrete units, while articulation is continuous. In the speech wave, as viewed from a time-frequency-intensity spectrogram, we find a mixture of continuous pattern variations and discrete breaks which are related to points in time of vocal tract opening or closing, i.e. at the release of an obstruction, or at the instant of a complete closure being reached. Onsets and offsets of the voice source or of a noise source may also provide distinct boundaries. Even though a single phoneme occasionally may have its main physical counterpart in a single acoustic segment, this is usually not the case, Fant (1962). In the speech wave, we may find more segments or less segments than in the phonological string and segmentation criteria may loose their precision or become ambiguous due to the continuity of the underlying articulatory gestures.

Thus, once a segmentation has been attempted, we find that one phoneme generally influences several successive acoustic segments. Conversely, due to coarticulation, a single physical segment is generally influenced by several successive phonemes. A typical instance of this temporal spread of information-bearing elements is that patterns of formant transitions into or out of a physical segment become important cues to the identity of the phoneme conventionally ascribed to the segment. An example is the positive transition in all formants at the release of a labial closure and the systematically different patterns of dental/alveolar or velar/palatal release. A stylised version of the speech code along these lines was developed at Haskins Laboratories (Liberman, Ingemann, Lisker, Delattre, and Cooper, 1959). It was based on experiments with simplified synthetic speech stimuli. An attempt to structure patterns of real speech was performed by Fant (1962, 1968).

However, in spite of an accumulating acoustic-phonetic knowledge, we still lack an insight in the full code sufficient for speech recognition purposes. Even if we could develop a perfect talent of reading spectrograms of unknown texts, we would still have difficulties in transforming our visual recognition strategies to schemes suitable for automatic computer analysis. Thus, in the lack of well-defined visual strategies and invariance criteria, most people involved in speech recognition research have now turned to non-phonetic approaches, employing statistical methods of pattern matching of units, usually larger than single phonemes. HMM and neural analog networks have taken over.

This is a real challenge. Shall we give up our attempts to force the speech code, or shall we continue to rely on computers to acquire a recognition competence that is as inaccessible to objective analysis and understanding as our competence as listeners? At least, we could make better use of insights in acoustic-phonetic structure.

But how shall we gain access to the complete code? The issue of invariance versus variability was the subject of a whole conference, Perkell and Klatt (1986). The view I expressed (Fant 1986), is that instead of worrying about the lack of invariance, we should pay more attention to means of studying and structuring variability. Spectrographic patterns of the same utterance produced by different speakers may vary considerably in spite of relative small differences in apparent qualities (Fant, Nord, and Kruckenberg, 1986). An example is shown in Fig. 2. The three subjects have produced the voiced stop /g/ in the Swedish word “legat” with

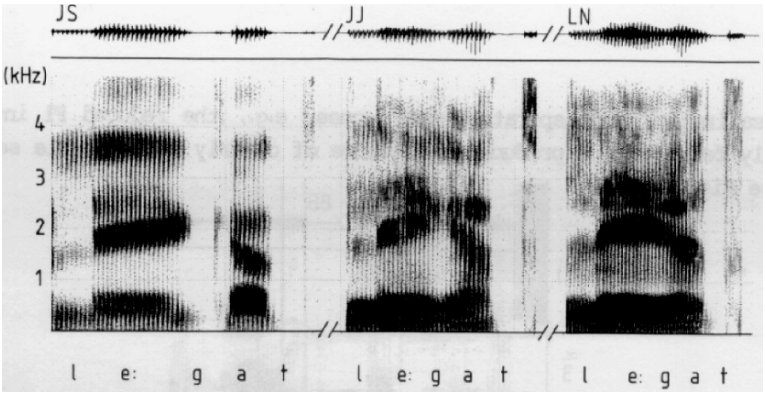


Figure 4.4.2. Normal and two degrees of target undershoot of a voiced stop.

varying degree of articulatory closure, which is complete for subject JS, incomplete for subject JJ and even more reduced for subject LN. The weak gesture towards closure for subject LN is sufficient to evoke the percept of a voiced stop but one may argue that the "top-down" expectancy in listening adds to the identification.

Another instance of incomplete articulatory closure is illustrated in Fig. 3. Here the subject AA at the left of the figure produces a welldefined boundary between the vowel /ö/ and the following nasal consonant /n/. The subject JS, on the other hand, does not produce a complete oral closure during the /n/ which therefore becomes realised as a nasalised vowel with no apparent boundary between the /ö/ and the /n/. This is a quite common phenomenon in a sequence of a vowel followed by a nasal and an obstruent.

These two examples illustrate indivual pronunciation patterns but they could also represent two levels of prominence, i.e. of stress along the hyper/hypo dimension as discussed by Lindblom (1990), see also (Fant, Kruckenberg and Liljencrants, 2000A).

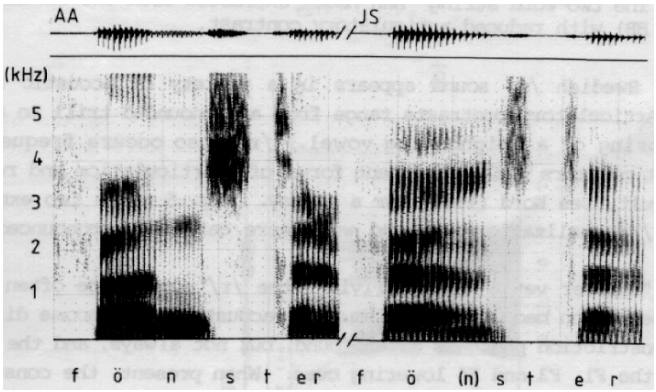


Figure 4.4.3. Complete and incomplete alveolar closure of a nasal consonant [n].

Some phonemes are produced with a complex combination of articulatory activities, while a minimal phonetic distinction often involves a single selective modification which may change several aspects of the spectrographic pattern. Such a case is illustrated in Fig. 4. Here the contrastive nonsense words [kakɑ:ka], [gagɑ:

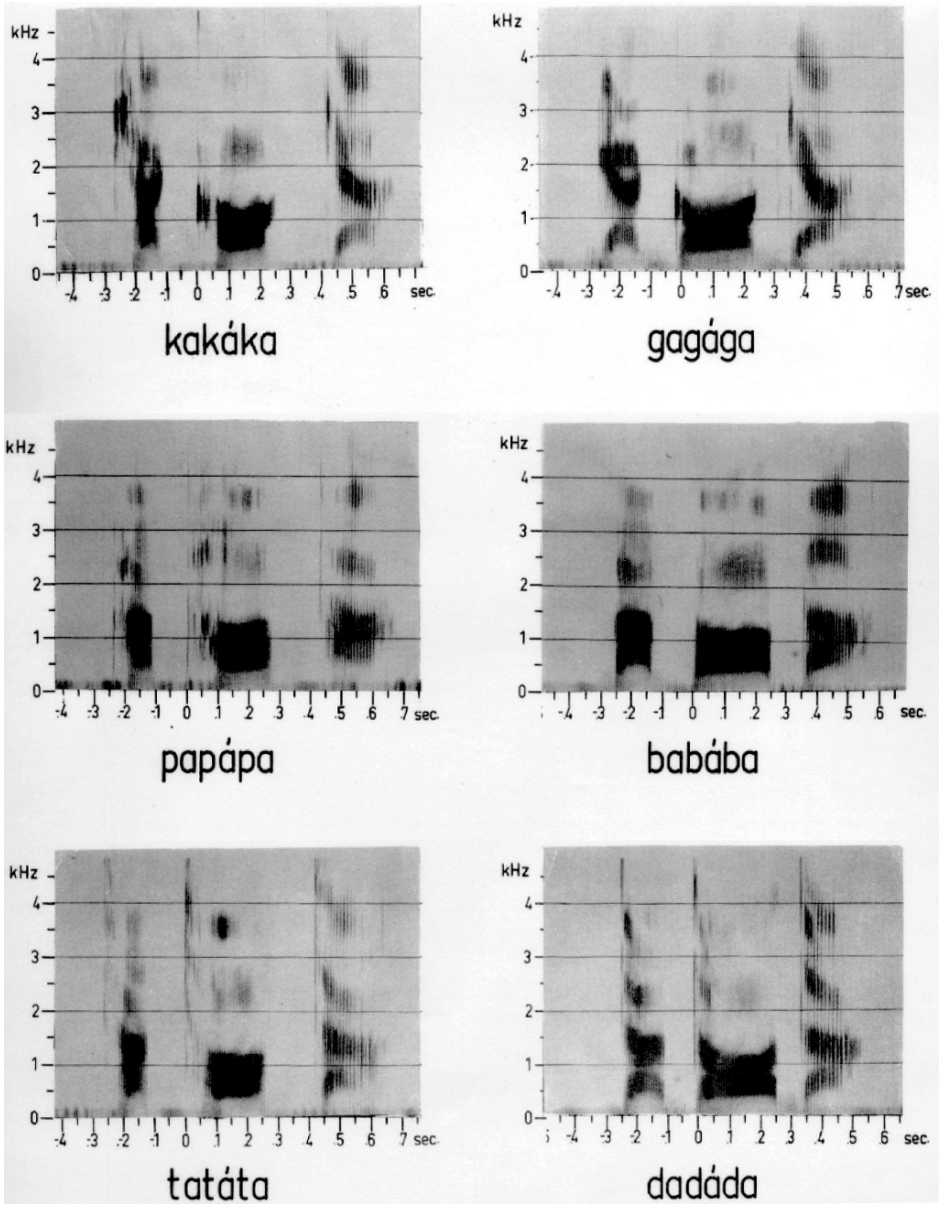


Figure 4.4.4. Glottal and supraglottal opening and closing in voiced and unvoiced (aspirated) stops determine duration patterns.

ga], [papα:pa], [babα:ba], [tatα:ta], [dadα:da] illustrate the pair wise relations between unvoiced (aspirated) and voiced stops. The major articulatory correlate of this distinction is the difference in the opening and closing gestures of the vocal cords. The corresponding acoustic modifications cover several dimensions such as F1 cut-back, aspiration noise and voice source dynamic variations. On the other hand there is perfect synchrony in the 6 words of the timing of the termination of the second vowel with respect to the onset of the burst of the preceding stop, i.e. the articulatory release-opening gesture is the same. The duration of the articulatory release-opening gesture is the same in all six words.

Prosody constitutes a major part of the speech code as imposed by language, dialect, social and individual patterns. The main categories are intonation, accentuation, prominence and grouping. It is by now well recognised that prosody constitutes the major difficulty for a second language learner and is the crucial determinant of the quality of text-to-speech systems. Earler work on Swedish prosody have been reported by Fant and Kruckenberg (1989, 1994). More recent publications are Fant, Kruckenberg and Liljencrants (2000A, B); Fant, Kruckenberg, Liljencrants and Hertegård (2000), Fant, Kruckenberg, Liljencrants and Botinis (2001) and Fant and Kruckenberg (2001). The prosody of poetry reading has been studied by Kruckenberg and Fant (1993).

It is found that the encoding of stress and relative emphasis affects not only duration, pitch and intensity but also most aspects of speech production, segmental realisations and the voice source (Fant 1993, 1995, 1997).

Advanced work on articulatory modelling will be a major tool for structuring the speech code. Effective vocal tract models and synthesisers have been developed (Lin and Fant, 1992; Engvall 2000) but we lack articulatory data and control strategies within a phonetic frame. The complexity of this task is obvious and has remained an inhibiting factor for the development.

Large data banks of recorded speech have been collected at many places but these are generally used for training speech recognition systems or for sampling of phonetically tagged speech fragments for synthesis, or for specific dialect studies. They are rarely being used for systematic studies of the speech code. Development work on speech synthesis is often carried out on a trial and error basis without a proper foundation in speech analysis. Exceptions are the now classical studies of Klatt (1987) and the corresponding work in Sweden by Carlson, Granström, and Hunnicutt (1990); Carlson and Granström (1986, 1997). However, much of the information is hidden in computer programs and is limited to specific contextual frames.

6. CONCLUSION

Speech research is facing a challenge. Without a broad integrated expansion of fundamental knowledge we can not realise far reaching aims such as communicating with computers as freely as with human beings. To achieve this insight is not a matter of a single intellectual undertaking of finding the key of a code. Brilliant ideas can not substitute hard work. In quest of the speech code we need, in the first place,

much more information from speech analysis. It requires multilevel investigations and modelling over a substantial period of time.

Mankind is making much progress in mapping the genetic code. We need some of the same patience and persistence in mapping the speech code, and of course a more appropriate funding. The reward will be not only a substantial gain in design and performance of speech systems but also a more solid theoretical basis of linguistics and phonetics and applications in a vast field of human communication.

NOTE

* The main body of the present article is an updated excerpt of (Fant, 1989), a contribution to a symposium on speech and brain functions sponsored by the Wennergren Foundation. It is dedicated to the memory of one of the editors, the late Ulf von Euler, a pioneer in studies of dyslexia.

REFERENCES

Carlson, R. and Granström, B. Eds. (1982). *The Representation of Speech in the Peripheral Auditory System.* Elsevier Biomedical Press, Amsterdam.

Carlson, R. and Granström, B. (1986). A Search for Durational Rules in a Real-Speech Data Base. *Phonetica*, 43, 140–154.

Carlson, R., Granström, B., and Hunnicutt, S. (1990). Multilingual text-to-speech synthesis with applications. In B. Ainsworth (ed), *Advances in Speech Hearing and language Processing*, Vol. 1 JAI Press Ltd, London.

Carlson, R., and Granström, B. (1997). Speech synthesis. In: Hardcastle W.J. & Laver J. *The Handbook of Phonetic Science.* Oxford: Blackwell Publ. Ltd, 768–788

Dahl, I. and Galyas, K. (1987). Experiences with the use of computer programs with speech output in Teaching Reading and Writing. In *European Conference on Speech Technology, Edinburgh*, II. (eds. J. Laver and M.A. Tack). CEP Consultants, Edinburgh.

Engvall, O (2000). Replicating three-dimensional tongue shapes synthetically. *TMH-QPSR* 2–3/2000.

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton, The Hague.

Fant, G. (1962). Descriptive Analysis of the Acoustic Aspects of Speech. *Logos*, 5, 3–17.

Fant, G. (1967). Auditory Patterns of Speech. In *Symposium on Models for the Perception of Speech and Visual Form, 1964.* (ed. W. Wathen-Dunn). The MIT-Press, Cambridge, MA.

Fant, G. (1968). Analysis and Synthesis of Speech Processes. In *Manual of Phonetics*. (ed. B. Malmberg). North-Holland, Amsterdam, 173–287.

Fant, G. (1978). Vowel perception and specification. *Rivista Italiana di Acustica* II, 69–87.

Fant, (1980). The Relation Between Area Functions and the Acoustical Signal. Phonetica 37, 55–86.

Fant, G. (1984). Human speech and communication aids. In, *Proc. of the II Intl. Conf. on Applications of Physics to Medicine and Biology.* Singapore, World Scientific Publ. Co., 3–19.

Fant, G. (1986). Features—fiction and facts. In J. Perkell and D. Klatt, (eds,) *Invariance and Variability of Speech Processes*. Lawrence Erlbaum Ass. Publ. 482–491.

Fant, G. (1989). The speech code. In C. von Euler, I. Lundberg and G. Lennerstrand (eds.) *Brain and Reading*. MacMillan, London, 1982, 171–182.

Fant, G. (1993). Some problems in voice source analysis. *Speech Communication* 13: 7–22 (1993).

Fant, G. (1995) The LF-model revisited. Transformations and frequency domain analysis. *STL-QPSR 2–3/1995:*119–155 (1995).

Fant, G. (1997). The voice source in connected speech. *Speech Communication* 22 :125–139.

Fant, G. and Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR* 2/1989, 1–83.

Fant, G. and Kruckenberg, A. (1994) Notes on Stress and Word Accent in Swedish, *STL-QPSR* 2–3/1944, 125–144. Also published in *Proc. Int. Symp. on Prosody, 18 Sept 1994, Yokohama,* 19–36

Fant, G. and Kruckenberg, A. (2001). F0 analysis and prediction in Swedish prose reading. In Nina Grønnum and Jørgen Rischel (eds.) *To honour Eli Fischer-Jørgensen.* Travaux du Circle Linguistique de Copenhague. Reitzel Copenhagen, 124–147.

Fant, G. Kruckenberg A and Liljencrants J. (2000A) Acoustic-phonetic analysis of prosody in Swedish. In (ed.) A. Botinis, *Intonation. Analysis, Modelling and Technology,* Kluwer, Academic Publishers, 55–86.

Fant, G. Kruckenberg A. and Liljencrants J. (2000B). The Source-Filter Frame of Prominence. *Phonetica* **57**, 113–127

Fant, G., Kruckenberg, A., Nord, L. (1991). Prosodic and segmental speaker variations. *Speech Communication* 10: 521–531.

Fant, G. Kruckenberg, A., Liljencrants J. and Botinis, A. (2001). Prominence correlates. A study of Swedish. *Proc. Eurospeech*-2001. (Revised version).

Fant, G., Kruckenberg, A., Liljencrants, J. and Hertegård, S. (2000). Acoustic phonetic studies of prominence in Swedish. *TMH-QPSR* 2–3/2000, 1–52

Galyas, K., Fant, G. and Hunnicutt, S. (1992). *Voice output communication aids*. A study sponsored by the International Project on Communication Aids for the Speech Impaired, IPCAS. The Swedish Handicap Institute, Stockholm (86 pages).

Hunnicutt, S. (1986) "Lexical Prediction for a Text-to-Speech System," in Communication and Handicap: Aspects of Psychological Compensation and Technical Aids, E. Hjelmquist & L.-G. Nilsson, (eds.), Elsevier Science Publishers.

Jakobson, R., Fant, G. and Halle, M. (1952). *Preliminaries to Speech Analysis. The Distinctive Features and their Correlates.* Acoustics Laboratory, Massachusetts Inst. of Technology, Technical Report No. 13 (58 pages). Published by MIT press, seventh edition, 1967.

Klatt, D. (1987). Review of text to speech conversion for English. *J. Acoust. Soc. Am*., 82, 737–793.

Kruckenberg, A., Fant, G. (1993). Iambic versus trochaic patterns in poetry reading. *Nordic Prosody* VI, Almqvist & Wiksell International, 123–135, (1993).

Liberman, A.M., Ingemann, F., Lisker, L., Delattre, P., and Cooper, F.S. (1959). Minimal Rules for Synthesizing Speech. *J. Acoust. Soc. Am*., 31, 1490–1499.

Liberman, A.M and Mattingly, I.G. (1985) The Motor Theory of Speech Perception. *Cognition*, 21, 1–36.

Lin, Q. and Fant, G. (1992). An articulatory speech synthesizer based on a frequency domain simulation of the vocal tract, *IEEE-ICASSP*, paper 173, San Francisco, March 23–26, 1992.

Lindblom, B. (1989). Phonetic invariance and the adaptive nature of speech. In B.A.G. Elsendoorn and H. Bouma (eds), *Working models of human perception* (pp. 139–73). London: Academic Press.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle and A. Marchal (eds.) *Speech Production and Modelling* (Kluwer Academic Publishers, Netherlands: 403–439 (1990).

Perkell, J., and Klatt, D. (1986) (eds,) *Invariance and Variability of Speech Processes.* Lawrence Erlbaum Ass. Publ.

CHAPTER 5

SPEECH PERCEPTION

Several studies of vowel perception were performed at KTH in the 1970's and earlier. These are reviewed in the first article (Fant, 1978), which also contains basic auditory theory and applications to two-formant approximations, with outlooks on neuro-physiological modeling. The basic issue is that the percept of vowel colour can be related to spectral shape rather than to individual formants, and that two or more formants can combine in shaping a relevant part of the spectrum. A two-formant approximation is based on the true F_1 of the vowel and an F_2' (F_2 prime) substituting F_2 and higher formants. F_2' comes close to F_2 in back vowels and close to F_3 in the vowel [i]. A simple experiment on the association of pure sinewaves and vowels reveals some of these relations.

The second article in chapter 5 deals with an application of speech analysis to audiology. My data on the distribution of vowel and consonant formants, in frequency and intensity level from my early work at Ericsson (Fant,1959), have been adopted for assessing how much of the spectral information is potentially available, assuming a specific pure tone audiogram and speaking distance. This has been possible by the absolute calibration of sound pressure levels in my speech analysis. An estimate of the speech reception ability is performed by applying weights to the separate audiogram frequencies, and integrating over the audible part of the speech spectrum. The shape of the formant distribution has motivated the term "the speech banana", frequently used in audiology.

SELECTED ARTICLES

[5.1] Fant, G. (1978). Vowel perception and specification. *Rivista Italiana di Acustica* II, 69–87.

[5.2] Fant, G. (1995). Speech related to pure tone audiograms. In G. Plant and K.E. Spens (eds), *Profound deafness and speech communication*. London: Whurr Publ. Ltd, 299–305.

ADDITIONAL READING

Lidén, G. and Fant, G. (1954). Swedish word material for speech audiometry and articulation tests. *Acta Oto-Lar.* Suppl. 116, 189–210.

Wedenberg, E. and Fant, G. (1949). Auditory training of deaf children. *Acta Oto-Lar*. XXXVII, Fasc. 5, 462–469.

Fant, G. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. *Ericsson Technics* **1**-1959, 1–106.

Chistovich, L., Fant, G., de Serpa-Leitao, A. and Tjernlund, P. (1966). Mimicking and perception of synthetic vowels. *STL-QPSR* 2/1966, 1–18, and part II in *STL-QPSR* 3/1966, 1–3.

Fant, G., Carlson, R. and Granström, B. (1975). The [e]-[ø] ambiguity. In G. Fant, (ed.) *Speech Communication,* Proc. Speech Communication Seminar, Stockholm 1974, Stockholm, Almqvist and Wiksell. Vol **3**.

Carlson, R., Fant, G. and Granström, B. (1975). Two-formant models, pitch, and vowel perception. *Auditory Analysis and Perception of Speech* (G. Fant and M.A.A. Tatham, eds.). London, Academic Press Inc., 55–82.

Bladon, A. and Fant, G. (1978). A two-formant model and the cardinal vowels. *STL-QPSR* 1/1978, 1–8.

Fant, G. (1983). Feature analysis of Swedish vowels—a revisit. *STL-QPSR* 2–3/1983, 1–19.

CHAPTER 5.1

VOWEL PERCEPTION AND SPECIFICATION

More than twenty years of research efforts in the field of vowel specification in a perceptual scale are summarized here.

Phsychoacustical concepts are adopted to explain how our auditive system utilizes the many acoustical characteristics to define vowel quality. Particular attention is devoted to the problems of fundamental frequency-timbre interaction and of male-female differences.

1. INTRODUCTION

In studies of connected speech as well as of isolated vowels we are plagued by a variability of speech data and with a variability of percepts and of notational systems. The speech code varies with language, dialect, age and sex and specific physiological constraints of the speaker. Even within the frame of a single speaker the contextual constraints, such as immediate phonetic context, stress, intonation, place within a sentence, speaking tempo, and voice effort condition large variations. Also, there is the pathology of speech and hearing to take into account in a detailed analysis. Our knowledge of some of these factors is still rather incomplete, especially their implication to perception theory. If we instead look for essential principles of the decoding mechanism much of this variability is overcome by reference to relational invariance. This is the basic principle of the theory of distinctive features.

When it comes to defining absolute vowel quality we get in troubles since a phonetician's judgement and labeling is much influenced by particular language frames. According to P. Ladefoged (1967) [9] the only secure basis for defining absolute vowel quality is to refer to judgements of a jury of phoneticians trained in the Daniel Jone's school of transcription. But this is a very exclusive reference and Daniel Jones's voice, still available on grammophone record, has a high-pitched rather extreme quality.

Now, why could we not use a synthesizer as a reference and a computer to replace the expert jury? This is indeed a challenge. Synthesis techniques are technically sufficiently well developed to satisfy high demands of quality, at least as far as male voices are concerned. One limitation in present know-how is that of insufficient flexibility in synthesis to cover various speaker categories such as male, female, adult, child but this is not a major obstacle. Further work in speech production and speech analysis can provide these insights but we still need a solid basis of knowledge about speech perception. Such knowledge need not be physiologically true as long as we capture functional essentials and can formulate the constraints we have to apply to speech-wave data. A functional model of auditory perception would be the requirement for a computerized assessment of vowel quality.

In general then, in order to cope efficiently with our specification problems, we need to consider the basic constraints on speech-wave data related to both production

mechanisms, choice of parameters and their mathematical interrelations, and the auditory and perceptual mechanisms.

A basic problem is that of dimensionalities. The finer observations we want to make, the larger number of parameters do we need which increases the complexity of the task. With a reduction of the number of parameters, we loose accuracy of specification but gain simplicity of description. This is one part of the experimental phonetician's dilemma. Even if we knew everything about dimensionality of vowel percepts we would still demand various levels of approximations for various descriptive purposes.

2. DIMENSIONALITY

A very detailed physical specification of a vowel would include the frequencies, bandwidths and amplitudes of three or four formants, the frequency of the voice fundamental and some additional information about the voice source spectrum. With our present insight in the theory of speech production and mathematical analysis of speech signals, we can state important interrelations, such that formant bandwidths and formant amplitudes are highly predictable given the complete pattern of formant frequencies up to the fourth and assuming normal voice production, speaker category, and absence of nasalization. On the basis of a statistical analysis of variance between the spectra of different vowel sounds R. Plomp (1975) [11] of the Inst. of Perception in Soesterberg, The Netherlands, has calculated the following order of significance: the frequency of the second formant F_2, the frequency of the first formant F_1, the frequency of the third formant F_3, the level of the third formant L_1.

The predictability of formant amplitudes from formant frequencies or rather the great dependency of overall spectral shape on formant frequencies is systematic and of considerable importance in specificational theory (Fant, 1960 [3]).

These interrelations are preserved in speech synthesis with a serial analog, for instance the Swedish OVE-type synthesizer. A first and primitive step in adjusting stimulus specifications to an auditory scale is to exchange frequency for mel scale and the logarithmic dB scale for a semilog scale. This is exemplified in fig. 1.

Observe the dimension of spectral energy represented by the area under the spectrum curves which increases with increasing first-formant frequency F_1. This is the phonetic dimension of closed and open vowels and also that of voiced consonants to vowels. A part of the syllabicity of speech is thus related to the frequency of the first formant.

One conclusion we may draw from this demonstration is also that synthesis can provide naturally sounding vowels. By changing the parameters of synthesis we may accordingly evaluate the perceptual importance of various speech-wave parameters. One observation to be made is that moderate changes of formant bandwidths do not significantly change the vowel quality except under the most sensitive conditions, that is when a vowel is located at the boundary between two response categories. An increase of first formant bandwidth is generally associated with a feature of superimposed nasalization.

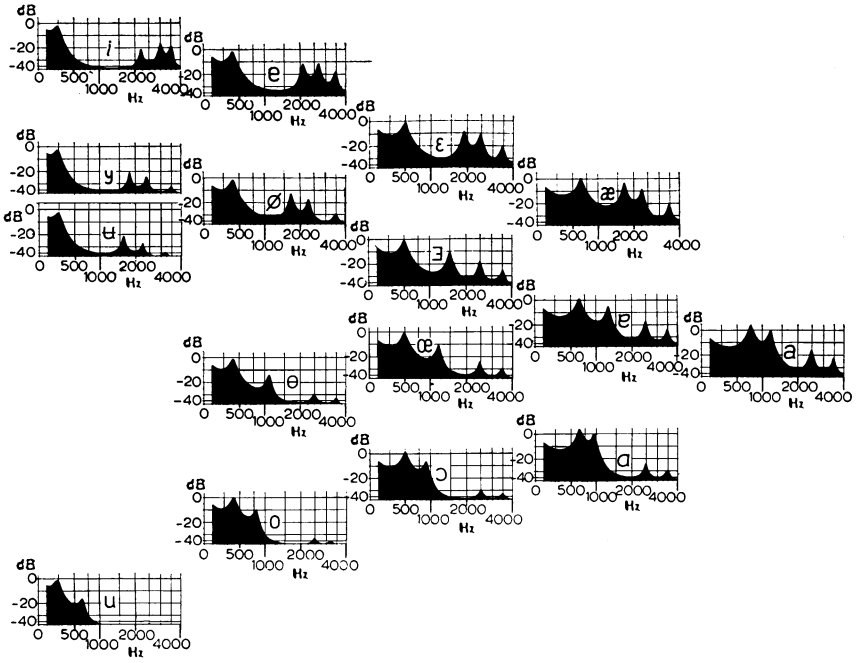


Figure 1. Spectra of synthetic vowels arranged in quantal steps of F_1 and F_2 (from Fant, Proc. 5th int. Congr. phon. Sci., Münster 1964).

We may go one step higher in ambition and ask ourselves: which are the most important spectral attributes for vowel identification? An extreme form of synthesis is to ask an audience to identify each of a number of sine waves of different frequencies with specific vowels. You listen to a tone and write down a vowel symbol.

This is a very easy experiment to administer. The outcome is a series of graphs, one for each vowel, showing the probability of its response as a function of the sine-wave frequency (see fig. 2). This figure pertains to Swedish vowels (Fant 1959 [2]). For the back vowels [u] [o] and [*a*] there is a marked similarity between these probability curves and the vowel spectra. We are thus in a position to claim that the most important part of the spectrum of back vowels is the F 1 F 2 region, and in the vowel [i] it is the F 3 F 4 region. Similar but more specific results would have been obtained with one-formant synthetic speech sounds.

3. F_2' Analysis

The next level of approximation is to use two formants, not simply F 1 and F 2, but F 1 and something I have labeled F_2' (F_2 prime), which is the preferred setting of a second formant of a synthetic two-formant sound which a subject can vary in a matching experiment to approximate the quality of the complete vowel.

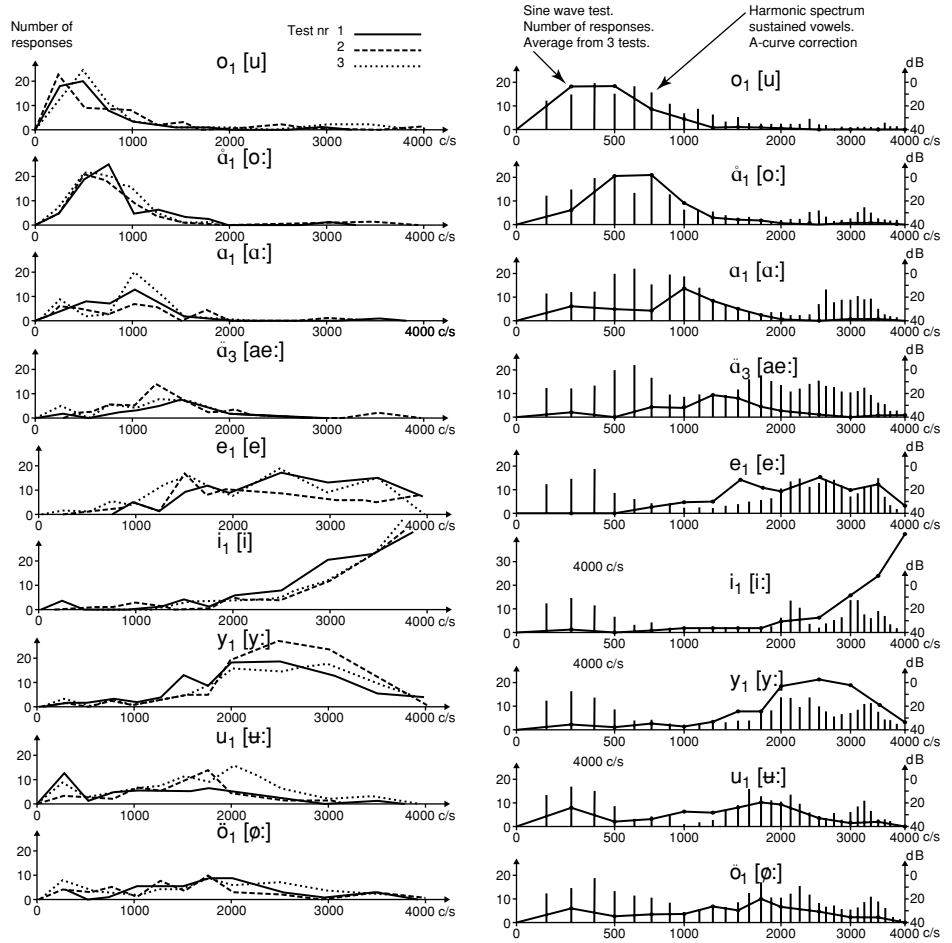


Figure 2. Results from single sine-wave tone-vowel association tests. The distributional curves representing probability of vowel responses as a function of frequency are compared with vowel spectra (from Fant, 1959).

As seen in fig. 3 subjects tend to place F_2' close to F_2 in back vowels and some centralized vowels. In front vowels F_2' occupies a position between F2 and F3 except in the vowel [i], where F_2 prime is situated in the F3 F4 region.

This kind of two-formant synthesis is nothing new. It was a routine in the early work at Haskins Laboratories. Now in Stockholm my colleagues Rolf Carlson and Björn Granström have been looking more closely into the auditory significance of this kind of representation (Carlson et al., 1975 [1], and in Leningrad a corresponding research has been carried out by Ludmilla Chistovich and her associates (Karnickaya et al., 1975 [7]).

I shall return to the Russian work later and proceed to discuss some of the work within our group in Stockholm. Let us go back to the linear scale broad-band spectra of the vowels and search for some correspondence to the F_2'. We see that in the [i]

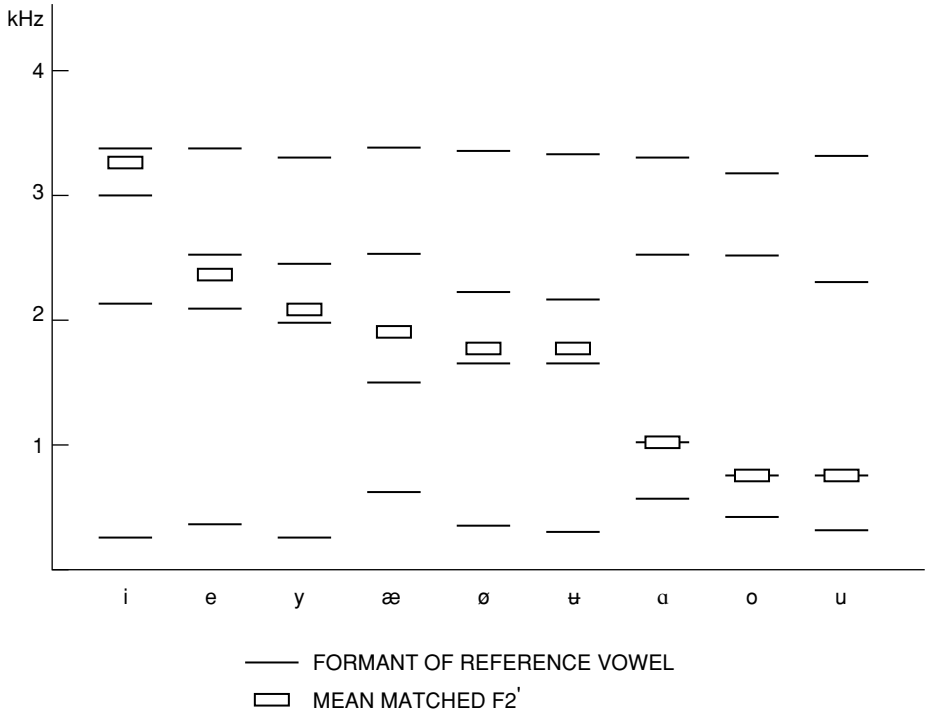


Figure 3. F_2' indicated by rectangles derived from tests with two-formant synthetic vowels matching four-formant synthetic vowels of the same F_1 and F_0 contour (from Carlson et al., STL–QP SR 2–3/1970).

spectrum F_4 and F_3 dominate but that the spectral weight is shifted to F2 and F3 in [e] and in the rounded vowel [y]. Actually the main difference between the vowel [y] and the vowel [i] is a 500-Hz lower F_3 in [y]. In addition F_2 and F_4 of [y] are 100 Hz lower than in [i]. The difference in F_2', however, is 1200 Hz, which is about twice the sums of the difference in formant frequencies.

The differential effect of moving F_3 alone retaining constant frequencies of other formants in the [i]-[y]-domain is shown in fig. 4. In the boundary region between [y] and [i] a small shift in F_3 causes the matched F_2' to change with a larger amount. This is a highly non-linear relation which can qualitatively be explained by the interrelation between formant-frequency patterns and formant levels and the assumption that the auditory system picks out a peak of spectral prominence in a group of formants such as F2 F3 and F4 of front vowels. When F_3 is close to F_2 the matched F_2' falls in a region between F_2 and F_3 and often closer to F_2. When F_2 is closer to F_1 than to F_3 the level of F_3 will decrease and F_2' matches F_2 rather closely. General considerations of this sort have enabled us to construct a formula for predicting F_2' from formant frequencies F_1 F_2 F_3 and an average F_4. The agreement between predicted and matched F_2 primes is good.

Another approach supporting the F_2' concept started out from a filtering with the v. Bèkèsy-Flanagan cochlea model (Flanagan, 1965 [5]). This provides a rather

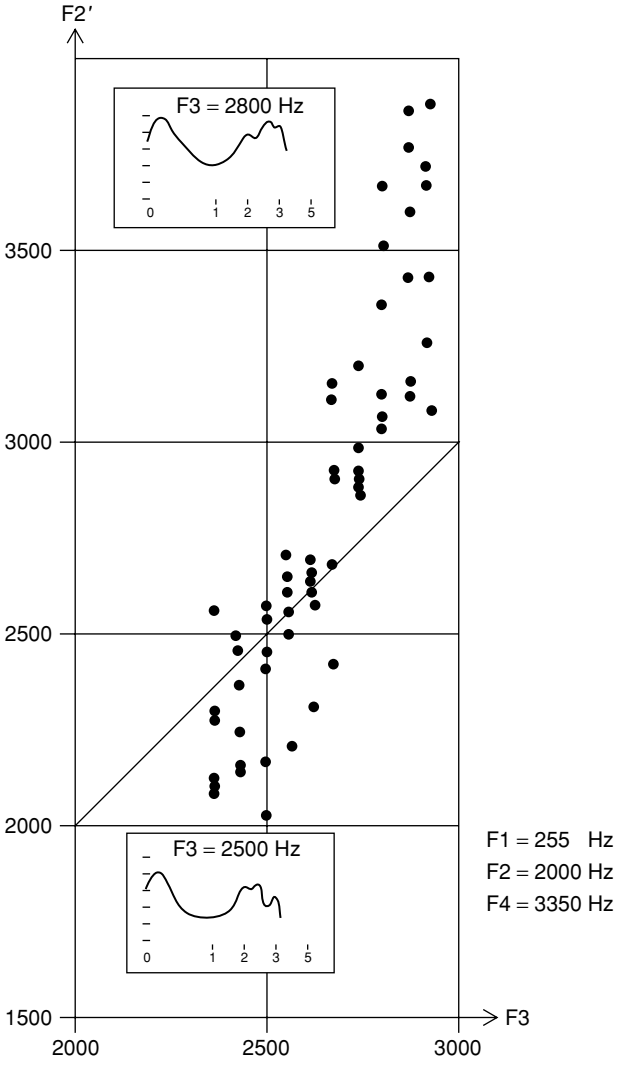


Figure 4. A moderate shift in F_3 along the [i] [y] continuum provides a large shift in F_2' (from Carlson et al., STL-OPSR 2–3/1970).

unselective analysis. Spectral shapes are heavily smoothed out. However, it occurred to Carlson and Granström that one should study the time-domain statistics of the ouput of each cochlea tap of the model. In doing so they calculated the zero-crossing frequency at each output tap and constructed a graph showing how many taps displayed the same zero-crossing frequency within a small range of frequencies. Now, this distribution becomes sharply selective with two prominent peaks appearing, one for F_1 and one higher peak which proved to coincide closely with F_2' prime from the matching experiments. (See fig. 5 pertaining to the reference synthetic vowel [i]). The cochlea-analog analysis has been tested on human vowels [o] and [e], see

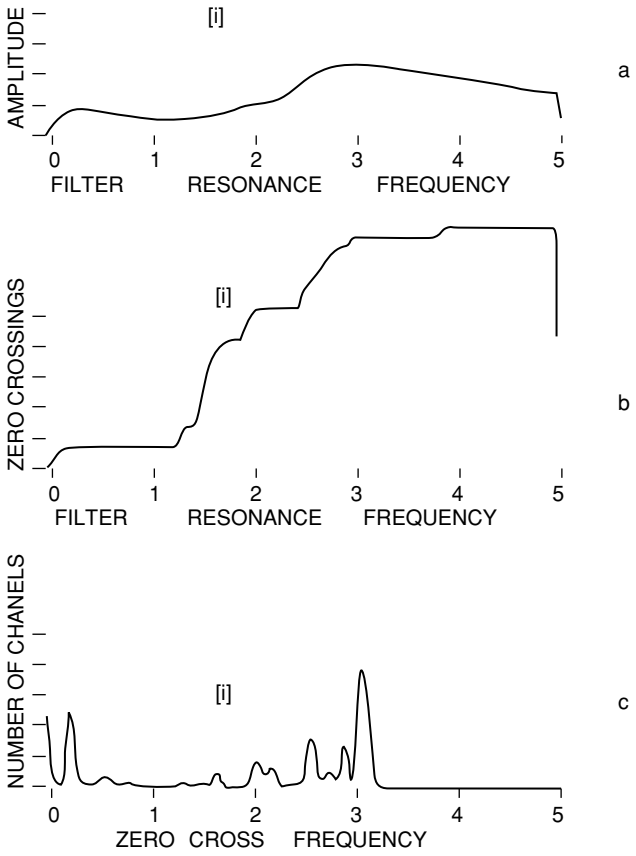


Figure 5. Cochlear model processing of a synthetic vowel [i]. a) amplitude envelope on the basilar membrane; b) zero-crossing frequency distribution and c) density along the basilar membrane.

fig. 6: the two peaks appear as expected. However, the results are not always this fine and more experience is needed with this model before we can recommend it.

We have now anyhow selected evidence from three independent studies which support the validity of the two-formant F_1 and F_2' approximation and provide a numerical agreement (Carlson et al. (1975 [1]).

4. Loudness Density Analysis

In spite of the success of the cochlea model we can claim functional relevance only. Another approach is to look for a psychoacoustic model. A third-octave filter bank provides a rather good approximation to a set of critical band filters, as specified by Zwicker and Feldtkeller (1967) [12]. Such a filter bank is the 8051A Hewlett & Packard analyzer for measurements of spectral loudness density. Our reference four-formant synthetic vowels were analyzed by such a filter bank. Figure 7 shows the loudness density patterns of the vowels [u] [o] [ɑ] [æ] [e] [i] [y] [ʉ] [ø]. Except for

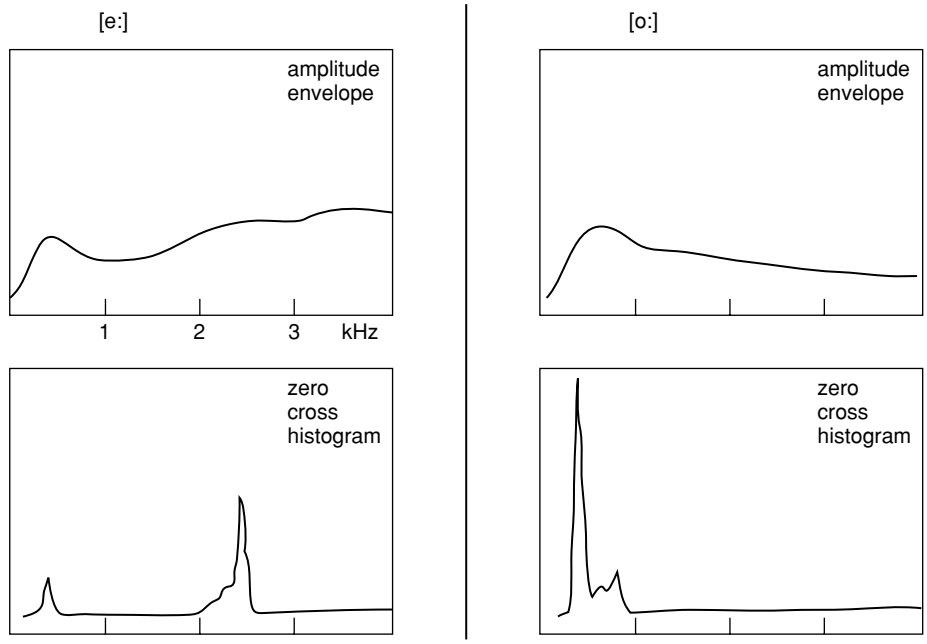


Figure 6. Same as Fig. 5 with natural vowels [e] and [o]. Unpublished data by Carlson and Granström.

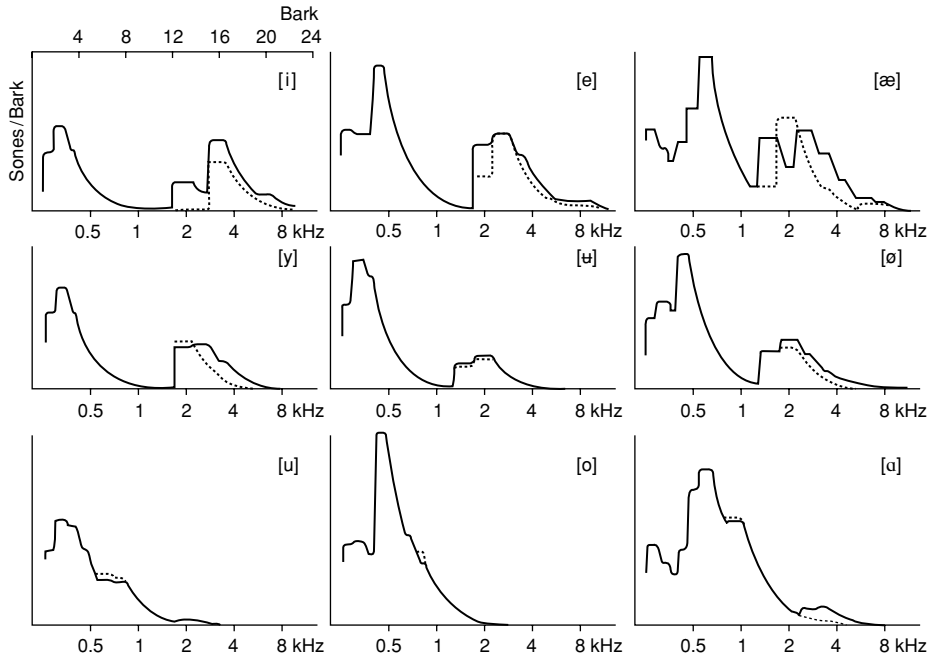


Figure 7. Loudness density spectrograms of Swedish vowels obtained by a HP third octave analyzer.

[æ], which has separate peaks for F 2 and F 3, we have two dominating peaks for front vowels and essentially one peak only for back vowels, which conforms with studies at Bell Laboratories and Haskins Laboratories about 20 years ago and with the single sine-wave association test that I discussed earlier. More exact simulations of loudness-density processing should be undertaken.

Figure 8 shows the masking contours of narrow-band maskers of various frequencies (Zwicker and Feldtkeller, 1967 [12]). The masking from low-frequency sounds extends with appreciable amounts to higher frequencies. This effect is grossly taken care of in the Hewlett and Packard analyzer. The phonetic implication of such masking is that as we gradually decrease the intensity level of F_2' nothing radical happens with the response until F_2' falls below the masked threshold set by F_1. At this instance the perceptual response shifts rather suddenly to that of a one-formant sound set by F_1 alone. Thus a vowel [e] or [ø] will shift to [o] when F_2' is reduced sufficienlty. Relative formant amplitudes within a cluster of formants, such as F 2 F 3 F 4 F 5 are important for giving weight to a particular F_2', whereas the overall level of the formant group can be changed within wide limits before the response changes. This conclusion also appears from the work of Karnickaya et al. (1975) [7]. They claim that the auditory system picks out and measures the location of the two major peaks even if there are relative prominent other peaks present. Besides the loudness-density processing she postulates a specific lateral inhibition mechanism for emphasizing a lower formant in a group of two adjacent formants.

5. AUDITORY ORIENTED VOWEL DIAGRAMS

For phonetic descriptive work it seems rational to specify vowels in terms of the sum and difference of F_2' and F_1 on a mel-scale. These two parameters which we may label center of gravity and spectral spread constitute a rotation of the M_1 M_2' plane by 45 degrees. A presentation of Swedish vowels in terms of these parameters $(M_2' + M_1)$ and $(M_2' - M_1)$ is given in fig. 9.

We see how all back vowels are efficiently separated from the rest as a group with minimal spectral spread, $M_2' - M_1$ and how we have additional levels of $(M_2' - M_1)$ for mid and front vowels. The unrounded front vowels have approximately the same center of gravity whilst the rounded front vowels [y] [ʉ] and [ø] occupy a lower level of $(M_2' + M_1)$.

There is some tendency of equal spacing between adjacent vowel symbols. This principle has been adopted for generating a set of two-formant synthetic vowels constructed from a rule of approximately 0.8 critical band-quantal steps in the center of gravity and about 2.5 critical bandwidths in the domain of spectral spread (see fig. 10 a tape-recording of these vowels can be supplied by request).

Phonetic (IPA) symbols have been arbitrarily assigned to the samples. We can hear that the quality is distinct and fairly natural in most parts of the system except perhaps in the region of the vowel [i] which sounds too "thin". These kinds of synthetic stimuli might find some applications as standards for percentual experiments. We may also conclude that there is a need for a third parameter for improving the quality. This could be a parameter of spectral spread in the F 2 F 3 F 4 region.

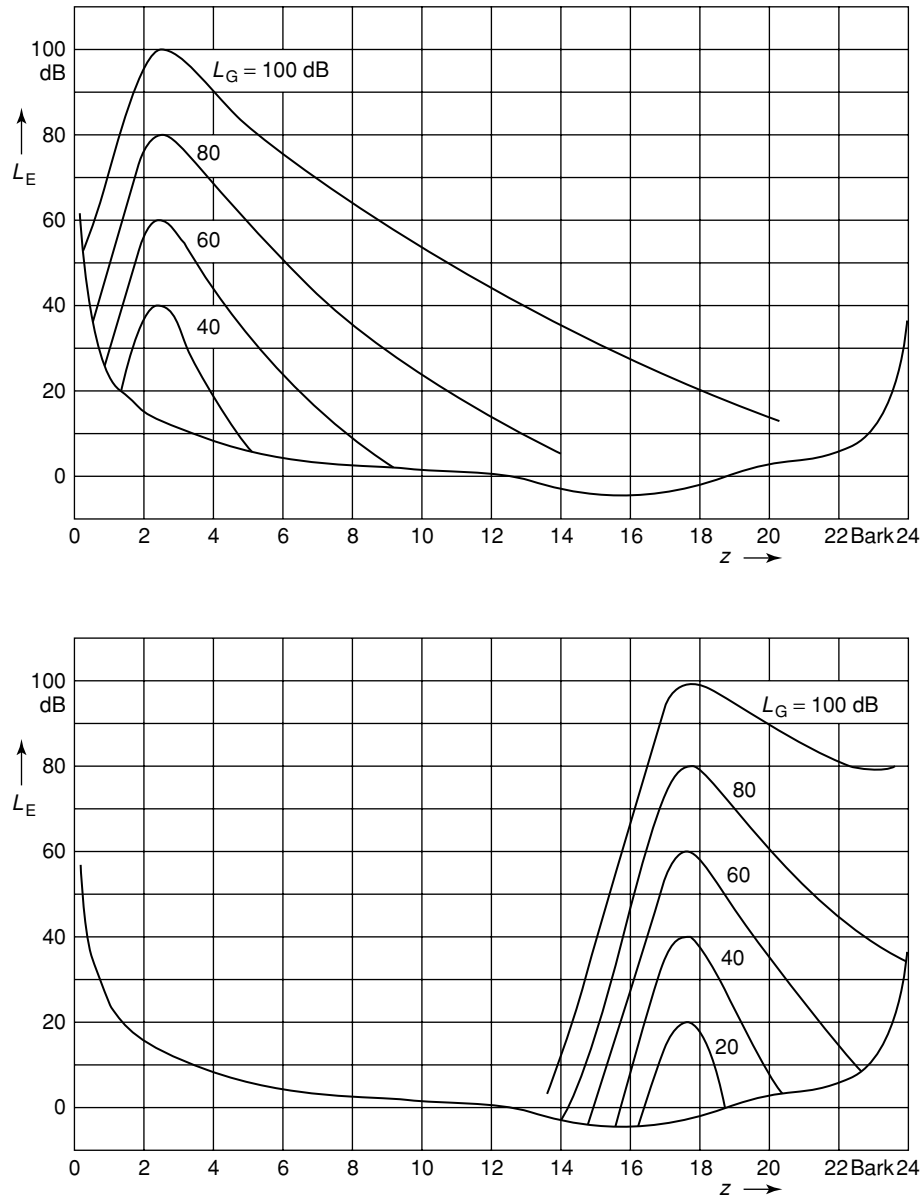


Figure 8. Spatial excitation patterns along the basilar membrane for a low frequency and a high frequency narrow band of noise (from Zwicker and Feldtkeller, 1967)

The conclusion we can draw from these studies is that F_2' (F_2 prime) has a greater specificational power than F_2 alone and that F_2' appears to be the better parameter to be used together with F_1 in a two-dimensional representation. It remains to see how great the need will be for a third dimension. A transformation of F_1 and F_2' to corresponding measures on a mel-scale is motivated but I would go one step

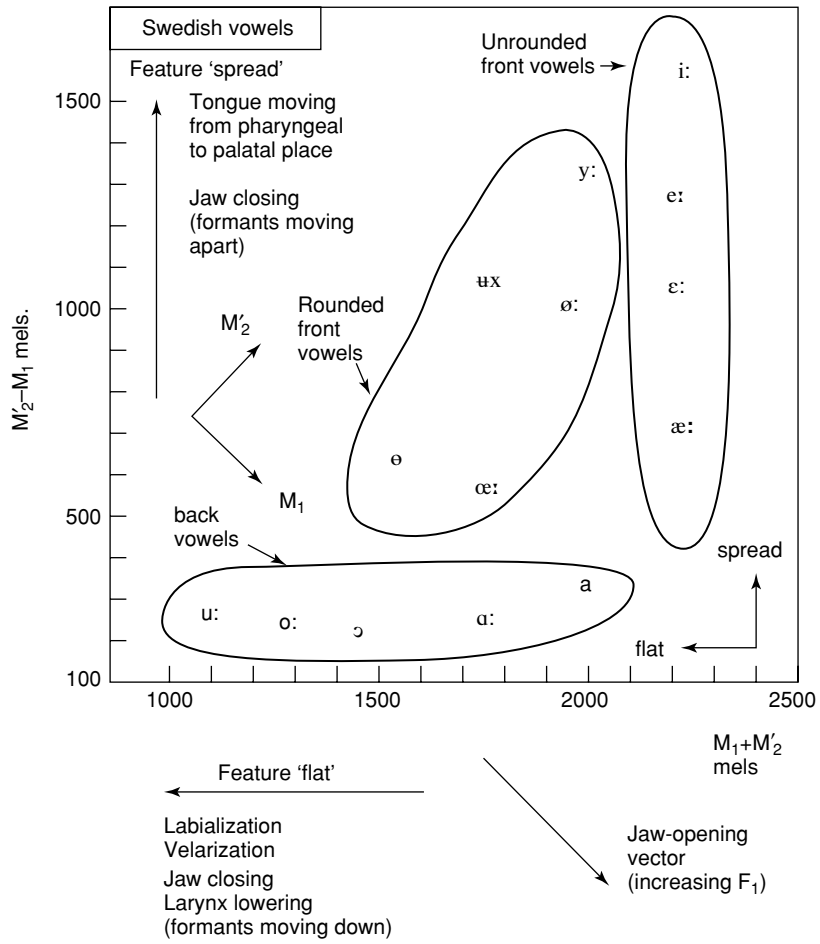


Figure 9. Swedish vowels in a $M_2' - M_1$ "spread" venus $M_2' + M_1$ "Flat" mel-scale; from Fant, STL-QPSR 2–3/1969.

further and recommend the Zwicker Bark-scale for critical bands of hearing which is approximately proportional to the mel-scale but more significant to problems of frequency-to-auditory space transformations.

6. F_0 Timbre Interaction

We are now approaching the central theme of the constituents of vowel timbre. Besides the auditory space-frequency distribution of sound there is the frequency of the voice fundamental to consider. It is known from experiments with synthetic vowels (Miller, 1953 [10], Fujisaki and Kawashima, 1968 [6] and Carlson et al., 1975 [1]), that given a constant spectral envelope a change of F_0 from about 120 Hz to 240 Hz shifts the perceived quality to that of a more rounded vowel, that is, a vowel of lower F_1 and F_2'. In the domain of back vowels the octave increase in

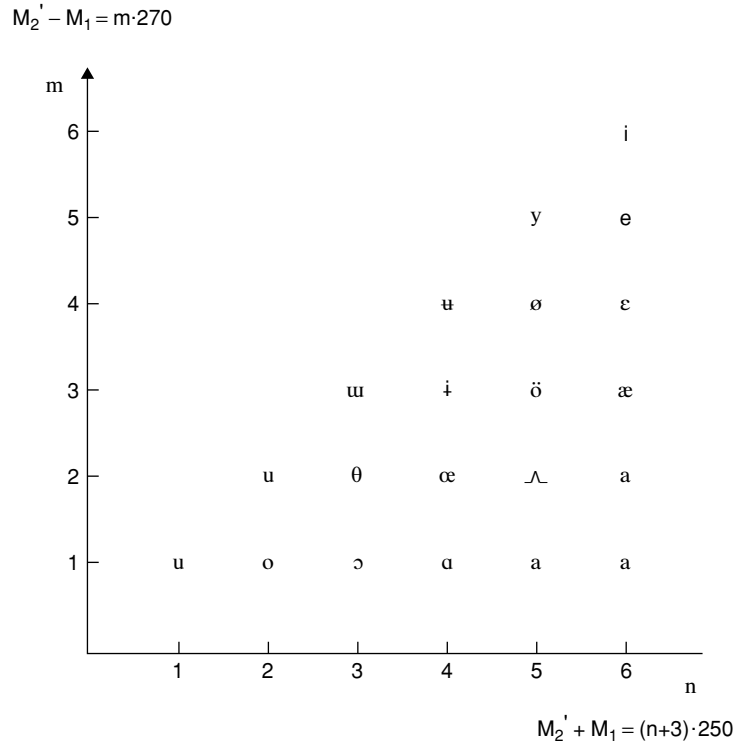


Figure 10. Two-formant synthetic vowels arranged in equal increments of $M_2' - M_1$ and $M_2' + M_1$. The step sizes are of the order of a critical bandwidth.

F_0 may be compensated by a 60 Hz rise in F_1 and F_2'. Without this compensation a vowel [ɔ] attains a quality closer to [o] and [ɛ] is shifted in the direction of [ø].

What are the perceptual mechanisms behind this effect? It could well be a learned response by associating higher F_0 with female voices. Although it has been found that females and males tend to differ less in terms of F_2' than in terms of F_2, there still exists a difference also on a mel-scale or on a Bark-scale. The perceptual normalizer could then be the value of F_0, some overall features related to the length of the vocal tract or to source spectrum. This is then a matter of inherent normalization and not the kind of sequential conditioning demonstrated by Ladefoged and Broadbent (1957), in which the identification of a vowel differs dependent on the scale factor of formants in a carrier phrase.

We can detach a female vowel from its contexts and still perceive the intended phonetic value. In contrast to a theory of learned response we could imagine a lower-level mechanism by which the complex timbre is influenced by the voice fundamental in such a way as to cause a compensation in the perceived timbre. One most hypothetical suggestion I have made is that the perceived measure of center of gravity in the spectrum becomes a weighted average between the space-frequency

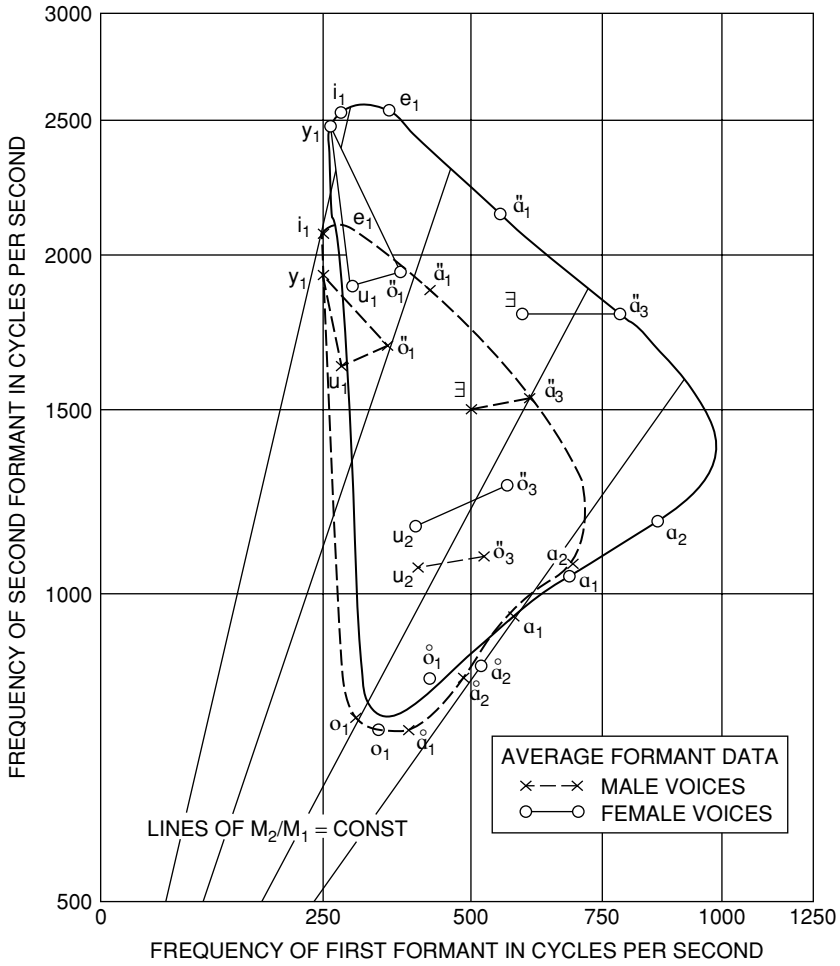


Figure 11. F_2 versus F_1 of Swedish female and male sustained vowels (from Fant, 1959).

representation of the formants and the voice pitch. The effect would increase with increasing F_0. The averaging of F_0 and formants would cause the total center of gravity to shift to slightly lower frequencies as observed. However, this effect, if present, does not resolve all the female-male or the child-male differences. I have not undertaken any investigation of whispered speech comparing different speaker categories but I assume that the invariance is inherently preserved independent of the fundamental.

7. MALE-FEMALE DIFFERENCE

Figure 11 shows the average formant contours in the F_1 F_2 plane for males and females in the Fant (1959) [2] Swedish study. I shall not go into details of the

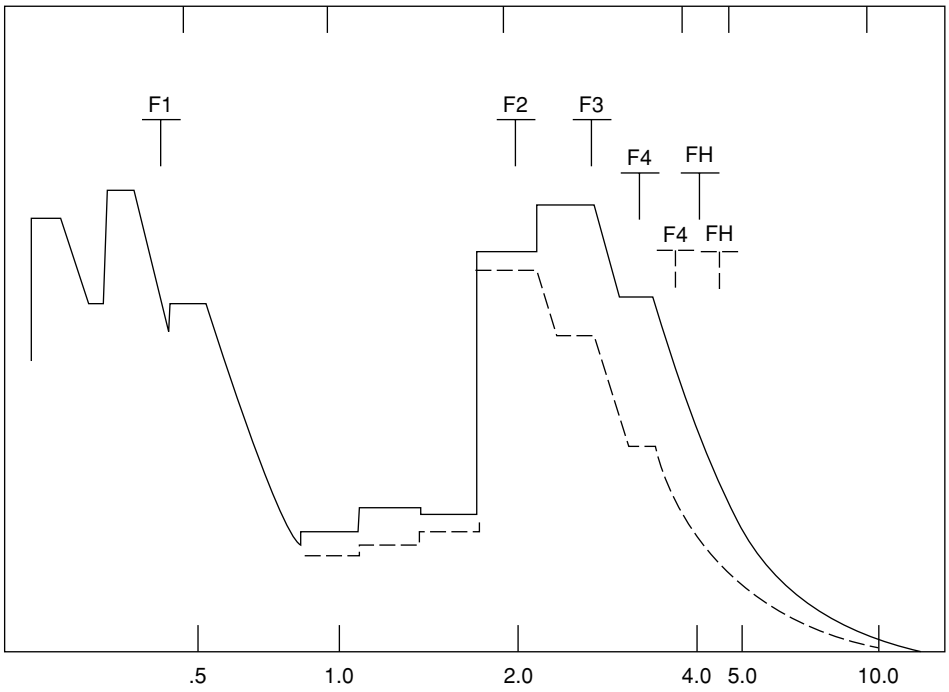


Figure 12. Shift in loudness density spectrum of a vowel when F_4 and higher poles are shifted from average male to average female location (from Fant et al., 1975).

relation but merely exemplify the ambiguity that an F1 F2 point in the area of a male's [e] vowel represents a vowel [ø] in the female distribution. One would expect this ambiguity to be resolved by specification of F3. This is not always the case. There exist [ø] vowels with exactly the same F_1 F_2 and F_3 as in a male's [e]. What additional differences do we have? What happens with F_2 prime? Assuming that the female really has a shorter vocal-tract length than the male, we would expect higher location of F4 and higher poles. But would this not raise the F_2 prime to a value higher than for a male? The answer is no and this is a paradox unless we understand the particular relation between formant location shifts and corresponding spectral shape and intensity changes.

Loudness-density curves for two front vowels differing in F4 and higher poles only are shown in fig.12. The shift to higher location of the upper formants removes the support these formants provide to the intensity of F3 and in effect the center of gravity of the F2 F3 region is shifted down in frequency. In addition there is a narrowing of the frequency span of F2 and higher formants. Thus, in spite of F1 F2 and F3 being the same, the F_2' measure differs in a direction so as to favor [ø] response. Other factors in the [e] [ø] distinction to consider are of course the difference in F_0 and also a difference in formant bandwidths, the rounded [ø] having an appreciably lower bandwidth than [e] (Fant et al.; 1975).

There are other instances where a seemingly apparent paradox effect is resolved by reference to the F_2' concept. Thus, one component in the female-male difference in the composition of the vowel [e] in Swedish is a larger than average shift in F_2 However, the differential effect of an increase in the frequency F_2 in this area is not necessarily an increase in F_2'. Under the condition that F_2' was located in the F 3 F 4 area before the shift, an elevation of F_2 tends to emphasize the F 2 F 3 area typical of [e] and weaken the F 3 F 4 area typical of [i]. This is again an example of the non-linear but nevertheless systematic relation between formant positions and F_2'.

There is a simple physiological correlate of F_2' as the tendency of the mouth or the front cavity to determine the most important formant of the spectrum. With the exception of sounds which do not have a major constriction in the vocal tract we may thus state that F_2' is associated with a front-cavity resonance. Females and males differ less in the length dimension of this cavity than in the posterior part of the vocal tract. This is one eplanation why the F_2' concept reduces observed female-male differences. In a sense we are here back to the old view of vowels having two formants only, the second one being a front-cavity resonance. In many consonants the front cavity is the main determinant of formants and here the situation is similar.

(Ricevolo il 31 maggio 1978)

REFERENCES

[1] Carlson R., Fant G., Granstrom B.—*Two-formant models, pitch, and vowel perception*,—in Auditory Analysis and Perception of Speech (ed. by G. Fant and M.A.A. Tatham), Academic Press, London (1975), p. 55.

[2] Fant G.—*Acoustic analysis and synthesis of speech with applications to Swedish*—Ericsson Technics, No. (1959); also publ. in Speech Sounds and Features, MIT Press (1973).

[3] Fant G.—*Acoustic Theory of Speech Production*—Mouton, s-Gravenhage (1960) (2nd edition 1970).

[4] Fant G., Carlson R., Granstrom B.—*The* [e]—[ø] *ambiguity*,—in Speech Communication, vol.3 (ed. by G. Fant), Almqvist & Wiksell, Stockholm (1975), p. 117.

[5] Flanagan J.L.—*Speech Analysis Synthesis and Perception*—Springer Verlag, Berlin (1965); 2nd extended edition 1972.

[6] Fujisaki H., Kawashima T.—*The roles of pitch and higher formants is the perception of vowels*—IEEE Transactions on Audio and Electroacoustics AU - 16, (1968), p. 73.

[7] Karnickaya E. G., Mushnikov V. N., Slepokurova N. A., Zhukov S. Ja—*Auditory processing of steady-state vowels*—in Auditory Analysis and Perception of Speech (ed. by G. Fant and M.A.A. Tatham), Academic Press, London (1975), p. 37.

[8] Ladeforget P.—*Three Areas of Experimental Phonetics*—Oxford University Press, London (1976).

[9] Ladeforged P., Broadbent D. E.—*Information conveyed by vowels*—Journ. Acoustical Society of America, vol. 29(1967), p. 98.

[10] Miller, R.L.—*Auditory tests with synthetic vowels*—Journ. Acoustical Society of America, vol. 25(1953), p. 114.

[11] Plomp, R.—*Auditory analysis and timbre perception*,—in Auditory Analysis and Perception of Speech (ed. by G. Fant. and M.A.A. Tatham).—Academic Press, London (1975) p. 7.

[12] Zwicker E., Feldtkeller R.—Das Ohr als Nachrichtenempfänger—Hirzel Verlag, Stuttgart (1976).

CHAPTER 5.2

SPEECH RELATED TO PURE TONE AUDIOGRAMS

INTRODUCTION

A prerequisite for speech perception is that a sufficient part of the speech signal is above the threshold of hearing. The extent to which this is the case, assuming reference speech power and spectral distribution at a certain distance to a human receiver, may be visualized by superimposing the spectral distribution of speech on a pure tone audiogram. An example is given in Figure 1, which is taken from Lidén and Fant (1954). The shaded frequency-intensity area is the reference distribution of speech power and is often referred to as the 'speech banana'. Although this representation of speech has been widely accepted and used in audiological circles, its origins are not well known and were not discussed in the Lidén and Fant (1954) work. The graph was the result of speech analysis work I carried out at the Ericsson Telephone Company in the period 1946–1949. Prior to this material being published (Fant, 1959) I made the results available to Erik Wedenberg who used them in his studies of auditory training programmes for severely hard-of-hearing children (Wedenberg, 1951, 1953). It is the purpose of this chapter to review the derivation and applications of the 'speech banana'.

THE SPEECH ANALYSIS DATA

One of the aims of my work at Ericsson was to provide a background for understanding the consequences of reserving parts of the available bandwidth in telephone lines for signalling purposes. The problem is analogous to the evaluation of the degradation caused by a specific hearing loss. In addition to direct perceptual tests it was considered important to study the actual frequency locations and intensities of the major formants of Swedish speech sounds.

The speech analysis was carried out with considerable care to preserve absolute calibration of the intensity data. A composite graph of speech formants from all Swedish vowels and consonants was constructed and this is shown in Figure 2 (from Fant, 1959). The frame is an intensity (sound pressure level) versus frequency diagram in which the standardized free field threshold of hearing and a 40 dB equal loudness contour are shown. All sound pressure levels pertain to a speaking distance of one metre. The French and Steinberg (1947) average speech spectrum, modified to represent intensity in successive 250 Hz bands, was also included. It fits well into the formant data which shows a distribution of approximately $+10$ dB to -20 dB around the long term average data. The maximum sound pressure levels were of the order of 65 dB corresponding to the first formant (F1) of vowels with F1 values around 500 Hz.

In Figure 3 (Fant, 1959) these data have been transformed to sensation levels above the standardized free field threshold at a distance of 1 metre and summarized to show regions occupied by voice fundamental frequency (F0), and the first, second, third

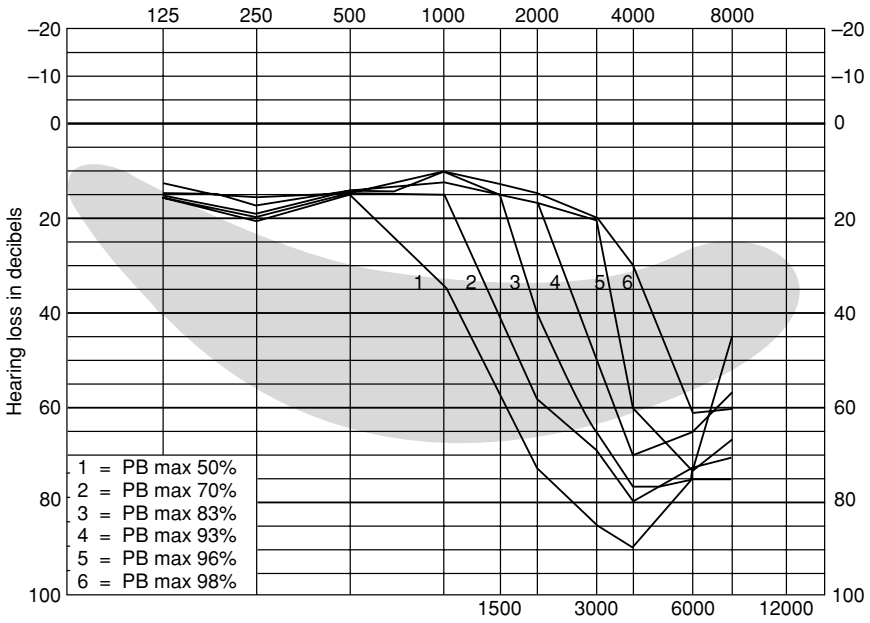


Figure 1. Maximum scores for PB (phonetically balanced word lists) in some cases of high frequency hearing loss. From Lidén and Fant (1954).

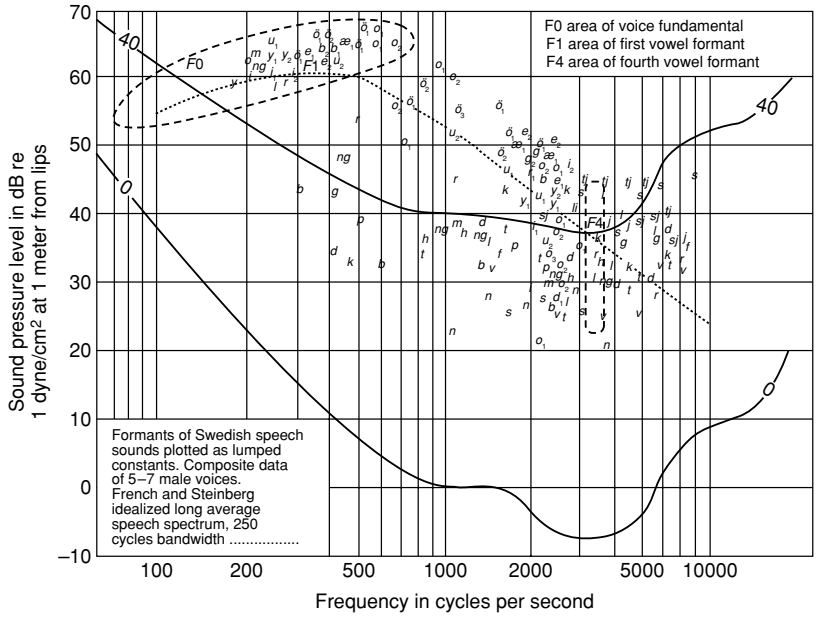


Figure 2. Sound pressure level versus frequency plot of Swedish vowel and consonant formant data. From Fant (1959).

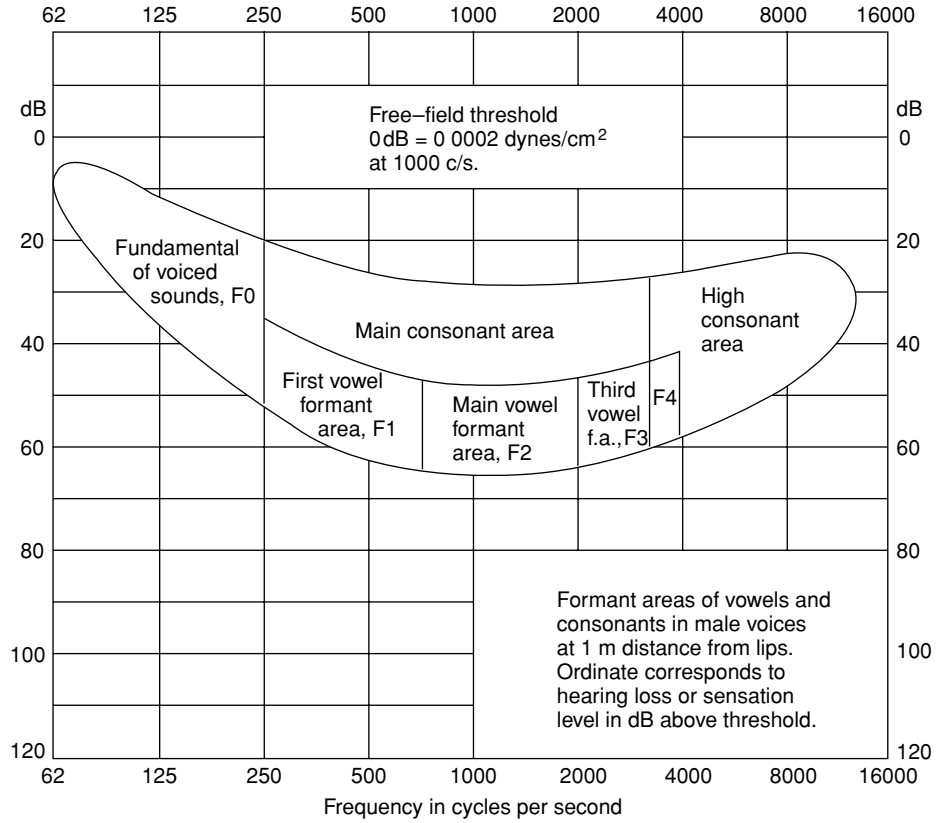


Figure 3. Speech spectrum data schematized in terms of formant areas. The ordinate in sensation level versus free field threshold at 1 metre distance. From Fant (1959).

and fourth formants. Also shown are the main consonant area and the high frequency consonants region. In audiological applications of this work the divisions are usually left out and only the outline contour of the 'speech banana' is retained.

APPLICATIONS

We can return to Figure 1. This includes data on articulation scores for various high frequency hearing losses derived from speech audiometry with phonetically balanced monosyllabic (PBM) word lists. It may be seen that only when a substantial part of the formant region is below the subject's threshold of hearing is there an appreciable reduction of the articulation score.

The potential value of the spectral information transmitted in a speech communication link is quantified by the Articulation Index (A_i, French and Steinberg, 1947) which includes the receivers hearing threshold. A_i, which equals 1 for an ideal communication system, and which can be calculated from the weighted sum

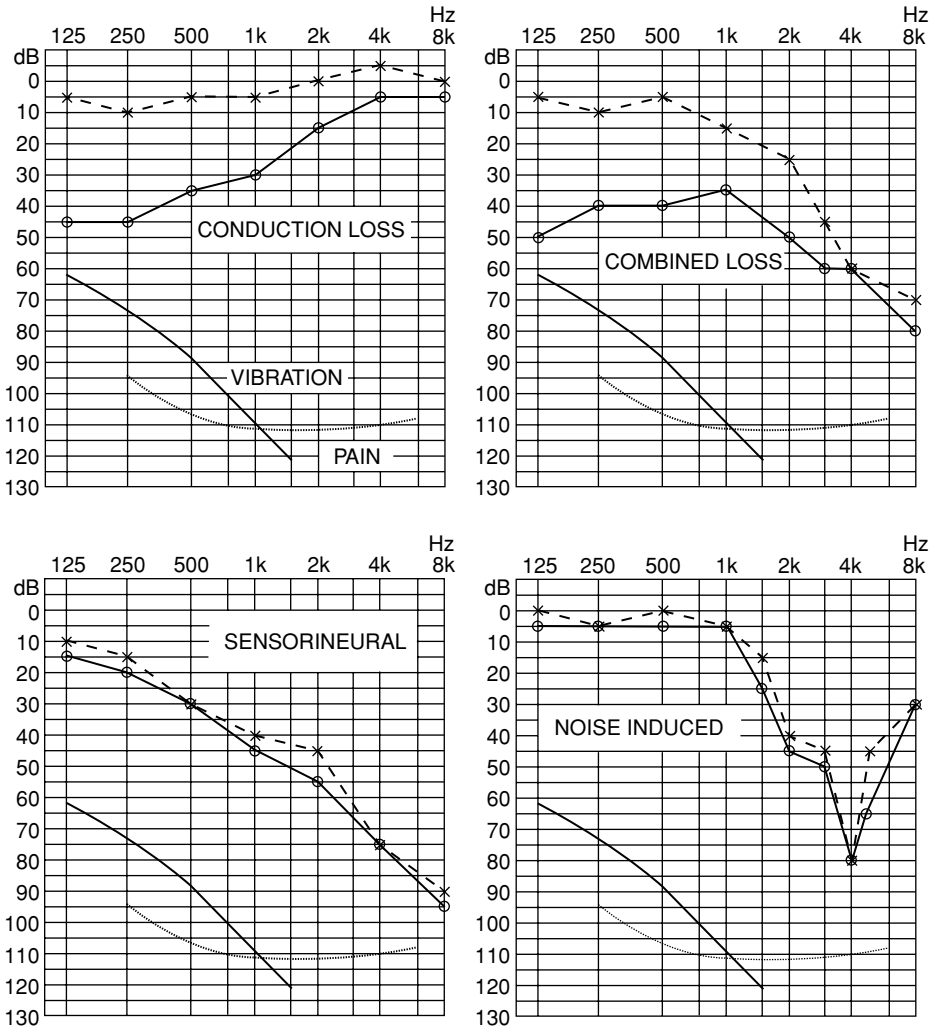


Figure 4. Pure tone audiograms illustrating four typical types of hearing loss. Crosses and broken lines pertain to supplementary bone conduction measures.

of contributions from each part band of the speech spectrum in proportion to its relative importance and the extent to which the intensity in a band exceeds the subjects masked or unmasked threshold. Thus,

$$A_i = \Sigma a_n W_n$$

where a_n is the relative importance of the band number n and W_n is an intensity factor which varies between 0 and 1, the latter when the sound pressure level in a band is 30 dB or more above the threshold.

The following tabulation derived from data published by Beranek (1947) shows the Articulation Index a_n per octave band centred at each of seven audiometric frequencies.

Frequency (Hz)	125	250	500	1000	2000	4000	8000
a_n (%)	2	7	14	23	32	19	3

Corresponding data for Swedish are not available but should not differ too much from these. Given a total Articulation Index the corresponding articulation score depends upon the specified test used, its vocabulary, and its level of difficulty (Beranék, 1947; French and Steinberg, 1947).

The simplified procedure above has been applied to typical hearing losses of four categories: (1) conductive loss, (2) sensorineural loss, (3) combined loss and (4) noise induced loss. The examples shown in Figure 4 are taken from a statistical survey I carried out in 1944 in connection with my electrical engineering thesis work. They were extracted from a sample of 377 audiograms taken in a Stockholm audiological clinic. Thirteen per cent were classified as conductive hearing loss, 39% as sensorineural, 37% as mixed whilst 11% were noise induced.

A calculation of the Articulation Index for the examples presented in Figure 4 yielded the following values. Conductive loss $A_i = 0.9$, sensorineural loss $A_i = 0.5$, mixed loss $A_i = 0.4$ and noise induced loss $A_i = 0.7$.

This is merely a formal exercise stressing the basic fact that a certain signal-to-noise or signal-to-threshold distance is required for speech reception and that the relative importance of different frequency regions varies. In addition we have to consider the qualitative impairments in several auditory functions and in more central stages of speech information processing and, cognitive functions. But this is another domain in which Arne Risberg and his associates contributed a lifetime of productive studies.

REFERENCES

Beranek, I. (1947) The design of speech communication systems. Proceedings I.R.E. 35: 880.

Fant, G. (1959) Acoustic analysis and synthesis of speech with applications to Swedish. Ericsson Technics No.1, 1–108.

French, N.R. and Steinberg, J.C. (1947) Factors governing the intelligibility of speech sounds. Journal of the Acoustical Society of America 19: 90–119.

Lidén, G. and Fant, G. (1954) Swedish word material for speech audiometry and articulation tests. Acta Otolaryngologica, Suppl. 116, 189–204.

Wedenberg, E. (1951) Auditory training of deaf and hard of hearing children. Acta Otolaryngologica, Suppl. XCIV.

Wedenberg, E. (1953) Auditory training of severely hard of hearing pre-school children. Acta Otolaryngologica.

CHAPTER 6

PROSODY

Our work on speech prosody started in 1986 and was then largely directed towards temporal aspects. It has gradually incorporated intonation and developed into rules for speech synthesis. In recent years, the ambition has been to cover broad aspects of production and perception including the foundation in the respiratory system.

Our studies of the duration of speech sounds, syllables, words, inter-stress feet and pauses have revealed regularities that underlie the notion of a quantal nature of speech timing, which is the theme of the first article (Fant and Kruckenberg, 1986).

In Swedish and other stress timed languages there is a clear alternation between stressed and unstressed syllables, which accounts for a quasi-rhythmical element. Although the duration of inter-stress intervals increase with the number of syllables or speech sounds involved, their average value in a few seconds of reading, about 0,5 seconds, appears to function as a temporal quantum for the planning of the duration of pauses at major clause and sentence boundaries. Pauses tend to occupy a finite number of such quanta, maintaining a rhythmical continuity. This is most often found in the speech of trained readers.

These quantal patterns are even more apparent in the reading of metrically structured poetry, as shown in the second article (Kruckenberg and Fant, 1983) which deals with the timing and intonational properties of iambic and trochaic verse.

Our studies of several subjects, males and females, reading a prose text have revealed a large individual spread in the realization of prosodic boundaries, with respect to where pauses are inserted, their duration and associated properties. Prepause lengthening alone is the common realization of less prominent phrase boundaries. On the other hand, there was a small individual spread in the duration of pauses between sentences, which averaged 1 second for prose reading and 0,5 seconds for news reading in support of our quantal theory. These data appear in the third article (Fant, Kruckenberg and Barbosa-Ferreira, 2003).

The fourth article (Fant and Kruckenberg, 2004) is a broad résumé of our work in prosody, incorporating a production base from studies of the voice source and the respiratory system. Syllabic and word prominence, continuously scaled from our prominence parameter RS, have been related to acoustic parameters. We have developed a unique method of intonation analysis, and a strategy for predicting complete prosodic realization in text-to-speech synthesis of Swedish. It has been successfully tested on an Mbrola platform. Our rule system incorporates some language universal features, which have enabled promising results in synthesis of French and English. The article also reviews an experiment on straight line versus smooth continuous representation of F0 contours.

Among additional work not reviewed in the four articles above is a study (Fant, Kruckenberg and Nord, 1991) based on the readings of a Swedish text translated into English and French, with a view of finding aspects related to the concept of stress timing versus syllable timing. A significant dimension is the difference in duration between stressed and unstressed syllables, which is greater for Swedish and English than for French. Our statistics employ separate averages for one- two- three and four phoneme syllables. In each case the difference is of the order of 100–120 ms and increases with the individual's degree of distinctiveness. An interesting finding is the stability of the duration of unstressed syllables across speakers and also across languages.

A backbone in our studies of prominence is the continuously scaled RS parameter. It was introduced already in Fant and Kruckenberg (1989), which also provides an all-round analysis of prosodic parameters, mainly related to durational properties, e.g. our early observations of quantal effects. Compensatory tendencies in speech tempo variations were noted. Early work on quantifying F0 and duration as correlates of stress and Swedish tone accents have been reported in Fant and Kruckenberg (1994).

Our first major article on prosodic analysis and synthesis was Fant, Kruckenberg, Gustafson and Liljencrants, (2002). Relevant data from this publication have been included in article number 4 (Fant and Kruckenberg, 2004).

Individual patterns in the realization of prosodic boundaries have been studied in some detail. We have performed a perceptual scaling in terms of a boundary prominence parameter, scaled from 1 to 5 (Fant, Nord and Kruckenberg, 1986; Fant and Kruckenberg, 1989; Fant, Kruckenberg and Nord, 1990).

SELECTED ARTICLES

[6.1] Fant, G. and Kruckenberg, A. (1996). On the quantal nature of speech timing, *Proc. ICSLP 1996*, 2044–2047. (Revised version).

[6.2] Kruckenberg, A. and Fant, G. (1993). Iambic versus trochaic patterns in poetry reading, *Nordic Prosody VI*, Stockholm, 1993, 123–135.

[6.3] Fant, G., Kruckenberg, A and Barbosa-Ferreira, J. Individual variations in pausing. A study of read speech. (2003). Proc. of the Swedish Phonetics meeting in Umeå, Phonum 2003.

[6.4] Fant, G. and Kruckenberg, A. (2004). An integrated view of Swedish prosody. Voice production, perception and synthesis. To appear in Gunnar Fant, Selected Writings.

ADDITIONAL READING

Fant, G., Nord, L. and Kruckenberg, A. (1986). Individual Variations in Text Reading. A Data-Bank Pilot Study. *STL-QPSR* 4/1986, 1–17.

Fant, G. and Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR* 2/1989, 1–83.

Fant, G., Kruckenberg, A. and Nord, L. (1990). Acoustic correlates of rhythmical structures in text reading. *Nordic Prosody V*, Turku, 1990, 70–86.

Fant, G., Kruckenberg, A. and Nord, L. (1991). Durational correlates of stress in Swedish, French and English. *Journal of Phonetics* **19**, 1991, 351–365.

Fant, G. and Kruckenberg, A.(1994). Notes on Stress and Word Accent in Swedish, *STL-QPSR* 2–3/1944, 125–144. Also published in *Proc. Int. Symp. on Prosody, 18 Sept 1994, Yokohama,* 19–36.

Kruckenberg A. and Fant G. (1995). Notes on syllable duration in French and Swedish. *Proc. ICPhS 95,* Vol II, 158–161.

Fant, G., Kruckenberg, A., Gustafson K. and Liljencrants, J. (2002). A new approach to intonation analysis and synthesis of Swedish, *Speech Prosody 2002, Aix en Provence.* 283–286. Also in *Fonetik 2002, TMH-QPSR* 2002, 161–164.

CHAPTER 6.1

ON THE QUANTAL NATURE OF SPEECH TIMING

ABSTRACT

This is a review of regularities we have observed in the analysis of text reading, mostly Swedish, directed to the timing of vowels and consonants, syllables, inter-stress intervals and pauses. We have found tendencies of quantal aspects of temporal structure, superimposed on more gradual variations, which add quasi-rhythmical elements to speech. A local average of inter-stress intervals of the order of 0.5 sec appears to function as a reference quantum for the planning of pause durations. A recent study, confirming our previous findings of multiple peaks with about 0.5 sec spacing in histograms of pause durations, provides support to this model. It is well established that pause duration tends to increase with increasing syntactic level of boundaries. However, these variations tend to be quantally scaled even within a specific boundary category, e.g. between sentences or between paragraphs. Relatively short pauses, as between phrases or clauses, show durations in complementary relation to terminal lengthening. There are indications of approximately 1, 1/2, 1/4, 1/8 ratios of average duration of inter-stress intervals, stressed syllables, unstressed syllables and phoneme segments which adds to the observed regularities. The timing of syllables and phonetic segments with due regard to relative distinctiveness and reading speed are discussed and also tempo variations within a sentence.

1. SEGMENTS AND SYLLABLES

The main source of data to be discussed here derives from our own studies [1–7]. The text was a passage of about 8 minute duration from a Swedish novel read by our reference subject, a Swedish language expert. A databank search system organized within a linguistic frame was developed for the processing. Our analysis has been concerned with individual vowels and consonants, syllables, inter-stress intervals and pauses. In addition we have data from 15 other subjects reading a limited part of the text. Durations were measured by hand from broad band spectrograms.

The concept of quantally structured duration data is not new. Gårding [8], in a study of contrastive prosody, proposed a timing model for read Swedish in which the duration of an unstressed CV-syllable is the unit. Syllables with either a long vowel or a long consonant, i.e. stressed syllables were given two such units and phrase final lengthening one extra unit.

Our accumulated experience from speech analysis, including a recent unpublished databank survey, allows a more extensive modelling. We find a clear tendency of factor 2 relations between major categories. Inter-stress intervals, measured from the onset of the vowel in a stressed syllable to the onset of a vowel in the next stressed syllable, excluding those spanning a pause or a syntactic boundary, averaged 540 ms. The average duration of primary stressed syllables as well as those of secondary stress in compound words was 270 ms. Unstressed syllables averaged 132 ms. Mean phoneme duration was 70 ms. Unstressed vowels averaged 59 ms and unstressed consonants 51 ms. There are thus approximately 1, 1/2, 1/4, 1/8 relations in the timing of inter-stress intervals, stressed syllables, unstressed syllables and phonemes.

The data above refer to contexts excluding prepause locations. Within this regular frame there exists a continuity of variations of segment duration and positional variants, but one still finds regularity traits. Thus, consonants after short stressed vowels are of about twice the length of unstressed consonants, which holds for voiced as well as for unvoiced consonants, according to [1] a ratio of 87/44 for voiced and 135/67 for unvoiced consonants.

A basic distinction in Swedish phonology is that of "quantity". A stressed syllable contains either a long or a short vowel. A stressed short vowel is followed by a long consonant or vice versa. The relation of the duration of a long stressed vowel to a short stressed vowel is not 2 to 1 but of the order of 1.6 to 1. Lexically stressed vowels in function words generally lose their stress in connected speech. As a mean trend over all contexts and tempos and several data corpora we derived in [1] a relation between long and short stressed vowels of

$$V_{long} = 1.9V_{short} - 45 \text{ ms} \quad (1)$$

The duration distinction is lost when V_{short} approaches 50 ms. A fully stressed VC: is about 10 % shorter than a V:C and of the order of 210 ms.

The average number of phonemes per syllable is close to 2.9 for stressed and 2.2 for unstressed syllables, but text specific variations occur. In our standard prose passage we noted 3.0 phonemes per stressed syllable. Alternatively, with a non-conventional definition of syllables, constrained by root morphemic criteria such that the word "legat" would be segmented as [leg-at] opposed to the conventional [le-gat], the average number of phonemes per stressed syllable in the corpus was found to increase to 3.3. One argument in favour of the morphemic definition would be that the duration of the consonant following the stressed vowel is lengthened. In our statistics, retaining the conventional definition of syllables, we have accordingly introduced a special category for initial consonants of unstressed syllables that are preceded by a stressed vowel in an open syllable. The mean duration of such syllables is 192 ms and the average number of phonemes is 2.55, i.e. substantially greater than for the main category of unstressed syllables.

2. PAUSES. RHYTHMICAL CONTINUITY

Lea [9] introduced the concept of rhythmical continuity of stress intervals (feet) spanning a pause, stating that mean values of such intervals equaled an integer of the average duration of inter-stress intervals that are not interrupted by a syntactic boundary. From readings of the Rainbow Passage he noted quantal steps of 0.5 seconds. Pauses before clauses averaged 0.5 seconds and before a new sentence 1 second, which implies pause spanning feet of 1 and 1.5 seconds.

This model was further developed by us [1]. We noted a complementary relation between the duration of pauses of the first quantal order within sentences and prepause lengthening.

As shown in Fig. 1 from [4] the sum of these two components tend to match the average free-foot duration derived from a short time memory span of about 8 free feet, or 4 seconds

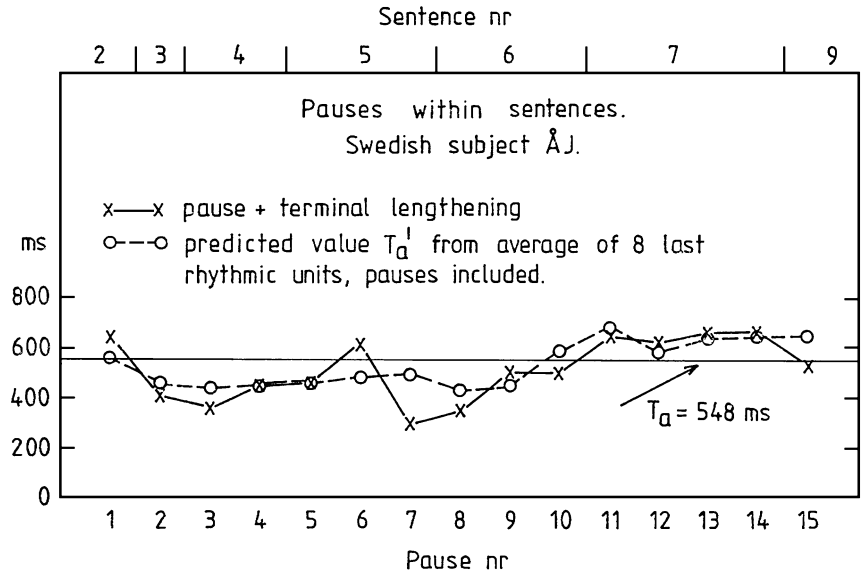


Figure 1. Inter-sentence pause duration plus final lengthening compared to local, 8 feet average inter-stress intervals.

We have observed quantal distributions of pause duration within one and the same boundary category, for sentence as well as for paragraph endings. These findings are speaker specific, some producing more rhythmically coherent patterns than others, and some favoring a larger number of quanta than others. An example from [4] is shown in Fig. 2.

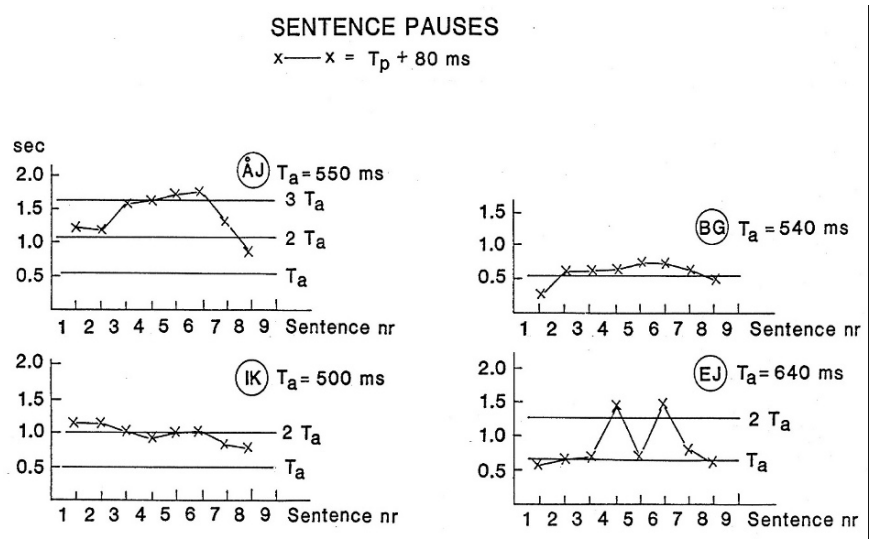


Figure 2. Examples of four Swedish subjects inter-sentence pauses showing quantal tendencies.

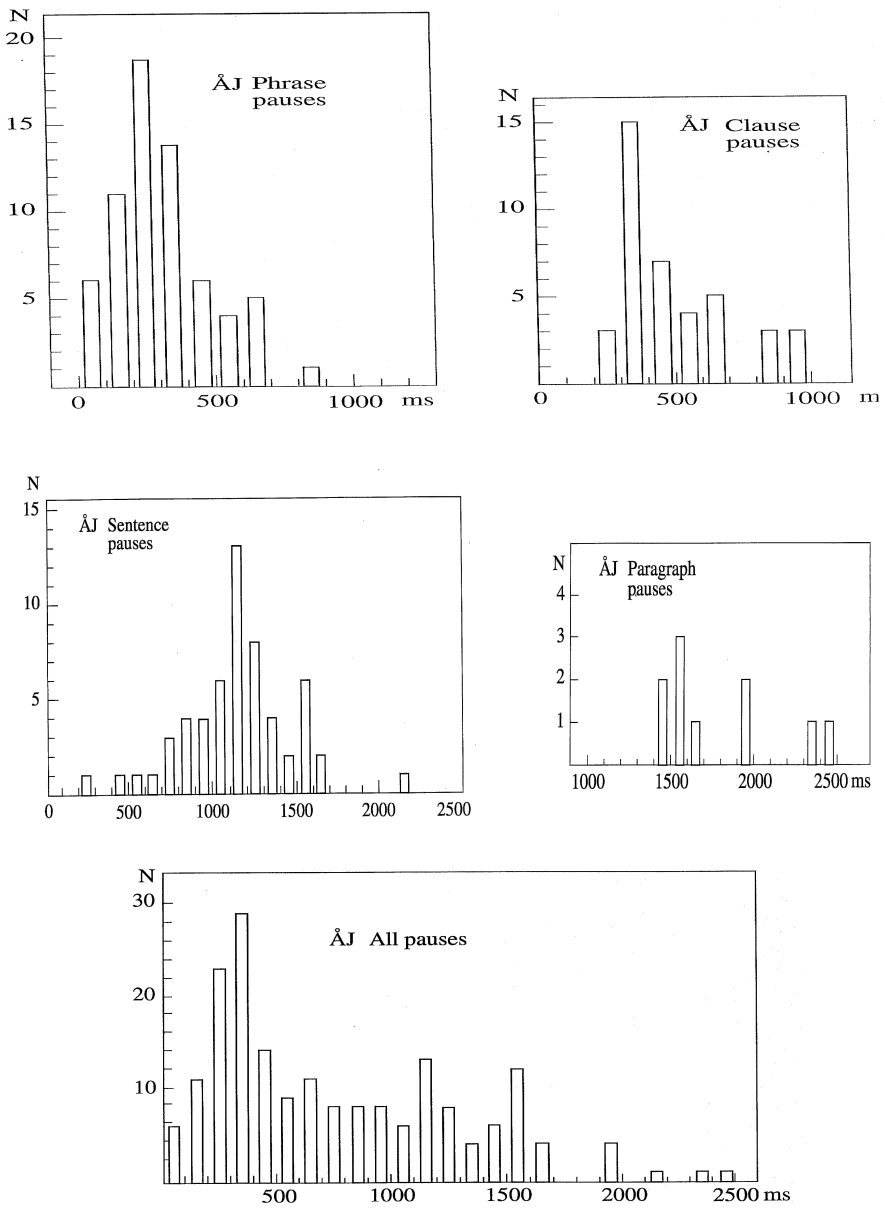


Figure 3. Phrase, clause, sentence, paragraph and all pauses in an 8 minute long text reading, subj ÅJ.

Here the duration of the pause plus a standard value of prepause lenghtening equals an integer of the subject's average inter-stress interval. We have also exemplified such trends for English as well as for French, [4].

Examples of multi-mode pause durations are illustrated in Fig. 3. pertaining to our reference speaker and in Fig. 4 to some recently collected data for a female

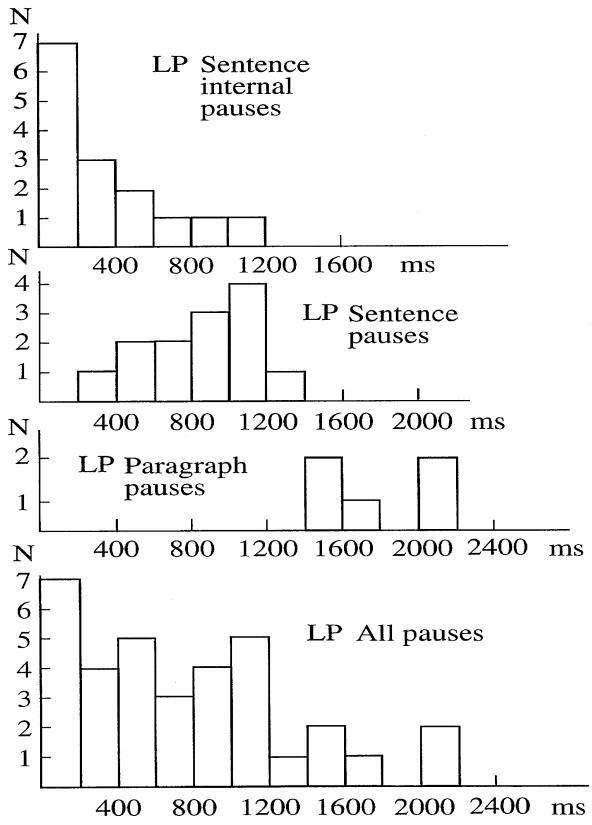


Figure 4. Sentence-internal, sentence, paragraph and all pauses in a 3 minute introductory reading for a Linguaphone course.

subject introducing a Linguaphone course. Distances between peaks are of the order of 0.5 seconds. Observe how this trend is apparent in terms of pause durations of 2 and 3 quanta and at paragraph boundaries 3, 4 and 5 quanta.

Strangert [10] found pause duration systematically increasing with syntactic level comparing phrase, clause, sentence and paragraph boundaries. These are comparable to our data, but the possibility of quantal effects are hidden in the averaging process.

An analysis of histograms supplied by Heldner and Strangert [11] also shows some trends of bimodal distributions but they are less apparent than in our reference data.

An additional support for the quantal nature of pausing derives from data on breathing, [13], see [1] page 36. In this study sentence pauses showed a histogram peak at 500 ms without breathing and 1000 ms with breathing. Clause boundaries peaked at 500 ms with breathing and at 300 ms without breathing.

3. INTER-STRESS INTERVALS

It is by now well established that the duration of an inter-stress interval (foot) T_n increases with the number of phonemes, n, or with the number of syllables, m, in the foot.

For a passage read by our reference subject, excluding boundary spanning feet, we noted

$$T_n = 158 + 53n \tag{2}$$

where n is the number of phonemes in the foot. Alternatively, in terms of the number of syllables, m

$$T_m = 190 + 120m \tag{3}$$

The average foot length was $T_n = 548$ ms corresponding to $n = 7.5$ phonemes or $m = 3$ syllables

A sequence of stress beats is thus quasi-rhythmical only [14], and the standard deviation is of the order of 40%. However, the prosodic importance is considerable for stress timed languages such as Swedish.

From our databank study of 8 minutes of prose reading we recently found average values of final lengthening of unstressed syllables ranging from 95 ms for junctures without a pause, 95 ms for sentence internal boundaries, 70 ms for sentence boundaries and 35 ms for paragraph boundaries. Final lengthening of stressed syllables was of the order of twice that of the unstressed values above. These data are comparable to those in [1] where we reported a mean value of final lengthening of $T_f = 110$ ms for within-sentence pauses of the order of 300–500 ms and more specifically

$$T_f = 190 - 0.2T_p \quad (R = 0.4) \tag{4}$$

Clause or sentence initial shortening was found to be of less magnitude and of the order of 40 ms for stressed and 15 ms for unstressed syllables.

Ideally, our model of rhythmical continuity across a pause implies that the sum of pause duration and final lengthening (initial shortening or lengthening included) of segments within a pause spanning foot equals an integer of the average foot length. The remaining part of the foot containing segments before and after the pause (excluding preboundary lengthening and postboundary shortening effects) has a duration which is determined by the number of phonemes according to Eq. 2 and which averages a free foot quantum.

However, as confirmed in [12] the complementary relation of final lengthening and pause duration appears to hold for relatively short pauses only. The decrease of final lengthening before longer pauses should also be considered. More extensive data are needed for a refinement of these durational models.

4. Distinctiveness and Tempo

The durational contrast between stressed and unstressed syllables is a major correlate to distinctiveness [2,5,7]. A part of the contrast derives from the larger average number of phonemes per syllable in stressed than in unstressed syllables. However, in Swedish the major part of the contrast, of the order of 100 ms, derives from the lengthening of vowels and consonants in stressed positions.

For our reference subject we noted an increase of the unstressed syllable duration Du with the number of phonemes n in the syllable by a linear regression [5,7].

$$D_u = 9 + 51n \tag{5}$$

Correspondingly for stressed syllables

$$D_s = 62 + 72n \tag{6}$$

In a more distinct reading mode, stressed syllables increased relatively more than unstressed syllables which remained rather stable. The relative constancy of unstressed syllable duration also holds true of individual variations.

The stressed/unstressed contrast is speaker and language dependent. It is smaller in French than in Swedish [7] and is largely carried by stressed syllables. The statistics for unstressed syllables were rather similar with respect to both speakers and languages. The same trend is also maintained comparing lower tempo and normal tempo speech. However, there is a reversal in fast speech where the unstressed syllables are relatively more reduced than stressed syllables [3].

4.1. *Tempo Variations*

The local tempo in terms of average segment duration within a sentence or a phrase is considerably influenced by the density of content words and thus of potential stresses within the text. In addition there exist deviations from normally predicted duration that reflect reductions and expansions around and within focal regions. Such deviations tend to cancel within a sentence, [3, 5] and reflect a finite pulmonary and articulatory energy at disposal [15]. In addition, alternating slowing down and speeding up of the tempo within a paragraph adds to the naturalness of reading.

5. Concluding Remarks

We have demonstrated two regularity aspects of speech timing. (1) Quantal steps of the order of 500 ms in pause duration related to the average duration of inter-stress intervals.

(2) Average duration of stressed syllables, unstressed syllables and phoneme segments are of the order of 250 ms, 125 ms and 62.5 ms. which suggests 1/2, 1/4 and 1/8 ratios of the basic 500 ms quantum.

(3) The trend of rhythmical continuity across pauses is speaker specific in manifestation.

(4) The choice of quantal level is influenced by syntactic criteria, breathing and individual habits.

Much more work could be devoted to problems of statistical significance and influence of tempo and reading style and specific language dependencies.

6. ACKNOWLEDGEMENTS

This work has been financed by grants from the Bank of Sweden Tercentenary Foundation, the Swedish Council for research in the Humanities and Social Sciences and the Carl Trygger Foundation.

Gunnar Fant and Anita Kruckenberg

REFERENCES

[1] Fant, G. and Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style, *STL-QPSR* 2/1989, 1–83.

[2] Fant, G., Kruckenberg, A. & Nord, L. (1991). Prosodic and segmental speaker variations, *Speech Communication* 10, 521–531.

[3] Fant, G., Kruckenberg, A. and Nord, L. (1991). Some observations on tempo and speaking style in Swedish text reading. *ESCA Workshop on "The phonetics and phonology of speaking styles",* Barcelona.

[4] Fant, G., Kruckenberg, A. and Nord, L.(1991). Stress patterns and rhythm in the reading of prose and poetry with analogies to music performance. In J. Sundberg, L. Nord, R. Carlson (eds.), *Music, Language, Speech, and Brain*, Wenner-Gren International Series Vol. **59**, 380–407.

[5] Fant, G., Kruckenberg, A. and Nord, L. (1992). Prediction of syllable duration, speech rate and tempo, *Proc. ICSLP 92,* Banff, Vol **1**, 667–670.

[6] Kruckenberg, A. and Fant, G. (1993). Iambic versus trochaic patterns in poetry reading. *Nordic Prosody VI*, Stockholm, 123–135.

[7] Kruckenberg, A. and Fant, G. (1995). Notes on syllable duration in French and Swedish. *Proc. XIIIth ICPhS* , 158–161.

[8] Gårding, E. (1981). Contrastive prosody: a model and its application. AILA Congr. 181. *Studia ling*. 35 146–166.

[9] Lea, W.A. (1980). *Trends in Speech Recognition*, Prentice Hall, Inc.

[10] Strangert, (1991). Pausing in texts read aloud. *Proc. XIIth ICPhS* Vol. **4**, 238–241.

[11] Heldner, M. & Strangert, E. (1996). Personal communication of data.

[12] Horne, M, Strangert, E and Heldner, M. (1995). Prosodic boundary strength in Swedish: final lengthening and silent interval duration. *Proc. XIIIth ICPhS,* Vol. **1**, 170–173.

[13] Base, A. (1983) *Pauser i tal*, EE thesis work, KTH, dept. of speech, music and hearing.

[14] Lehiste, I. (1977). Isochrony reconsidered. *J. Phonetics* **5**, 253–263, 1977.

[15] Öhman, S. (1967) Word and sentence intonation: a quantitative model. *STL-QPSR* 2–3/1967, 20–54.

CHAPTER 6.2

IAMBIC VERSUS TROCHAIC PATTERNS IN POETRY READING

INTRODUCTION

What are the characteristic features of the reading of poetry? What makes us feel and understand that what we listen to is just poetry? It is in order to find an answer to such questions that we have devoted some time and energy to the analysis of poetry reading—a research that we have reported on in earlier articles (Kruckenberg, Fant & Nord, 1991A, Nord, Kruckenberg & Fant, 1990, Fant, Kruckenberg & Nord, 1991A, Kruckenberg, Fant & Nord, 1991B).

In Saussure's view, the linguistic sign has two faces—the signifier and the signified, in French "le signifiant" and "le signifie'"—or, if expressed more simply, one side concerning form and one side concerning content. If you look upon language as a means of communication, as in ordinary prose, le signifie', i.e. the message and its implication, is the most important part. In Roman Jakobson's poetics, his theory on the means of poetic expression, the concept of le signifiant is just as important as the signifie', just as significant, since here form underlines and stresses and sometimes gives a particular distinction to the message. Moreover, the formal structure and organization of poetry conveys to us the idea that what we are dealing with is a poetic text, a poetic sign.

This is the foundation of Jakobson's idea of *the poetic function of language* which manifests itself in the structure of parallelisms on various linguistic levels, based on a principle of equivalence in relations of similarity and dissimilarity that can be seen in all his analyses of poetry, (Boström-Kruckenberg, 1979).

What we want to do is working along similar lines concerning the *reading* of poetry, and looking for *the poetic function of speech.*

When the reader knows that he is reciting poetry, he assumes a particular attitude—what Roman Jakobson refers to as "Einstellung"—towards the poetic text, and his speech changes in various ways; it is adapted, becomes more expressive, more significant. There are many aspects to be considered, and a few of them will be discussed here.

OUR STUDY

Two lines of approach have appeared to be important in finding out what makes us perceive a reading as poetical. One of them is of course a comparison with read prose, but we must also know the characteristic features that distinguish the poetic *text*, the written poem, from prose. We must also consider the poetic function of *language* in the text of the poem—what was pointed out above as one of Roman Jakobson's many great achievements.

As to the poetic function of language we will here only mention some problems we have approached so far. We ask ourselves how well the text of a poem agrees with the *meter* it is supposed to follow. What about one- and two-syllable words in an iamb or a trochee? What about the number of function words and content words in each of these meters? What about the origin of the unstressed syllable in one- versus two-syllable words, or the occurrence of accent 1 and accent 2 in various metrical positions?

As to the *choice of words* we intend in later studies to look into rhythmical figures in the combination of words conditioned by the meter, as well as into patterns of similarity of sounds in the choice of words in relation to the poem's structure of stressed and unstressed syllables.

As to *syntax* we will—also later on—look into the change of word-order in poetry as well as "the constraint of rhyme".

In short—there is a wide field for analysis of the structure of poetic texts in the spirit of Roman Jakobson.

But what about the poetic function of *speech*? Here the field of research is at least equally wide and rich, with problems dealing with rhythm, meter, pause, tempo, F0, distinctiveness of pronunciation, pitch, intensity, etc.

To sum up, we have so far taken interest in three aspects of performance:

1. Reading style, i.e. recital of poetry versus reading of prose.
2. Rhythmical regularity, continuity and phrasing.
3. Meter specific sound patterns of iamb and trochee.

Let us start with reading style. What is characteristic of the reading of poetry?

Our reference subject, ÅJ, was asked to read two Swedish poems, the trochaic "Näcken" (The water sprite) by E.J. Stagnelius and the iambic "Karl XII" (Charles XII) by E. Tegnér, first as poetry, then as if they had been written in prose.

Näcken.
Kvällens guldmoln fästet kransa.
Älvorna på ängen dansa,
och den bladbekrönta Näcken
gigan rör i silverbäcken.

Karl XII.
Kung Karl, den unga hjälte,
han stod i rök och damm.
Han drog sitt svärd från bälte
och bröt i striden fram.

What makes us feel that we are listening to poetry when he reads the poetic version? There are many aspects, but we have noticed the following features of poetry reading in comparison with prose reading. In the poetic reading the tempo, assessed by average syllable duration, is lower and more stable than in prose reading. Intensity and F0 are higher, while the modulation depth in local F0 variations is smaller in poetry than in the prose reading. We get a higher, more stable pitch, a "recital tone".

One can perceive a more even and stable rhythm in the poetic version—maybe also that pauses fit into the metrical pattern and support the rhythmical continuity across the lines.

EXPERIMENTAL RESULTS

Let us now turn to our measured data. From spectrograms with simultaneous registration of F0 and intensity we have measured the duration of conventional syllables, metrically classified as S = strong and W = weak, combined into metrical feet, that is to say S+W for trochaic meter (Näcken) and W+S for iambic meter (Karl XII). To some extent we have also measured interstress intervals, from the onset of the vowel of a stressed syllable to the onset of the vowel of the next stressed syllable. Such V-V intervals will primarily be used when studying rhythmical continuity across lines.

Tempo and Isochrony

For comparison we have data from the reading of normal prose, the novel "Änglahuset" by Kerstin Ekman (Fant and Kruckenberg, 1989). A general observation is the lower tempo in the recital of the poems, about 40% in average phoneme duration, compared to the reading of the novel.

In spite of the lower tempo, the duration of a foot in the reading of a trochaic or iambic poem is roughly the same as that of a stressed foot in normal prose, about 550 ms. This is explained by the fact that a stress foot in prose contains on the average one strong and two weak syllables, thus one unstressed syllable more than the iamb or the trochee.

The duration of a metrical foot shows somewhat less spread than a corresponding unit in prose. We noted a standard deviation of the order of 30% in poetry versus 40% in prose reading. The spread is basically related to the number of phonemes in a foot, which is more stable in poetry than in prose. It is also influenced by the relative emphasis and position of the foot in a phrase or a verse, i.e. a metrical line.

Systematic variations within a line are brought out in Fig. 1. In these examples we have omitted feet that are metrically inverted or otherwise depart from the meter. A general trend in both poems is that the length of the foot decreases with its position within a line. The first foot is longer than the second, and so on. This declination is broken by a final lengthening in the last foot of the line. In the poem "Karl XII", it is more pronounced than in the poem "Näcken". The same tendency of declination appears in the plots of average phoneme duration within successive feet.

Rhythmical Continuity Across Lines

How, then, is the transition performed from the end of one line to the beginning of the next one? Can we find a rhythmical continuity, and if so, how should it be described and explained? This is not the place for an exhaustive analysis. We will have to confine ourselves to stating that rhythmical continuity across the lines is

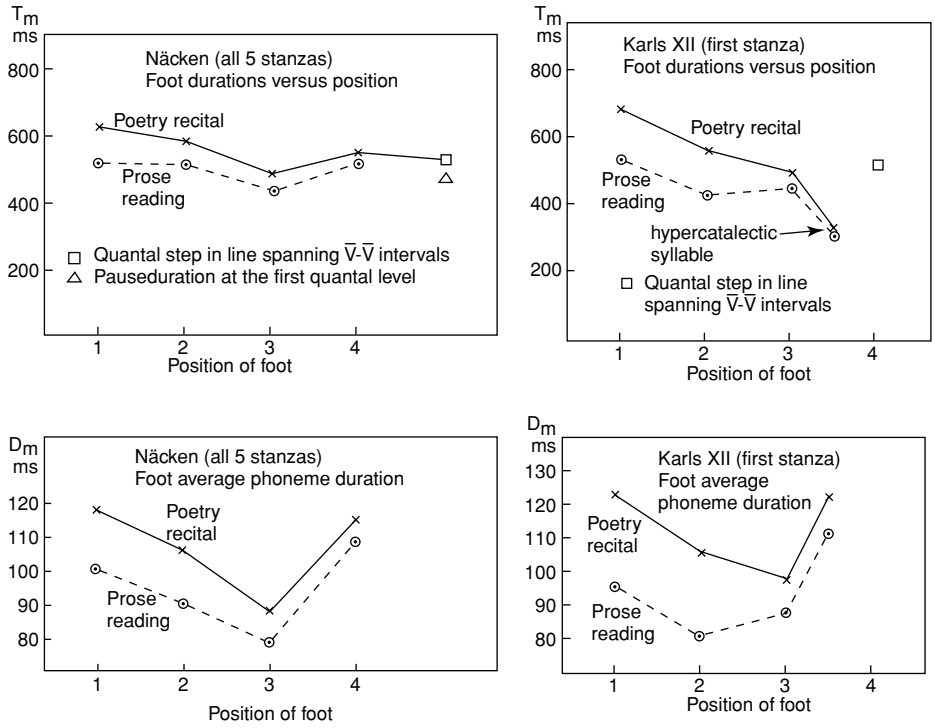


Figure 1. Metrical foot duration (above), and (below) foot average phoneme duration, versus position in a line. Poetic recital and prose reading of the trochaic poem (left) and the iambic poem (right).

best described by referring to V-V intervals that cover the last stressed vowel of a line and the following phonemes up to the end of the line, plus the pause, plus the phonemes that precede the first stressed vowel of the following line. On the average, V-V intervals within lines come close to average foot durations, but they appear to be more directly related to the sequence of stress beats that convey the basic rhythm. However, within lines, we prefer to refer to metrical feet as units.

Rhythmical continuity across lines implies that the duration of a spanning V-V interval is harmoniously related, either to some kind of average foot in reading poetry, or to the nearest preceding feet of the line in question. We have not been able to find such direct dependence within the line, but we have evidence of the spanning interval favouring certain quantal levels of 1, 2, 3 or 4 times a basic interval of about 525 ms, which is in better quantitative accordance with the average values of the last feet of a line than with an average foot, see Fig. 1 and 2. It seems that the internal clock of the speaker controls a stable average foot for the entire reading of a poem, which assures that local accelerations and retardations within the stanza average out. It may be possible that the line-spanning stress interval is directly controlled by this internal clock. Propositions along these lines were earlier put forward by Fant, Kruckenberg and Nord (1991A) and by Kruckenberg, Fant and Nord (1991B) and

concerning prose reading by Fant and Kruckenberg (1989) and by Fant, Kruckenberg and Nord (1990).

The rhythmical continuity across lines contributes to the isochrony in the reading of poetry. This tendency is supported by the inverse relation between the duration of the pause and the number of phonemes that are included in the spanning interval. When the number of phonemes increases, the duration of the pause tends to decrease. One example is the hypercatalexis, the metrically surplus, unstressed syllable in every second line of the poem "Karl XII", which, as far as we can judge from this limited material, conditions shorter pauses. This compensatory mechanism does not exist in prose. In prose reading the duration of pause-spanning intervals depends more on the number of phonemes than is the case in poetry reading.

Poetical Recital Versus Prose Reading

We will now make a more general survey of poetry reading versus prose reading, as well as of iambic versus trochaic reading of poetry. An external reference is the reading of the earlier mentioned novel. Here, the tempo calculated on the basis of average phoneme duration was about 20% faster than in the prose versions of the two poems. An important part of this difference relates to the more frequently occurring unstressed syllables in the novel.

If we now compare the poetic versions of the two poems with their prose versions, we can discern a common feature. When passing from prose to poetry reading, the duration of stressed syllables increases much more than that of unstressed syllables. An increased S/W (strong/weak) contrast follows. The same phenomenon has been observed earlier in clear and slow speech in comparison with normal speech (Fant, Kruckenberg & Nord, 1991B).

Meter Specific Features

What about meter specific differences between iambic and trochaic reading? We have earlier verified the statement of Newton (1975) that the weak syllable of the trochee is considerably longer than the weak syllable of the iamb, while the strong syllables of the trochee and the iamb are more equal, with perhaps some predominance for the iamb.

As illustrated in Fig. 2, we find a strong/weak durational ratio of S/W = 1.7 for the trochee and S/W = 2.5 for the iamb. This contrast promotes the more vivid impression we get from iambic verse than from trochaic verse. We have also shown that these durational patterns to a substantial part derive from the language material

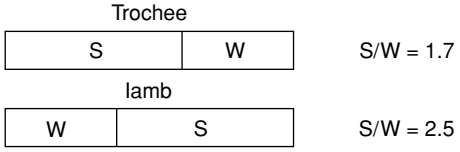


Figure 2. Temporal patterns of trochaic and iambic metrical feet.

selected for the poem. The weak syllable of the trochee contains on the average 2.62 phonemes while the weak syllable of the iamb averages 2.28 phonemes.

However, there remains an important difference partly dependent on other factors than syllable composition, i.e the final lengthening within the foot that affects the weak syllable of the trochee and the strong syllable of the iamb. This is especially apparent in line final positions. From a methodological point of view we have the possibility of eliminating the influence of the language material in the text by comparing separetely the prosaic and the poetic readings of the iambic and the trochaic poems. A closer analysis shows that the shift from the prosaic to the poetic version involves an increase of the duration of stressed syllables by 16.5% in "Näcken" and as much as 19% in "Karl XII". At the same time unstressed syllables show an increase of 6.5% in "Näcken" but only 2% in "Karl XII". Accordingly, the S/W (strong/weak) ratio of the iamb has gained in all by 6.5% compared to the trochee. This durational difference may be interpreted as a component of a meter specific pronunciation pattern.

Other meter specific characteristics may be derived from F0 and intensity data. A general precaution in all our data analyses has been to disregard passages of pronounced deviations from the metrical pattern such as inversions and emphatic stress on metrically weak syllables. In "Karl XII" such deviations are frequent.

In the spectrogram, Fig. 3, we find the first line of the trochaic poem "Näcken", at the top read as poetry and at the bottom read as prose. We may note an overall

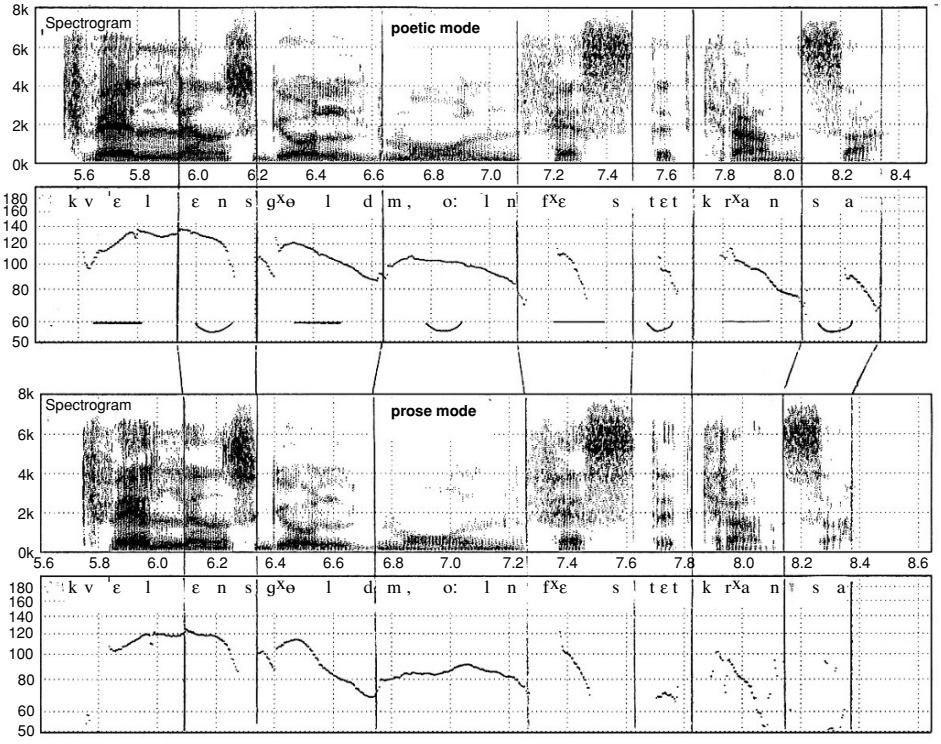


Figure 3. First line of the trochaic poem. Poetic recital and prose reading of the same text.

higher FO level in the poetic version, as well as a more stable FO contour. The FO of the second syllable of the accent 2 words "guldmoln", "fästet", and "kransa" have been raised to the same level as the main syllable. Another apparent feature is that the accent 2 FO drop in the main syllables of these words is reduced. Thus, the local FO modulations imposed by the word accent are less apparent in the poetic version. An obvious and general difference is found in the final syllable of the line, where the prose version shows a marked drop of FO, while the poetic version maintains a high FO, a continuation tone.

What has been said here about FO also applies to the intensity contour which is more even in the poetic version than in the prose version, see Fig. 4. Here the difference is enhanced by the short pause in the middle of the line of the prose version.

Readings of the iambic poem are illustrated in Fig. 5 and 6. All syllables in these examples derive from accent 1 words which are optimal for iambic verse. The spectrogram at the bottom of Fig. 5 shows the prose version of the two feet: "*Han* drog *sitt* svärd", (He drew his sword). Here we may observe the typical pattern of accent 1 with FO lower in the stressed syllable than in the preceding unstressed syllable, a well-known feature of Swedish as described by Bruce(1977). Thus in a stylized FO notation: "*Han* drog *sitt* svärd". In the poetic version we observe a different and partially inverted pattern, "drog" has higher F0 than "han", and "svärd" has an FO level in parity with "sitt", thus "han *drog* sitt svärd". In contrast to the prose version, the reduced accent 1 FO drop of the poetic version conditions a more level FO contour, an instance of reduction of the word tone modulation depth that we have already observed in the trochaic poem. Moreover, the high FO on the word "drog" indicates an emphasis which contributes to the rising character of the iambic meter. A rising trend within the foot was also found in the intensity contour of the utterance.

The spectrogram in Fig. 6 confirms these observations. The text is: "å bröt i striden fram", (and burst into the battle) The lower FO span of accent 1 in the poetic version is apparent, e.g. "i striden", as well as the continuation tone in the poetic version and the final fall in the prose version, thus the same phenomena that we observed in Fig. 6 and in the trochaic poem, Fig. 4. However, this does not apply to the end of the stanza, which is marked by an FO drop, both in the prose and in the poetic version of the two poems.

A common conclusion from these experiments with poems read in a poetic recital mode compared to a prose reading mode is a reduction of word accent modulation depth. This general finding also sheds light on meter specific characteristics, i.e. a relative emphasis of the weak syllable of accent 2 trochaic words and of the strong syllable of accent 1 iambic words. In both instances it is the second syllable of the foot, the W of the trochee and the S of the iamb, that is perceptually enhanced.

The analysis of intensity contours has provided supporting evidence. A stressed syllable in the reading of prose or poetry has, on the average, about 1–2 dB higher intensity than an unstressed syllable, but there is also a meter specific trend of a somewhat greater strong/weak intensity contrast in the iambic than in the trochaic foot. In other words, the weak syllable of the trochee and the strong syllable of the iamb show a relative increase of intensity. This strengthens the meter specific differences we have already observed in FO and in duration.

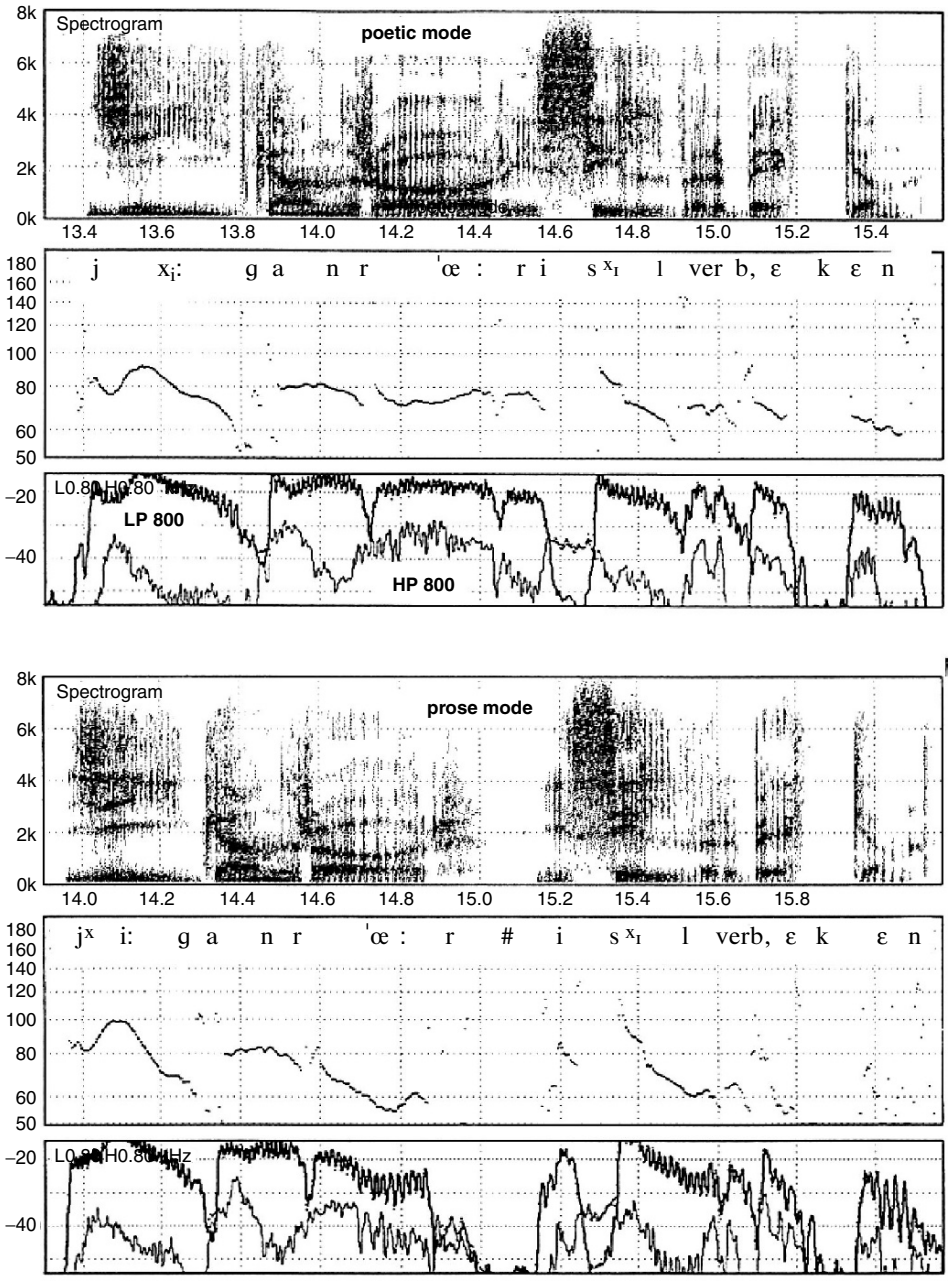


Figure 4. Spectrogram, F0 and intensity curves of a line within the trochaic poem. Poetic recital (above) and prose reading (below).

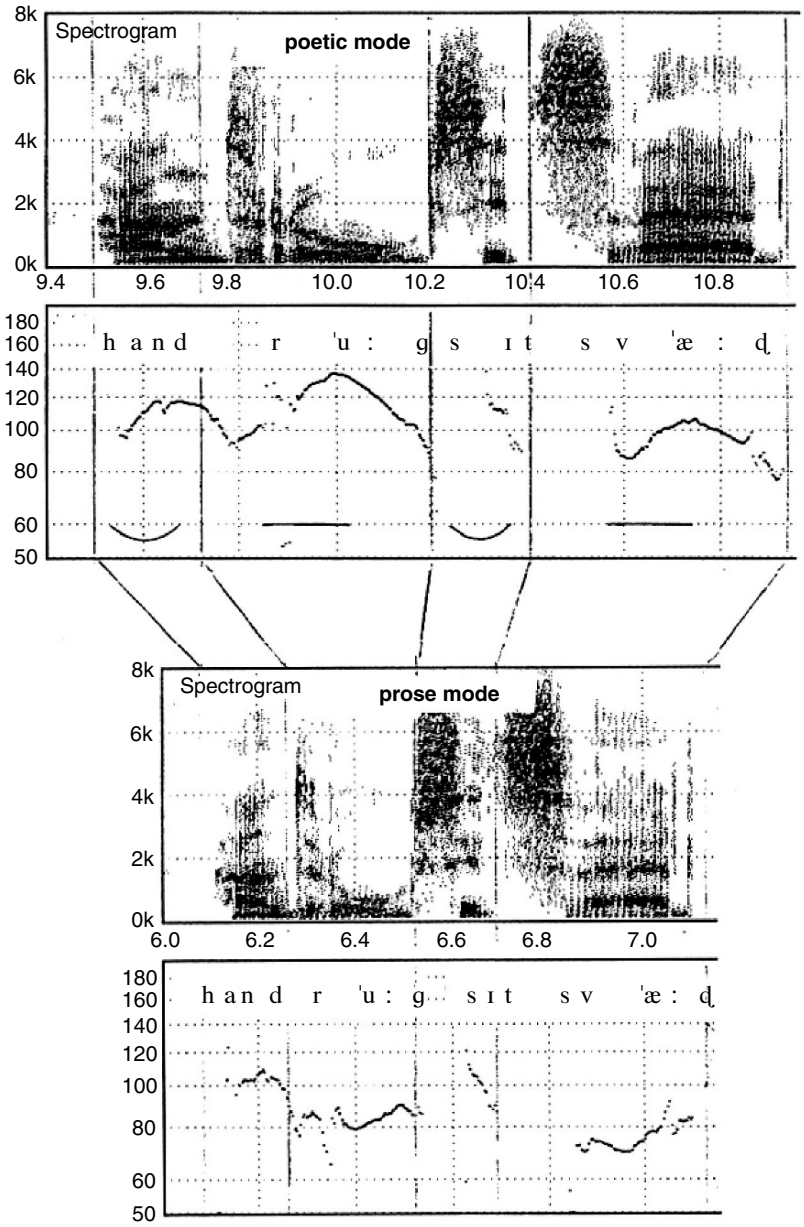


Figure 5. Two successive feet from the iambic poem. Poetic recital (above) and prose reading (below).

DISCUSSION

We have thus established that the reading of traditional Swedish poetry differs from the reading of prose in several respects, which, taken together, define a style of recitation. It includes slower tempo, a higher voice intensity level and greater clarity

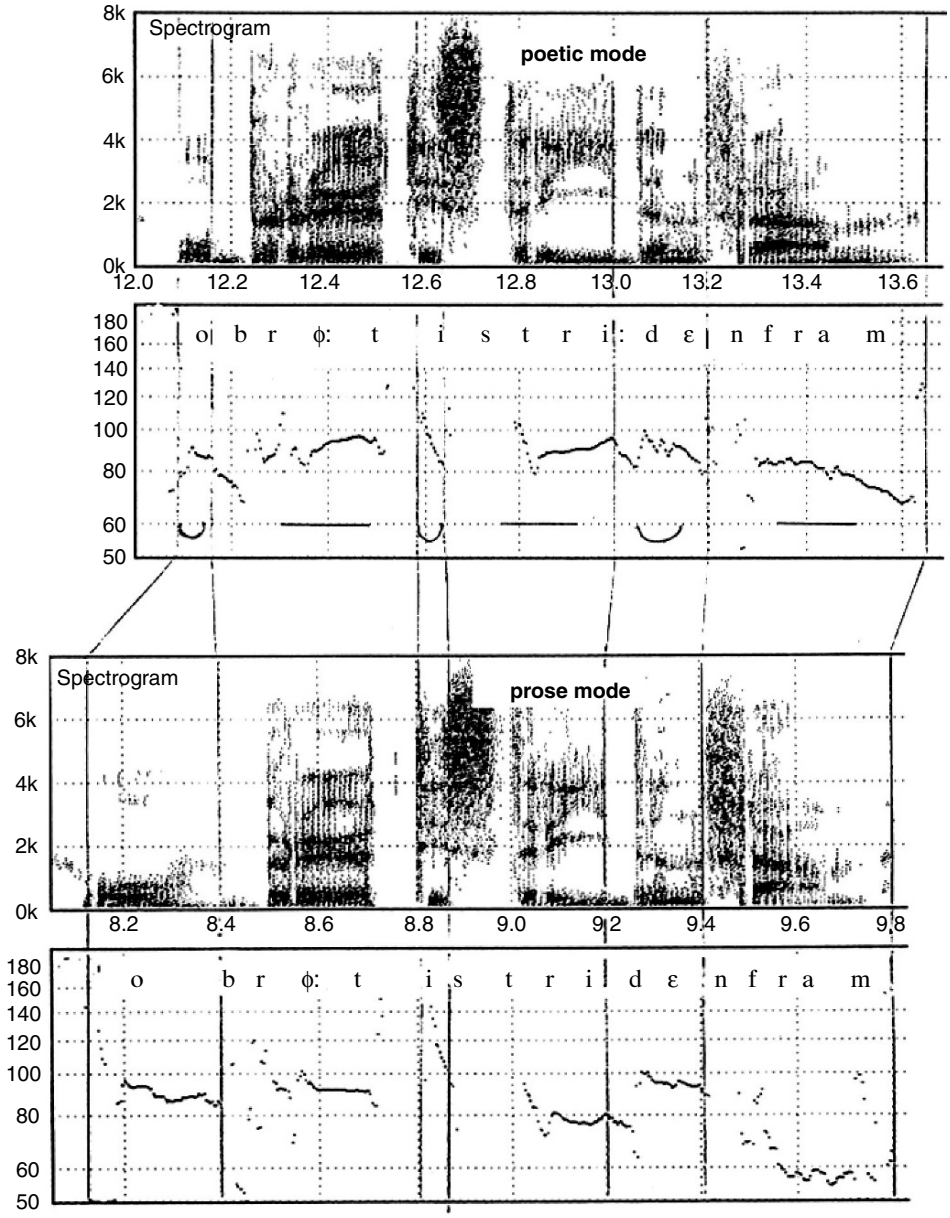


Figure 6. A complete line from the iambic poem. Poetic recital (above) and prose reading (below).

plus an increased contrast between the duration of stressed and unstressed syllables. Moreover, higher FO level, a reduced FO modulation depth of the tonal word accents, and consequently a more stable FO and intensity within a line. Furthermore, FO remains high at the end of a line except at the end of a stanza.

Absolute isochrony may occur in scanned readings only. In normal poetry reading foot durations vary somewhat less than in prose and we observe consistent patterns of declination of foot durations within a line. In poetry reading the tendency towards isochrony is reinforced by a rhythmical continuity in stress intervals that span pauses between lines, which is more pronounced than what is found in prose reading.

Iamb and trochee show meter specific characteristics in duration, FO and intensity. Some of these become apparent when comparing prose reading and poetry recitals of the same poem. The relatively heavy, unstressed syllables of the trochee at the end of a foot are associated with relatively greater duration and higher FO and intensity than in the prose version of the same poem, whereas the weak syllable of the iamb has lower FO, and the strong syllable relatively higher FO in the iambic recitation than in its prose version. We can also relate these features to the greater distinctiveness encountered in poetry recital than in prose reading. When emphasized, the unstressed syllable of the trochee is marked by a relative increase of both duration, FO and intensity, whereas in the iamb it is the stressed syllable that attains a corresponding increase.

These findings support a relational view of meter characteristics, e.g. the higher S/W contrast in the iamb than in the trochee. But it is only by considering the iambic or the trochaic foot as a unique group and feature domain that meter specific features may be conceived and defined. In this respect we may claim that specific iambic and trochaic patterns are not "metrical myths" (Loots, 1980), but a reality as proposed by earlier investigators (Risberg 1936, Newton 1975).

Meter specific characteristics are largely conditioned by the specific selection of language material found in the text, e.g. the more complex weak syllables of the trochee, usually containing more phonemes than the weak syllables of the iamb, and often originating from content words with a pronounced secondary stress on the second syllable. "Näcken" contains relatively more content words than "Karl XII". Moreover, the relative proportion of disyllabic words is greater in "Näcken" than in "Karl XII", while monosyllabic words are more frequent in "Karl XII". A relatively rich occurrence of monosyllabic stressed content words and unstressed function words is optimal for an iambic poem.

So both the particular selection of language material in the poetic text and the general characteristics of the poetic recital mode compared to prose reading determine the meter specific acoustic-phonetic and perceptual features. These features may also be reinforced by the performer's awareness of and wish to bring out the characteristic rhythm of each meter. However, in the future we need to look into an extended corpus to verify these claims.

ACKNOWLEDGEMENTS

These studies have been supported by grants from The Swedish Council for Research in the Humanities and Social Sciences and the Bank of Sweden Tercentenary Foundation,

Anita Kruckenberg and Gunnar Fant

REFERENCES

Boström-Kruckenberg, A. (1979): Roman Jakobsons poetik. Studier i dess teori och praktik (avh) Skrifter utgivna av litteraturvetenskapliga institutionen vid Uppsala Universitet 6, Uppsala 1979.

Bruce, G. (1977): *Swedish Word Accents in a Sentence Perspective*, CWK Gleerup, Lund 1977.

Fant, G. & Kruckenberg, A. (1989): "Preliminaries to the study of Swedish prose reading and reading style", STL-QPSR 2/1989, pp. 1–83.

Fant, G., Kruckenberg, A. & Nord, L. (1990): "Acoustic correlates of rhythmical structures in text reading", in *Nordic Prosody V*, Turku, pp. 70–86.

Fant, G., Kruckenberg, A. & Nord, L. (1991A): "Stress patterns and rhythm in the reading of prose and poetry with analogies to music performance", in *Music, Language, Speech, and Brain*, Wenner-Gren International Symposium Series Vol. 59, (eds. J. Sundberg, L. Nord, R. Carlson), pp. 380–407.

Fant, G., Kruckenberg, A. & Nord, L. (1991B): "Some observations on tempo and speaking style in Swedish text reading". ESCA Workshop on "The phonetics and phonology of speaking styles", Barcelona, 30 September—2 October, 1991.

Kruckenberg, A., Fant, G. & Nord, L. (1991A): "Från prosa till poesiens rytm och meter", in (eds. E. Lilja, J. Swedenmark, K. Wählin) *Vers-mått*. Studier framlagda vid Andra Nordiska Metrikkonferensen, Uppsala, 1989, pp. 147–162.

Kruckenberg, A., Fant, G. & Nord, L. (1991B): "Rhythmical structures in poetry reading", *Proceedings of the XIIth ICPhS, Aix-en-Provence*, 1991, pp. 242–245.

Loots, M.E. (1980), *Metrical Myths. An Experimental Phonetic Investigation into The Production and Perception of Metrical Speech*. 's Gravenhage

Newton, R.P. (1975), Trochaic and Iambic. *Language and Style*, No 8, 127–156

Nord, L., Kruckenberg, A. & Fant, G. (1990): "Some aspects of rhythm in prose, poetry and music", in *Nordic Prosody V*, Turku, pp. 256–265.

Risberg, B. (1936), *Den svenska versens teori, del 2*. Norstedt & Söners Förlag, Stockholm

CHAPTER 6.3

INDIVIDUAL VARIATIONS IN PAUSING A STUDY OF READ SPEECH

ABSTRACT

Earlier studies of pauses and associated prosodic boundary marking in text reading have been extended with respect to individual variability. The main part of the study pertains to text reading from a novel. A study of news bulletins tapped from the Internet was added. The main difference found was considerably shorter pauses between complete sentences in the reading of news reports than in the novel text. The five subjects involved in the novel reading showed consistent and different individual patterns of pause duration within complete sentences but more similar pause duration between sentences. The overall data collected support our quantal theory of speech timing. Special attention has been devoted to the influence of sentence length on pauses between sentences and total pause silence within sentences. Implications of our findings in prosody rules will be discussed in a separate article.

1. INTRODUCTION

Pause duration is merely a part of a complex of physical attributes signalling prosodic boundaries. Reorganisation of temporal structure, mainly as pre-boundary lengthening is usually present. It adds to boundary perception and can function alone as a "filled pause".

Additional boundary attributes are local modifications in F0 and in the voice source sensed as creaky voice. These have been studied by Fant, Nord & Kruckenberg (1986, 1989; Fant & Kruckenberg 1989) with reference to a perceptual scale of boundary prominence.

Furthermore, major prosodic boundaries such as before a new sentence or a clause are associated with an F0 reset, which according to Fant, Kruckenberg, Gustafson, Liljencrants (2002) is linearly related to pause duration.

The present study is devoted to a more detailed study of pause duration. A major object is individual variations but we shall also attempt to summarise tendencies with respect to position and sentence length.

One part of the speech material is from five subjects, two females, AÖ and IK and three men, SH, GF and JL, who read four selected parts from a novel, each of one minute's length. Two of these parts were the same as in Fant et al, (2002). The other two were analysed by one of the authors in connection with a university examination work. She also performed a parallel study of pausing in news reading from the Swedish radio, short passages read by female subjects.

2. PROSE READING

A major observation is that of uniform patterns of pause duration between complete sentences, whilst pauses within sentences showed large individual variations.

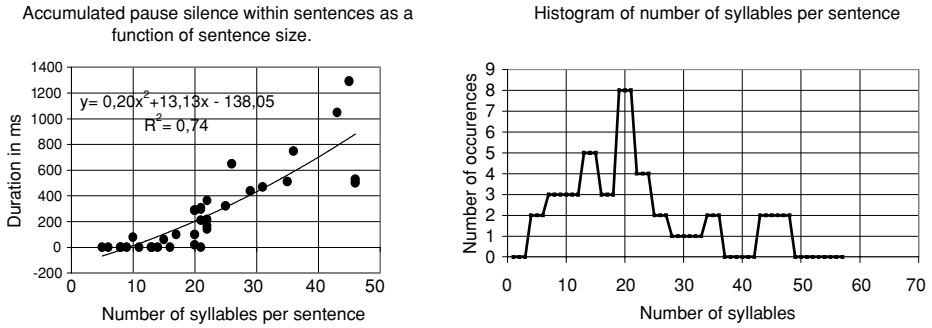


Figure 1. Accumulated pause silence within sentences as a function of sentence size and histogram of number of syllables per sentence.

There are also general trends. Sentence length has a definite influence on pauses between as well as within a sentence. Pauses between sentences averaged 1100 ms varying from 900 ms at 10 syllables length to 1300 ms at 40 syllables. Individual spreads were of the order of 150 ms. Subjects with a slow reading tempo tend to have longer than average sentence pauses, but the opposite was also found.

In contrast, as shown in the left part of Figure 1, the sum of accumulated pause length within a sentence as a function of sentence length covers a larger range. Sentences of less than 16 syllables are generally produced without any internal pause, but there is a rapid rise from 200 ms at 20 syllables to the order of 800 ms at 40 syllables. In the range above 20 syllables one can expect one or more major clause boundaries.

It is of some interest to note the many points around 20 syllables sentence length which has motivated the histogram at the right of Figure 1. The outstanding peak at 20 syllables suggests that related studies of language statistics could be of interest. Is this distribution specific to the author, Kerstin Ekman?

A sentence of 20 syllables is long enough to qualify for an internal pause. Since we are dealing with average data some speakers do not pause and others produce larger than average pauses.

Average data of total pause duration within sentences is shown to the left in Figure 2. Individual variations are considerable, the largest value for GF and the smallest for JL. AÖ, IK and SH occupy an intermediate range. In contrast, as shown to the right of Figure 2, duration between sentences varies much less. With the exception of subject JL, whose pausing pattern is rather special, there is a slight compensatory trend of larger within-sentence data combining with relative shorter between-sentence data.

A more detailed insight in individual pausing patterns is provided by Figure 3 which shows histograms for each of the five subjects of all recorded pauses within sentences, and at the lower right their sum. The individual variations are considerable and greater than expected. There is a common trend of a prominent region around 400–500 ms devoted to clause boundaries and one or two regions around 100–200 ms devoted to boundaries of lower syntactical levels. A marked exception is subject JL,

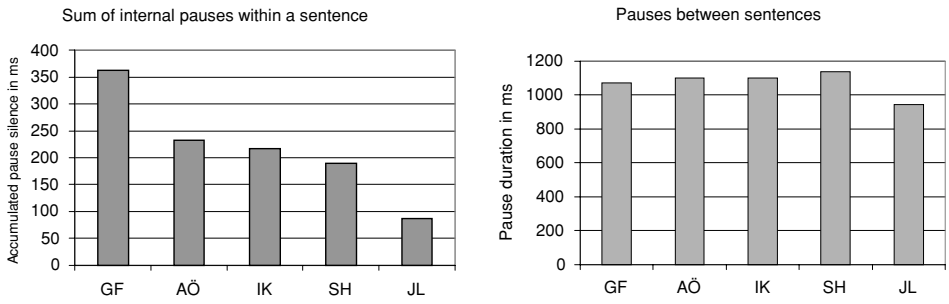


Figure 2. Average values of sentence internal total pause duration and pauses between sentences. Five subjects.

whose histogram is dominated by a single peak at 200 ms. A fairly prominent peak at 200 ms is also found in the SH histogram. The data averaged over all speakers shows three main peaks, at 450, 200 and 100 ms.

3. THE NEWS MATERIAL

The reading style in presenting news material over the radio differs considerably from that of our standard material of novel reading. Pauses between sentences are much

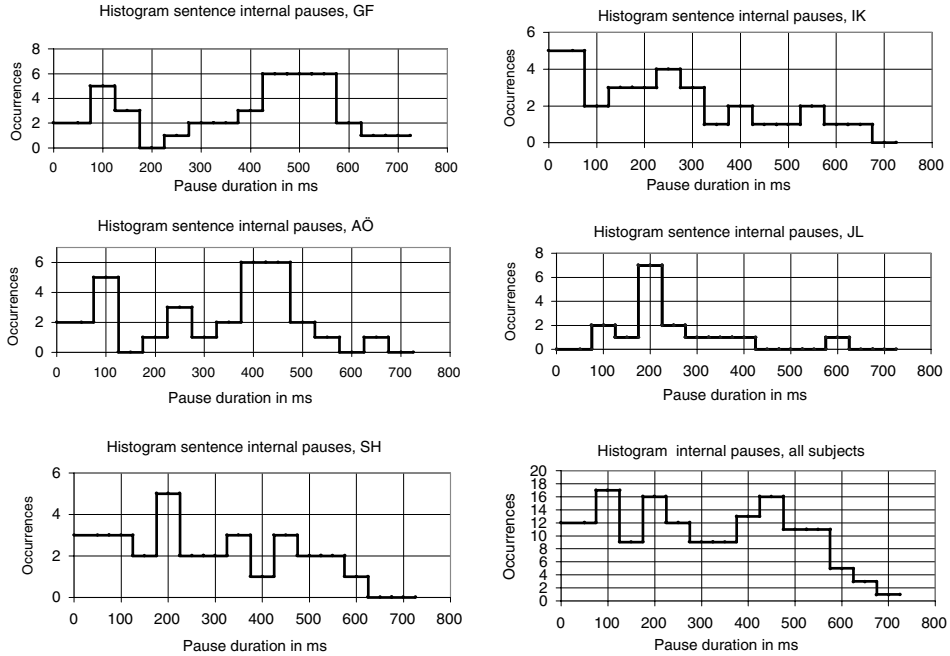


Figure 3. Histograms of sentence internal pause duration for each of the five subjects and their sum

smaller, averaging 530 ms, which appears as a dominating peak in a distribution of all pauses. Accordingly pauses occupy only 10% of the total reading time compared to 25% in the novel reading.

An attempt was made to identify major and subordinate clauses in the news material. These came out as 350 and 280 ms respectively.

4. SUMMARY OF PAUSE DATA

We have been concerned with two quite different speaking styles. One is in the reading of a novel and the other one is concerned with news reports over the radio. The main difference is in pauses between sentences which averaged 1100 ms for novels and 530 ms for news items. Paragraph pauses in the novel reading averaged 1550 ms. Sentence internal pauses were somewhat shorter in the news material than in the prose reading, 360 ms versus 450 ms for major clauses.

These patterns support our quantal theory of speech timing (Fant & Kruckenberg, 1996) according to which major pauses, with pre-pause lengthening added, favour discrete integers of a basic time constant of the order of 500 ms, usually 550 ms. A prototype pattern is thus pauses of 500, 1000, 1500, 2000 ms. A tendency in this direction is found with trained and rhythmically conscious speakers but it also executes some influence on average pause statistics.

Strangert (1990), in a study of the reading of a news related text, found pauses between paragraphs ranging from 1100 to1850 ms, pauses between sentences from 660 to 1000 ms and pauses at clause boundaries of the order of 250 ms and 130 ms at phrase boundaries. These data lie between our prose and news readings which could be expected from the specific text material.

5. EXPERIENCE FROM SPEECH SYNTHESIS

Special attention has been layed on how the syntactic frame conditions prosodic grouping and pause duration. The primary object is to predict junctures. A secondary object is to predict duration of pause silence and prepause lengthening. In view of the very large individual variations observed in our analysis data. the prediction is indeed a statistical process. It is necessary to select a prototype performance and choose a reasonable small inventory of possible pause durations matched to a limited set of positions within a syntactic frame. We have choosen 1500 ms for paragraph boundaries, 1000 ms for sentence boundaries, and the following discrete values for pauses within a sentence, 450, 175, 70, 40 and 0 ms. The choice of one of these depends not only on the syntactic frame but also on the length of the previous and the following prosodic group. Accordingly a clause boundary is not always 450 ms but can be shorter. Special rules apply to pre-pause lengthening. “Filled” pauses employ final lengthening only. Short pauses may be substituted by final lengthening.

REFERENCES

Fant, G., Nord, L.& Kruckenberg, A. (1986) Individual variations in text reading. A data-bank pilot Study. *STL-QPSR* 4/1986, 1–17.

Fant, G. & Kruckenberg, A. (1989) Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR* 2/1989, 1–83.

Fant, G., Kruckenberg, A. & Nord, L. (1989) Acoustic correlates of rhythmical structures in text reading. *Nordic Prosody V*, 23–25 Åbo.

Fant G. & Kruckenberg A. (1996) On the quantal nature of speech timing. *Proceedings of the International Conference on Spoken Language Processning*, 1996, 2044–2047.

Fant, G., Kruckenberg, A., Gustafson K, & Liljencrants, J. (2002) A new approach to intonation analysis and synthesis of Swedish, *Speech Prosody 2002, Aix en Provence.* 283–286. Also in *Fonetik 2002, TMH-QPSR* 2002, 161–164.

Strangert, E. (1990) Perceived pauses, silent intervals and syntactic boundaries, *Reports from the department of phonetics, University of Umeå, Phonum* 1, 35–38.

CHAPTER 6.4

AN INTEGRATED VIEW OF SWEDISH PROSODY. VOICE PRODUCTION, PERCEPTION AND SYNTHESIS

ABSTRACT

This is an account of our joint work on Swedish prosody and the development of prosody rules for text-to-speech synthesis. It includes a survey of voice production and multi-parameter analysis of parameter co-variation and perceptual salience. Attention is given to the auditory integration of local details of F0 contours. Accent 1 and accent 2 modulations are calculated from position and relative prominence, and added to syntactically derived intonation modules. The significance of our model to language universal rules is discussed.

1. INTRODUCTION

The purpose of our presentation is to provide a wide overview of our work in speech prosody, leading up to a novel system for intonation analysis and the prosodic base of text-to-speech synthesis.

On the production level, prosody has its roots in the human voice source and its dependence on the respiratory system and laryngeal articulation. A considerable amount of work has been devoted to voice source studies (Fant, 1982, 1993, 1995, 1997; Fant, Liljencrants and Lin, 1985) and the role of the subglottal pressure. We have outlined basic rules for the co-variation of subglottal pressure, voice intensity, duration and fundamental frequency F0 (Fant, Kruckenberg and Liljencrants, 2000A,B; Fant, Kruckenberg, Liljencrants and Hertegård, 2000). To what extent are these relations governed by physical constraints, and to what extent are they influenced by prosodic parameters? In what respects are prosodic parameters related to voice source characteristics?

As a contribution to the perceptual level of analysis, we have introduced a continuously scaled prominence parameter, labelled RS, which is applied to both syllables and words (Fant and Kruckenberg 1989, 1994, 2000B; Fant, Kruckenberg and Liljencrants 1999, 2000A,B). In speech analysis, RS determinations from listening tests, as well as subglottal pressure traces have been added to graphical displays of oscillogram, spectrogram, F0 and intensity curves. This is a maximally complete form of representation, see the detailed documentation in Fant, Kruckenberg, Liljencrants and Hertegård (2000).

From a production point of view, there is no clear separation between prosodic and segmental features. Prosodic phonological categories have their roots in all aspects of the speech wave.

An overall ambition of our work has been to present detailed accounts for the realization of prosodic categories. An example is relating prominence to duration, F0, subglottal pressure and two intensity measures, one with high frequency

pre-emphasis (Fant, Kruckenberg and Liljencrants, 2000AB; Fant, Hertegård, Kruckenberg and Liljencrants, 1997).

The final part of our presentation is devoted to intonation analysis and modelling, and to the realization of our findings in text-to-speech rules.

Of fundamental importance has been the choice of a semitone, i.e. a logarithmic scale for F0 representations. A novelty, made possible by the semitone scale, is the normalization of intonation contours in both frequency and time, which allows for a calculation of representative average values within a group of mixed male and female speakers as a foundation for speech synthesis (Fant and Kruckenberg, 2000A,B ; Fant, Kruckenberg, Gustafson and Liljencrants, 2002). Deviations from a mean reflect speaker specific intonation patterns.

Our overall strategy for synthesis has the aim of including all knowledge gained from our prosody studies. It appears to be more advanced and more detailed than in earlier proposals for Swedish, Carlson and Granström (1973) and later in Bruce et al. (2000).

It is a superposition system, in which local accent modulations in F0 are superimposed on smooth prosodic modules, i.e. base contours, applied to phrases or whole sentences. This property is shared with the influential model of Fujisaki (Fujisaki, Ljungqvist and Murata, 1993) with applications to Swedish (Fujisaki et al., 2000). Our system has the advantage of a closer tie to speech analysis data. The Swedish tonal accent 1 and accent 2 F0 modulations are given shapes and magnitudes that depend on the predicted prominence, as well as on the location of the accent contour within a prosodic group. Moreover, this dependency also applies to the calculation of segmental duration. Equations for quantitative shaping of accent, as well as of base contours, are derived by non-linear regression analysis supported by analysis-by-synthesis from our database of prose reading.

Special attention has been devoted to the prediction of prosodic grouping within a complete sentence from the grammatical structure. This is a weak point in rules for natural prosody. Here also, we have benefited from experience of the performance of our five speakers. We encountered a fair degree of conformity in gross features, such as overall declination and pausing before a new sentence. However, the lower the level of a syntactic parsing, the greater was the observed individual spread with respect to location of junctures, and of the duration and manifestation of pauses. We have ended up with a statistical view of probabilities that operate on sequences of syntactic events and their size.

To sum up, our approach incorporates several unique features, and appears to be more ambitious than other systems that we know of. The realization in an Mbrola diphone platform, which we refer to as the FK text-to-speech system, has provided a superior prosody.

The main architecture could be applicable to other languages. In addition to language specific rules, there appears to exist a base of language universal features. To sort out these relations is a challenge for future research. On the basis of our preliminary versions of English and French text-to-speech rules we have been able to demonstrate effects of code switching, e.g. simulating a French subject speaking English, retaining his French sound inventory and prosody.

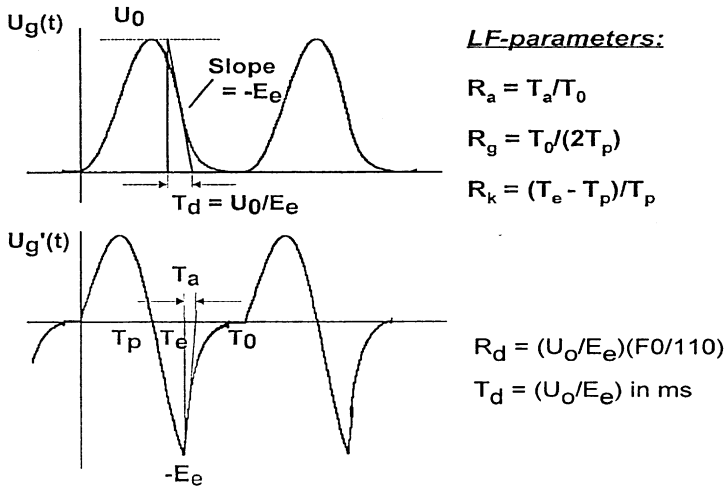


Figure 1. The LF-model.

2. Studies of Voice Production

It is the purpose of the following section to introduce subglottal pressure as a parameter in models of voice production, in specific the dependency of speech intensity on subglottal pressure and F0.

2.1. *The LF Model of Glottal Flow*

We shall start out with a brief introduction to voice production theory. The voice source is defined by the pulsating flow of air passing through the glottal slit. The LF-model introduced by Fant. Liljencrants and Lin (1985), has been widely used in speech synthesis and for general descriptive purposes. More recent developments are reported in Fant (1993, 1995, 1997). The model specifies a glottal flow pulse and the corresponding glottal flow derivative (Figure 1). The most important parameter is Ee, defined by the amplitude of the glottal flow derivative at the point of maximum discontinuity in the falling branch of glottal flow. It shows up as a prominent negative peak in the flow derivative waveform. Ee has the function of a scale factor for all formant amplitudes and is thus closely related to the sound pressure level.

The shape of the glottal flow pulse is defined by the parameters Rk, Rg and Ra. These also determine the open quotient OQ.

$$\mathrm{OQ} = (1 + \mathrm{Rk})/2\mathrm{Rg} + \mathrm{Ra} \tag{1}$$

The parameter $\mathrm{Rk} = (\mathrm{Te} - \mathrm{Tp})/\mathrm{Tp}$, describing the asymmetry of the flow pulse, is the inverse of the speed quotient, and $\mathrm{Rg} = \mathrm{T0}/(2\mathrm{Tp})$ is an inverse measure of the pulse rise time. The parameter $\mathrm{Ra} = \mathrm{Ta}/\mathrm{T0}$, is the relative time it takes for the glottis to reach maximal degree of closure after the instant of maximal discontinuity in the closing phase. Its contribution to OQ is small except in breathy voicing.

A useful parameter closely related to OQ is

$$\mathrm{Rd} = (\mathrm{Uo}/\mathrm{Ee})(\mathrm{F0}/110) \tag{2}$$

which can be interpreted as the relative declination time of the glottal pulse. It has been extensively used in our analysis of voice types.

Rd is related to the basic LF parameters by the following approximation

$$\mathrm{Rd} = (1/0.11)(0.5 + 1.2\mathrm{Rk})(\mathrm{Rk}/4\mathrm{Rg} + \mathrm{Ra}) \tag{3}$$

The descriptive role of Rd is similar to that of OQ. A regression analysis provides the relation

$$\mathrm{OQ} = 0.36 + 0.22\mathrm{Rd} \tag{4}$$

The frequency domain properties of the model are of primary importance as perceptual determinants. A high OQ promotes a high (H1 – H2), i.e. a dominance of the voice fundamental amplitude H1 over the amplitude of the second harmonic H2 and the rest of the source spectrum. The rate of decay of source harmonics in the middle and upper parts of the source spectrum is determined by a spectral tilt parameter

$$\mathrm{FA} = 1/2\pi\mathrm{Ta} = \mathrm{F0}/2\pi\mathrm{Ra} \tag{5}$$

which is the frequency at which the glottal flow spectrum begins to fall off by an extra 6 dB per octave towards higher frequencies. A low FA, i.e. a large Ta and thus a smooth and prolonged return phase of the glottal flow pulse, also contributes to increasing H1 – H2. These are typical features of breathy and weak phonations and of female versus male voices. For more detailed presentations, see Fant (1993, 1995, 1997); Stevens and Hanson (1994); Hanson, (1997A,B.); Hanson and Chuang, (1999).

Examples of voice source spectra generated by the LF model are shown in Figure 2. H1 – H2 is found to be proportional to Rd (Fant, 1997).

$$\mathrm{H1} - \mathrm{H2} = -7.6 + 11.1\mathrm{Rd} \tag{6}$$

or

$$\mathrm{H1} - \mathrm{H2} = -6 + 0.27\exp(5.5 * \mathrm{OQ}) \tag{7}$$

It should be observed that the data of Figure 2 pertain to glottal flow derivative spectra, which is a representative base for frequency domain specifications. Corresponding glottal flow spectra differ by an additional minus 6 dB/octave slope.

An increase of voice effort is usually accompanied by a decrease of OQ, Rd and H1–H2. In addition, the spectral slope parameter FA increases contributing to a relative dominance of upper formants, while on the other hand a decrease of voice effort may shift the spectral balance to a dominance of the voice fundamental. Under the latter condition, the overall sound pressure level will be less dependent on the formants than on the voice fundamental, and thus less dependent on Ee. This is a point to consider in studies of the covariation of Psub, F0 and SPL.

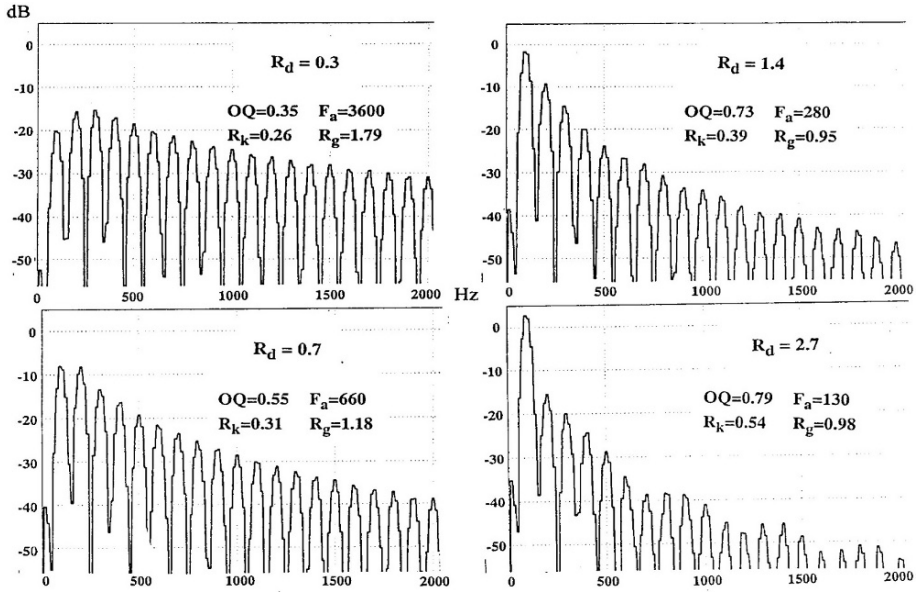


Figure 2. Representative samples of glottal flow derivative spectra generated from the LF-model.

An important fact, often overlooked, is that a manipulation of the voice fundamental amplitude H1 alone retaining H2 and all higher harmonics intact will not change the perceived quality much. What actually happens in a transition towards a breathy interval of vocal cord abduction in connected speech is that H1 stays rather constant whilst H2 and higher harmonics decrease in relative level. Additional negative spectral tilt is introduced by a decreasing FA and there is a broadening of formant bandwidths, which will be exemplified in Figure 15. This is but one example of voice source contextual variations found in connected speech (Fant, 1993, 1995, 1997; Fant and Kruckenberg, 1996).

2.2. *Static Aspects. Glissando Phonation*

As a general background we shall review an earlier study of glottal flow (Fant, 1982: Fant and Ananthapadmanabha, 1982) derived from inverse filtering of a vowel [ae] sustained at a gliding pitch from F0 = 80 Hz to 250 Hz, in Figure 3 which pertains to a subject JS.

An apparent feature is that the glottal flow amplitude Uo has a maximum in the subject's mid F0 range. The Ee parameter was extracted and is portrayed as a function of F0 in Figure 4 together with Uo. Data for a second subject, GF, is included. Both showed a distinct rise of Ee from 80 to 120 Hz of about 7 dB, which is of the order of 12 dB/oct. Above F0 = 120 Hz subject JS maintains an approximately constant Ee level, while subject GF shows a distinct drop. The peak volume velocity Uo shows a peak at F0 = 115 Hz. Its rising branch is less steep than that of Ee. Similar trends

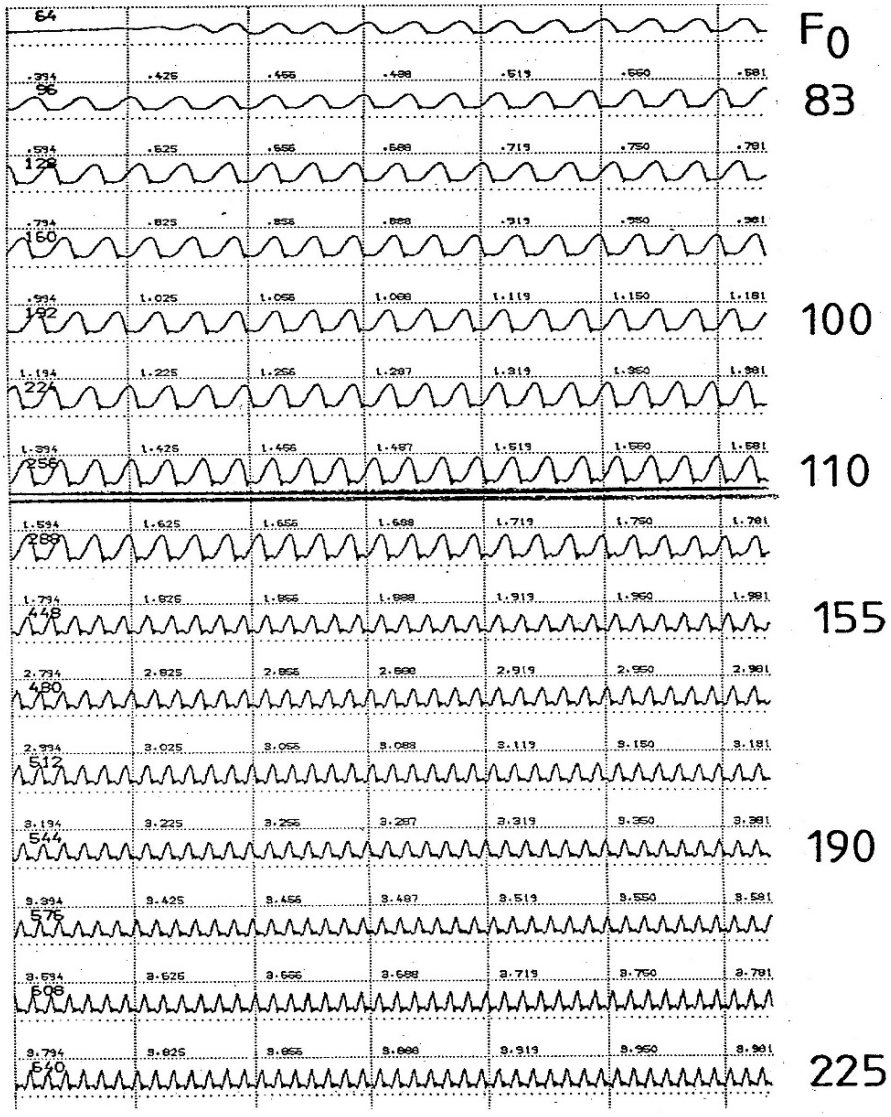


Figure 3. Glottal flow from inverse filtering of a glissando phonation.

can be observed in the lower part of Figure 4, which pertains to Ee and Uo as a function of F0 sampled from prose reading, subject ÅJ.

Data on Ee (dB) and Psub (cm H_2O) from four glissando phonations produced in the present study by subject SH are shown in Figure 5. These differ in voice effort but were produced with an intent to maintain a speaking rather than a singing mode of phonation. Ee and to a less extent Psub increase up to a reference mid-frequency at about 130 Hz, which we label F0r. Above F0r Ee tends to stay constant or, depending

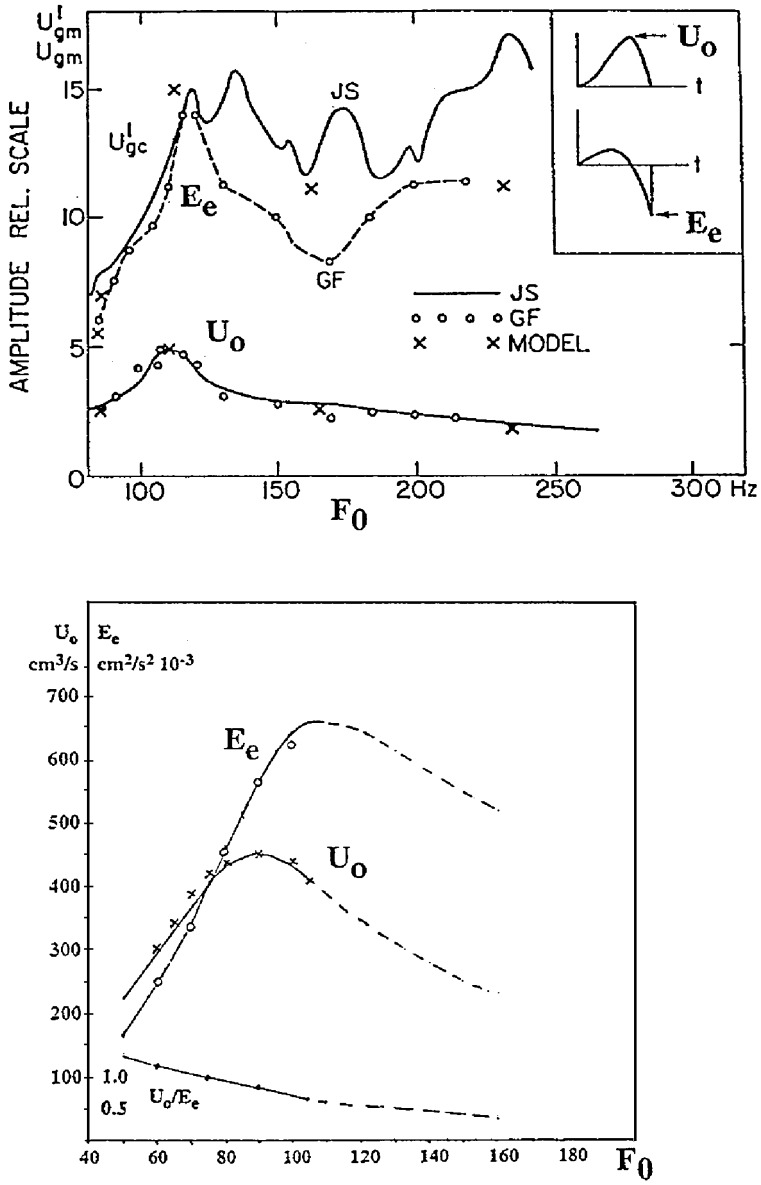


Figure 4. LF-parameters Ee and Uo as a function of F0. Above, subjects JS and GF in glissando phonation. Below, subject AJ data sampled from prose reading.

on the voice effort, shows a moderate rise or a fall. Psub exhibits similar but less distinct trends.

Averages for the four phonations appear in Figure 6 together with a prediction which was modelled from a statistical analysis of the covariation of Ee and Psub

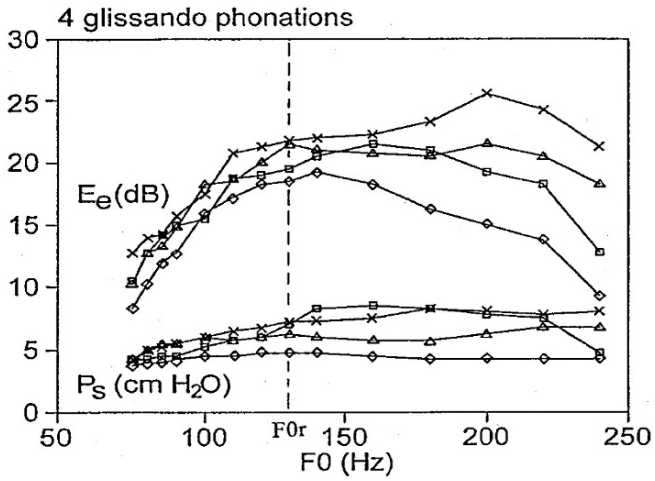


Figure 5. Ee and subglottal pressure Psub as a function of F0 in four glissando phonations.

with F0 in the range F0 < F0r, where we derived the following relations.

$$\text{Ee} \sim \text{F0}^{1.35} \quad (8)$$

(at constant Psub)

$$\text{Ee} \sim \text{Psub}^{1.1} \quad (9)$$

(at constant F0)

$$\text{Psub} \sim \text{F0}^{0.7} \quad (10)$$

(at constant Ee)

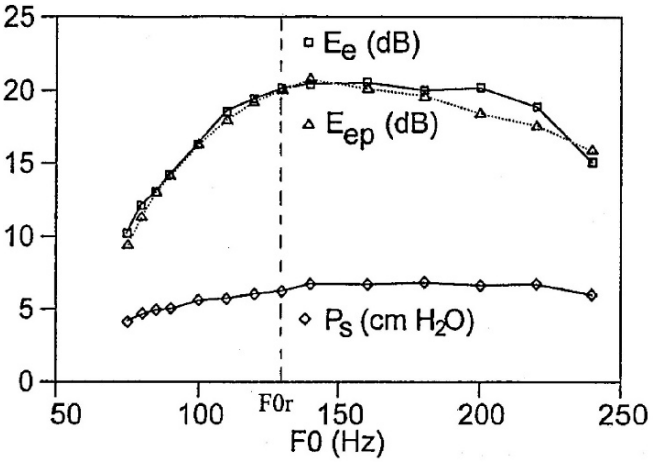


Figure 6. Average contours of Ee and Psub as a function of F0 and Eep, a prediction from F0 and Psub.

Accordingly, with co-varying Psub:

$$\text{Ee} \sim \text{F0}^{1.35*}\text{Psub}^{1.1} \sim \text{F0}^{2.1} \tag{11}$$

i.e. Ee is increased by 12.5 dB/oct increase in F0 when the covarying Psub is taken into account. On the other hand, in terms of Psub and a normal co-varying F0, we find

$$\text{Ee} \sim \text{Psub}^3 \tag{12}$$

The prediction of Ee being proportional to $\text{F0}^{1.35}*\text{Psub}^{1.1}$ provides a good fit up to F0r = 130 Hz which stands out as a distinct breaking point.

A more general model of Ee as a function of Psub and F0, valid for the entire F0 range has been developed

$$\text{Ee} = \text{K} + 20\log_{10}\left\{\text{P}_{\text{tr}}^{1.1}\text{x}_{\text{n}}^{1.35}[(1-\text{x}_{\text{n}}^2)^2 + \text{x}_{\text{n}}^2/\text{Q}^2]^{-0.5}\right\} \quad \text{dB} \tag{13}$$

where $\text{x}_{\text{n}} = \text{F0}/\text{F0r}$ and Q = 1.25. Here the transglottal pressure drop, Ptr = Psub-Psup, substitutes Psub, which is aerodynamically motivated under conditions of a finite supraglottal pressure drop, Psup. As shown in Figure 6, the overall fit to the measured data is good.

Assuming a fixed formant pattern or a dominance of the first formant, the sound pressure level SPL is proportional to Ee. The effect of a varying F0 can be approximated by reference to the density of pitch pulses exciting the vocal tract. An increase of F0 by ΔF0 would accordingly increase SPL by

$$\Delta\text{SPL} = 10\log_{10}(1 + \Delta\text{F0}/\text{F0}) \quad \text{dB} \tag{14}$$

i.e. 3 dB per octave increase of F0. This is but a minor part of the overall increase with F0. Some deviations from this rule may occur at high F0 values, depending on the location of F_1 with respect to the closest harmonics. We end up with the following expression for non-close vowels:

$$\text{SPL} = \text{K} + 20\log_{10}\left\{\text{Ptr}^{1.1}\text{x}_{\text{n}}^{1.85}[(1-\text{x}_{\text{n}}^2)^2 + \text{x}_{\text{n}}^2/\text{Q}^2]^{-0.5}\right\} \quad \text{dB} \tag{15}$$

which has been successfully tested in isolated spoken vowels, and also in connected speech as will be shown in Figure 13. However, according to more recent simulations, the accuracy can be somewhat improved by a larger exponent for Ptr. A factor $\text{Ptr}^{1.6}$ is recommended.

2.3. *The Significance of F0r*

The significance of F0r as the boundary between an upper and a lower part of a speaker's available pitch range is further enhanced by the graph of peak glottal area Ag, Figure 7, determined from fiber scope recordings in one of the glissando phonations. An average, analytically determined contour has been fitted to the data points of Figure 7, assuming an F0r = 130 Hz, which provides a reasonable fit.

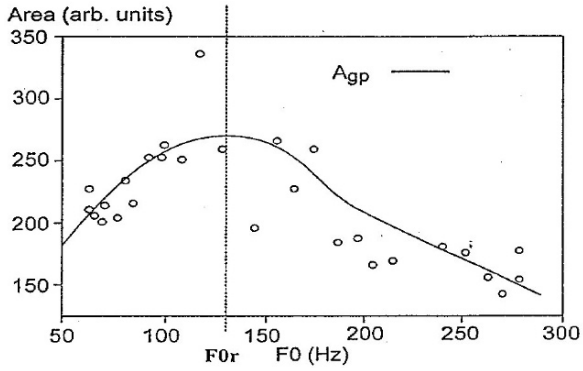


Figure 7. Glottal peak area as a function of F0 from one of the glissando phonations. High speed fiberscope photography.

We find the following approximate relations:

$$\text{Ag} \sim \text{F0}^{0.5} \text{ and } \text{Ug} \sim \text{F0}. \quad (16)$$
$$(\text{F0} < \text{F0r})$$
$$\text{Ag} \sim \text{F0}^{-1} \text{ and } \text{Ug} \sim \text{F0}^{-1} \quad (17)$$
$$(\text{F0} > \text{F0r})$$

In the upper frequency range the average trend is a constant Psub and a constant Ee, while Ag and Ug are inversely proportional to F0.

A smooth connection between the two domains in Figure 6 has been modelled by the following expression

$$\text{Ag} = \text{A}_\text{r}^\text{y} \quad (18a)$$

The exponent y is expressed as a function of $\text{x}_\text{m} = \text{F0}/\text{F0r}$

$$\text{y} = 0.5(1 - \text{x}_\text{m}^2) \quad (18b)$$
$$(\text{F0} < \text{F0r})$$

and

$$\text{y} = (\text{x}_\text{m}^{-2} - 1) \quad (18c)$$
$$(\text{F0} > \text{F0r})$$

We also have data on a corresponding normalized glottal width, expressed as the total area divided by the length of the glottal slit. As shown in Figure 8, it has a maximum at a somewhat lower frequency than F0r.

The concept of F0r outlined above deserves to be considered in phonatory theory. It is not claimed to mark a register break, but rather a region of changing parameter relations, and in practice it is a suitable mid-frequency reference for a voice.

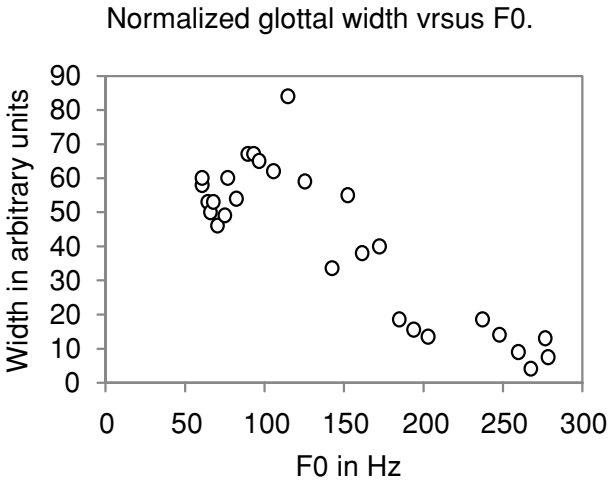


Figure 8. Normalized glottal width measured in connection with Figure 7.

Our measure F0r = 130 Hz is typical of a fairly low-pitched male. For a moderately high pitched female, to be portrayed in Figure 14, we have found an F0r around 220 Hz. A well developed voice has access to one octave below and one octave above F0r. The lower part is extensively used in speech, whereas the upper part is mainly reserved for intonation gestures and high prominence accentuation.

2.4. *Comments on Parameter Interrelations*

Our findings of relations between Psub, F0 and SPL from glissando phonations deserve some comments. How do they compare with data from sustained phonations in the literature and how do they relate to the production mechanism?

Although there are trends of a positive correlation, F0 is basically independent of Psub and mainly determined by the crico-thyroid muscle. It is known from earlier studies that a perturbation of Psub at constant laryngeal muscle activation causes a passive increase of F0. According to modelling performed by Titze (1989), this effect is largely confined to low F0 phonations, where the vocal folds are slack and lack stretching. Here, the F0 increase is of the order of 4 Hz per cm H_2O in Psub. Titze (1989) explains the F0 rise by an increase of the width of the glottal slit at constant length causing an elongation and stretching of the edge contour. However, this passively induced F0 rise is much smaller than observed from the average covariation of F0 and Psub in connected speech.

When calculating the gain in SPL from an increase of Psub, it is necessary to make specific assumptions concerning an associated F0 shift. Considering only F0 values below F0r, a differential version of Eq. 15 is

$$\begin{aligned}\Delta\mathrm{SPL} = {} & 20\log_{10}[(\mathrm{Psub}+\Delta\mathrm{Psub})/\mathrm{Psub}]^{1.1} \\ & +20\log_{10}[(\mathrm{F0}+\Delta\mathrm{F0})/\mathrm{F0}]^{1.35+0.5} \quad \mathrm{dB}\end{aligned} \tag{19}$$

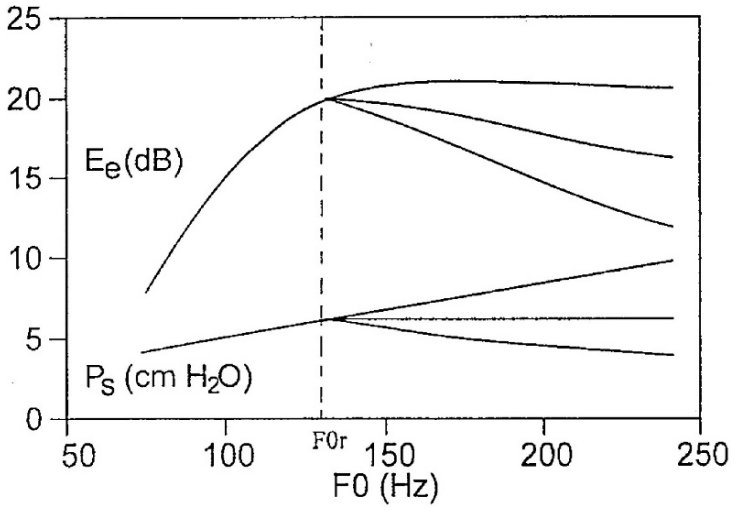


Figure 9. Three stylized alternatives of varying Ee and Psub. At frequencies above F0r, Psub is assumed to continue to rise, or be constant, or decaying. The decaying contour is often found in connected speech, the rising is typical of singing.

Thus, the gain in SPL by a doubling of Psub from 4 to 8 cm H_2O, assuming a passively induced F0 increase of 4 Hz per cm H_2O above F0 = 100Hz, provides a $6.6 + 2.4 =$ 9 dB increase which is in general agreement with several earlier studies, e.g. the theoretical derivation from pulse shape considerations of Fant (1982), and the experimental data of Ladefoged (1967); Sundberg et al. (1999); Strik and Boves, (1992). An uncertainty lies in the true amount of F0 shift in these experiments. Ladefoged's data is from utterances of single words at different voice efforts, with no restrictions laid on the pitch. He cites sound pressure being proportional to $Psub^{1.6}$ which is 9.5 dB per doubling of Psub. Sundberg et al (1999), claiming a constant pitch in their experiments from sustained phonations in a singing mode, also report a $Psub^{1.6}$ proportionality. The particular exponent may depend on the particular F0 level.

In a singing mode, Psub and Ee usually continue to increase with F0 above F0r. In a weaker voice effort, Ee could be falling which has implications for dynamic patterns of F0 and SPL in connected speech. These patterns are indicated schematically in Figure 9. Of special interest is the occurrence of an F0 peak in a domain of falling Psub at F0 > F0r. If the F0 peak overshoots F0r, the SPL reaches a maximum when F0 passes through F0r, and a minimum at the peak F0 value, see Figure 14.

3. SHORT UTTERANCES

3.1. *Vowels*

We shall now turn to the dynamic aspects of voice parameters. A part of our recordings were concerned with isolated vowels. The upper part of Figure 10 shows the

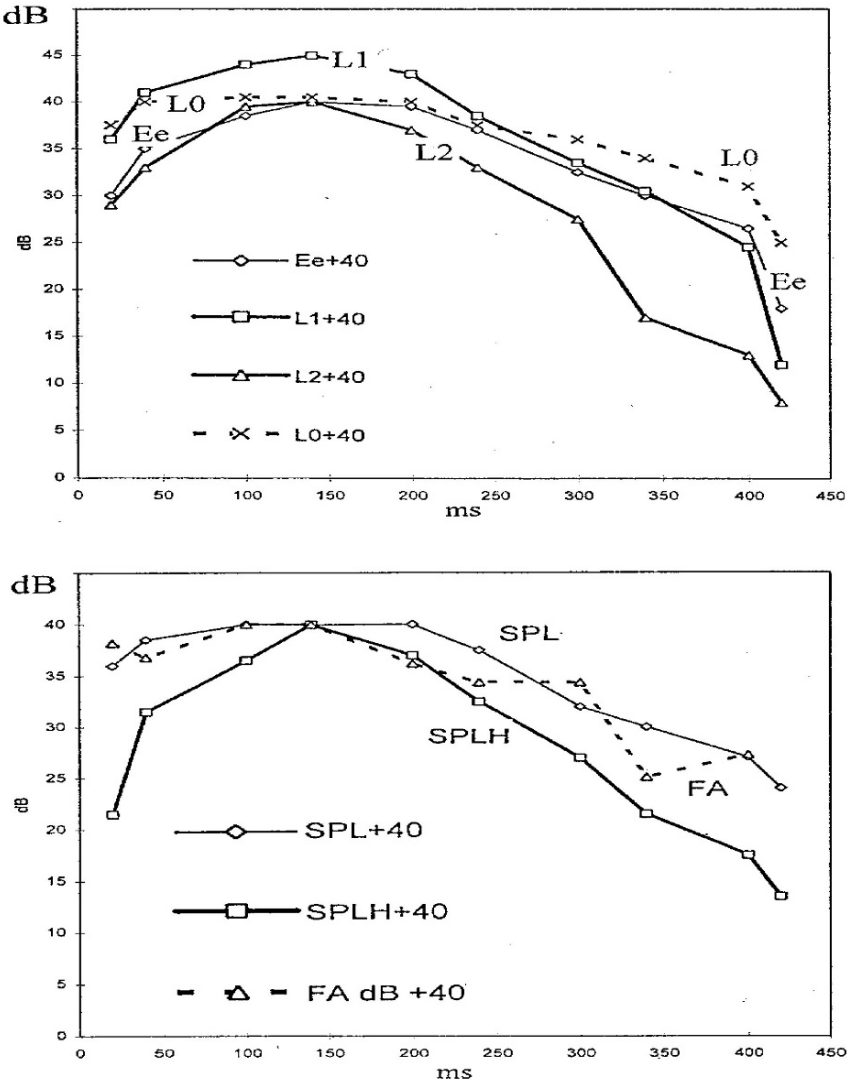


Figure 10. Temporal contours of a single vowel [ae] utterance. Above L0, L1, L2, and Ee, below SPL, SPLH and the source tilt parameter FA.

temporal varitions within a vowel [ae] of first and second formant amplitudes L1 and L2, and of the amplitude L0 of the voice fundamental, together with the voice source excitation amplitude Ee. The overall impression is a rise in the initial third part of the vowel, followed by a decay in which L0 gains dominance at the same time as L2 decays faster than L1. L0 is also prominent in the very first part of the utterance.

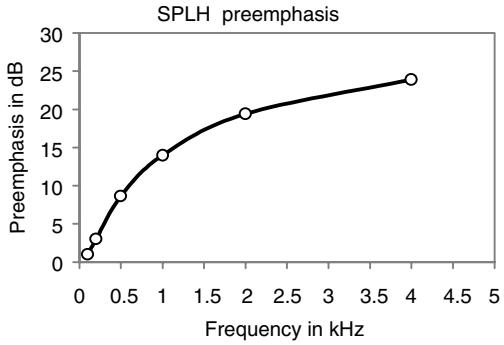


Figure 11. SPLH pre-emphasis.

The lower part shows SPL and a high frequency emphasized intensity measure SPLH which differs from the SPL by introduction of our standard pre-emphasis

$$G(f) = 10\log_{10}\{(1 + f^2/200^2)/(1 + f^2/5000^2)\} \quad \text{dB} \tag{20}$$

As illustrated in Figure 11, the pre-emphasis has a gain of 3 dB at 200 Hz, 14 dB at 1000 Hz and 25 dB at 5000 Hz. SPLH is more sensitive to variations in the region of the second and the third formant, F2 and F3, than is SPL and could accordingly match the concept of sonority.

The difference, SPLH-SPL, brings out the relative spectrum level of formants above F1 which in part is related to the source and in part to the filter function, i.e. the formant pattern (see section 5.7). At constant articulation, variations in the SPLH-SPL measure accordingly mirror variations in the high frequency contents of the source, which in turn is related to the concept of spectral tilt (Sluijter and van Heuven, 1996; Campbell and Beckman, 1997).

In Figure 10 it can be seen that SPLH is peaked towards the same position in time as L1 and L2 and falls off at a higher rate than SPL towards the end of the vowel. Included in Figure 10 is the LF spectral slope parameter FA, here transformed into an equivalent relative spectral level in dB at higher frequencies. The decay of L2 at a faster rate than Ee towards the end of the vowel, is coherent with the falling branch of FA, typical of breathy voicing. In the early part of the vowel, FA varies less and L2 follows Ee closely. The initial relative high L0 matching L1 is typical of a soft onset associated with other LF parameters than FA.

The covariation of F0, Psub, Ee (dB), SPL and a predicted Eep sampled from the average of five vowels is illustrated in Figure 12. SPL, Ee and Eep have been normalized to the same temporal position at a time $t = 0$ representing a common peak location. At $t > 0$ the three intensity parameters follow a common profile, indicating a good prediction from Eq. 15. Psub also has a maximum at $t = 0$. The F0 contour starts at F0 = 115 Hz with a small rise and extends down to F0 = 75 Hz. This profile is similar to that of a declarative phrase ending with complete vocal cord abduction.

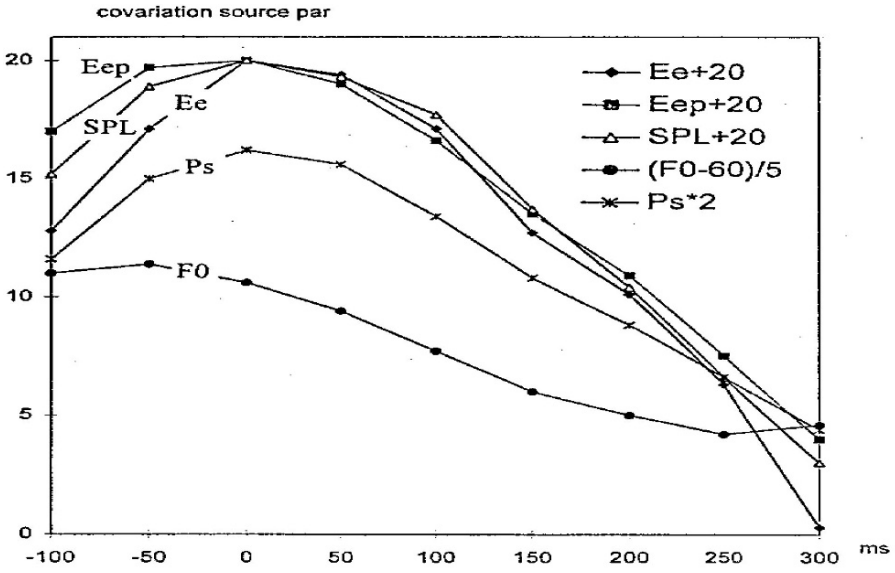


Figure 12. F0, Psub, SPL, Ee (dB), and predicted Ee, Eep, sampled from the average of five vowels.

3.2. *Isolated Sentences*

Our next object is the two word phrase, in Figure 13 "Ja, adjö" [jɑ:ajø]. It is pronounced with a juncture between the words marked by a drop in Psub, F0 and intensity. In both vowels the F0 peaks occur in a region of falling Psub, which is typical of long stressed vowels in Swedish. A prediction of SPL from F0 and Psub according to Eq. 15 has been carried out with the predicted SPLp contour vertically anchored in one SPL point. The overall fit is good, except in the juncture between the two words and in the [j] segment which have more extreme spectral patterns.

The covariation of Psub, F0 and intensity parameters illustrated in Figure 13 have a general significance. Typical of syllables carrying a long stressed vowel is a slow rise of Psub towards the onset of the accentuated syllable, followed by a fall within the syllable, combined with increasing F0. This is in particular evident in focal accentuation. An illustrative example from Fant and Kruckenberg, (1995); Fant, Kruckenberg and Liljencrants (2000B) is shown in Figure 14, pertaining to a female speaker's utterance: "Maria Le**nar** igen". Subglottal pressure was not recorded, but a falling branch of Psub can be inferred from the particular pattern of an SPL maximum as F0 passes through F0r = 220 Hz, and then a minimum of SPL at the F0 peak, and again a maximum at F0 = F0r in the descending branch. The same sentence with Psub included will appear in Figure 27B.

Figure 15 serves the dual purpose of illustrating Psub and Psup contours within a short sentence, and of voice source modifications associated with pre-occlusion aspiration in a highly stressed vowel [a]. The utterance is "E Axel här?" (Is Axel

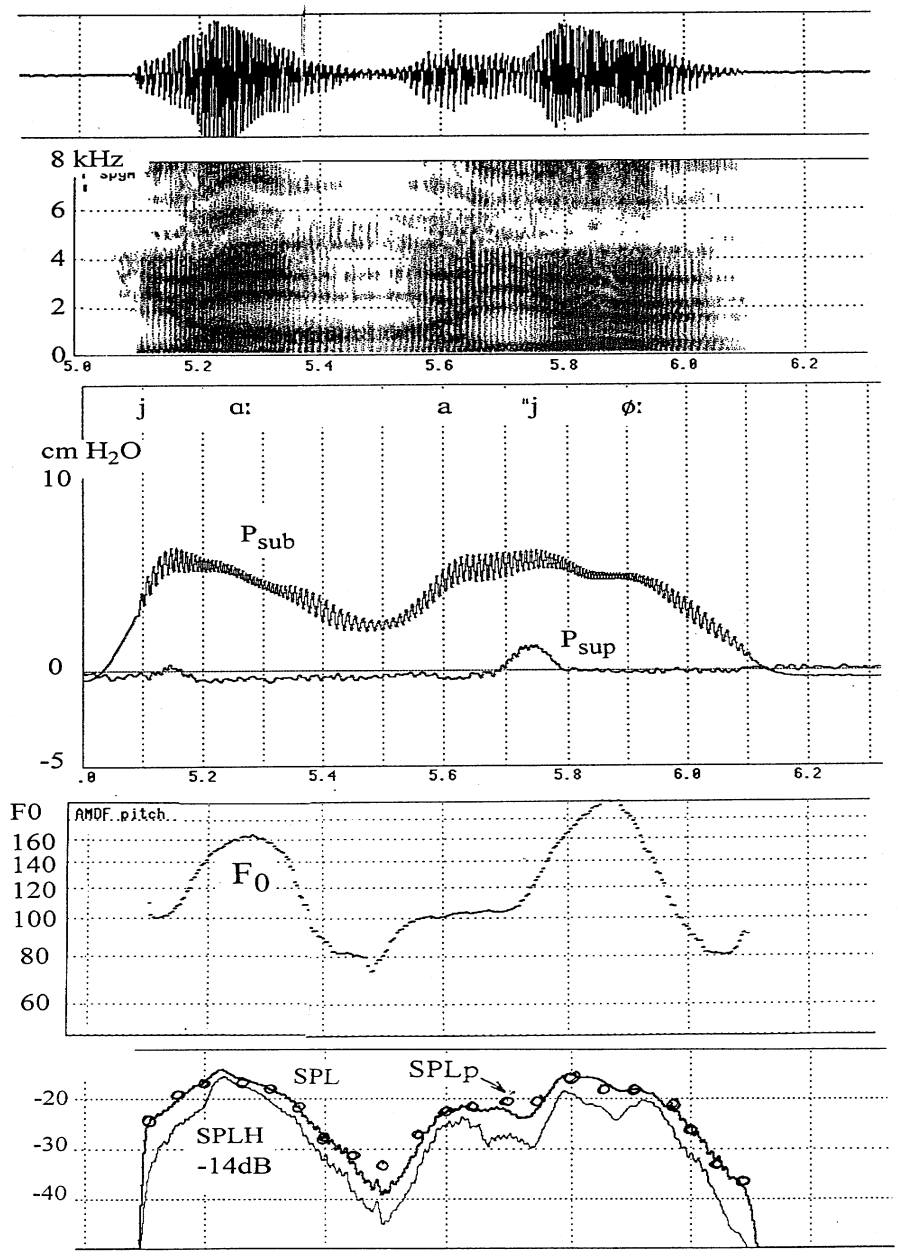


Figure 13. Illustrating the predictability of SPL in a two-word phrase from Psub and F0. Predicted SPLp is marked by open circles.

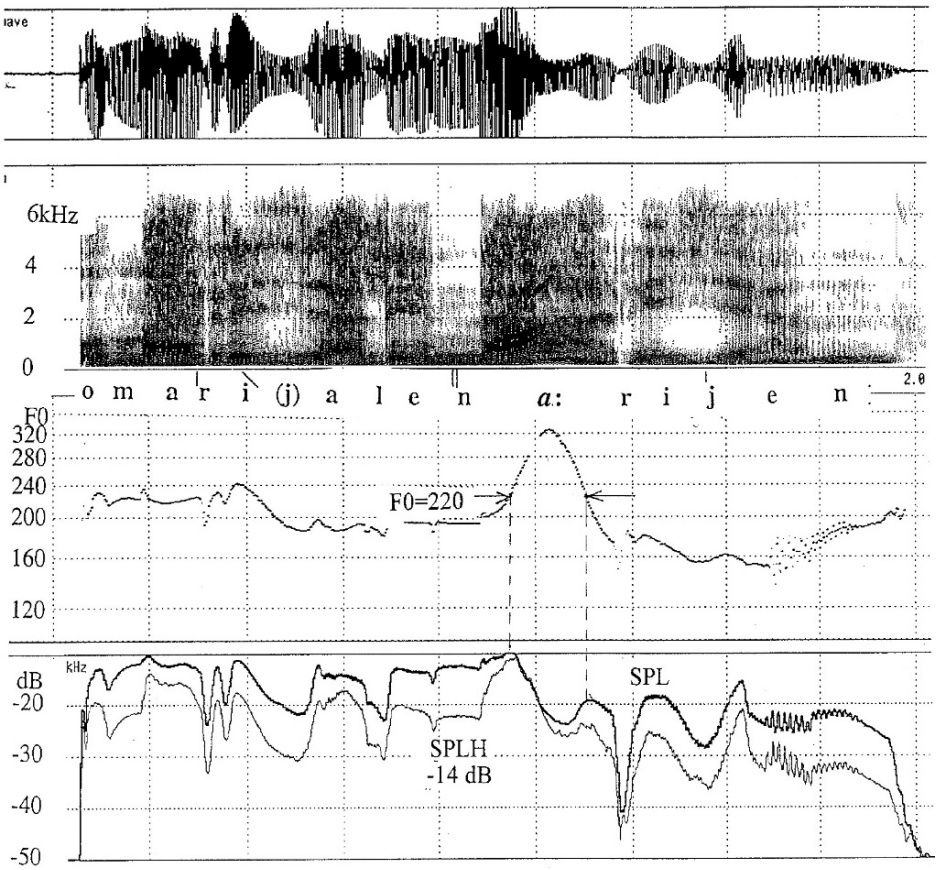


Figure 14. Illustrating the co-variation of SPL and SPLH with F0 in focal accent 2 [mariale na'r ijen], female subject. In the domain of the intonation peak, SPL passes through maxima at F0 = F0r = 220 Hz and a minimum at the F0 peak which suggests a decaying Psub, cf. Figure 9.

here?). As expected, Psub and Psup meet in the [ks] part, and there is a finite Psup build-up in the [l] and the [r]. The quite apparent dip in Psub at the end of the vowel [a], and the brief minor dip in the transition from [s] to [e] are due to rapid changes in glottal and supra-glottal openings in the boundary regions. A more detailed discussion of the aerodynamics of vowels and consonants will appear in sections 4 and 5.

Voice source dynamics in the pre-occlusion region of the vowel [a] have been visualized by narrow-band spectra sampled at 6 successive intervals 20 ms apart, from the middle of the vowel to the occlusion gap. A conjectured glottal area Ag has been drawn to illustrate the concept of glottal abduction as the cause of the increasing breathiness. The first spectrum sample displays a typical [a] pattern, but for a voice fundamental amplitude almost of the same level as that of the first formant. In the

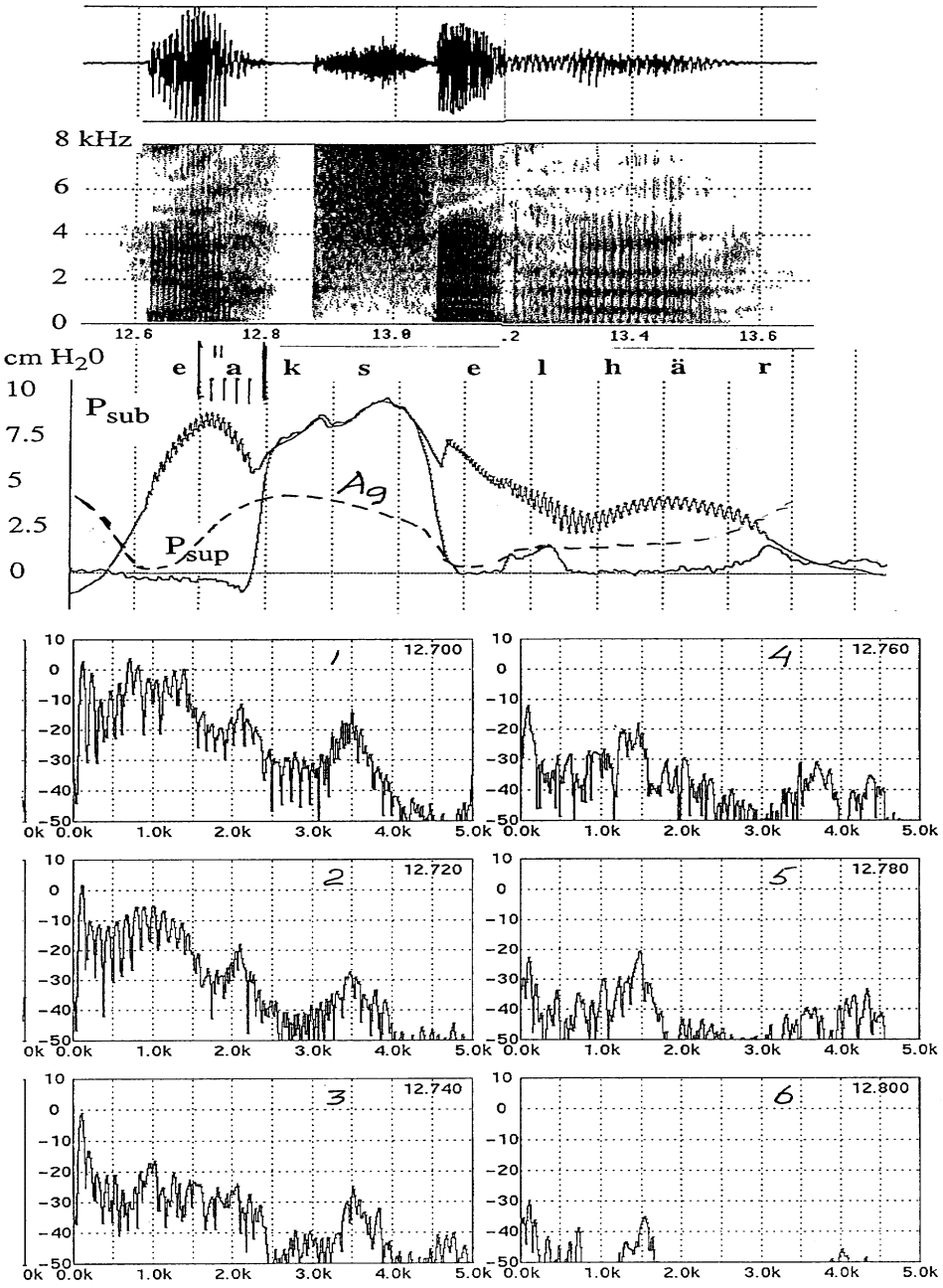


Figure 15. Pre-occlusion aspiration of a vowel [a] in the sentence "E Axel här?". [e " " aksEl hÏ†r]. Spectral sections at 20 ms intervals are shown. A conjectured Ag trace is inserted to indicate the underlying vocal fold abduction gesture.

next sample, the amplitude of the fundamental has stayed constant, while the level of the second harmonic and the rest of the spectrum has been effectively reduced. There is also a typical broadening of the bandwidths of F1 and F2 (Fant, 1997). In the third sample, there is a further reduction of the formant structure, and there is a fill in of noise and also of subglottal formants. The relatively small reduction of the amplitude of the voice fundamental from sample 1 to 3 can be expected to have a correspondence in a rather constant glottal flow amplitude Uo. In this sequence, the LF-parameter Rd and the open quotient OQ increase, Ee decays at a faster rate than Uo and FA will decrease. This pattern is also typical of breath group final abduction. Samples 4–6 are noise dominated and show a main peak at 1500 Hz, typical of the approaching [k] articulation.

3.3. *The Subglottal Pressure Contour*

The subglottal pressure contours show a high degree of regularity. At the onset of a breath group, voicing starts at a Psub of 3–5 H_2O. The duration of the rise to an initial value of the order of 6–8 cm H_2O is 120–180 ms. The main trend of the Psub contour is a declination to a value of the order of 4–5 cm H_2O at a reference point about 150 ms before the voicing offset, and then a faster rate of fall down to a final value of about 1.5 cm H_2O. Superimposed on the Psub and Psup contours are regions of maxima and minima, related to the stress pattern and to perturbations due to articulation-aerodynamic interaction, usually at voiced/voiceless boundaries. The initial voice onset requires vocal fold adduction and the normal voice offset is caused by abduction.

Observed regularities in an extended set of recordings (Fant, Kruckenberg, Liljencrants and Hertegård, 2000) deserve some comments. In a primary clause, usually a new sentence, the initial Psub has a mean value of 7.3 cm H_2O, with a standard deviation (sd) of 0.6 cm H_2O. The clause final value was 4.5 cm H_2O (sd = 0.6). For a following secondary clause within the same sentence, the corresponding values were Psub = 6.4 cm H_2O (sd = 1.0) initially and Psub = 4.0 cm H_2O (sd = 0.9) finally. A remarkable feature is that initial as well as final Psub values tended to be independent of the particular duration of a breath group, which averaged 1.9 sec (sd = 0.6) for a primary clause and 1.7 sec (sd = 0.5) for a secondary clause. As a matter of fact, regression equations relating Psub fall to duration showed a weak negative tendency. The Psub fall averaged 2.8 cm H_2O for a primary clause and 2.4 for a secondary clause.

A similar trend has been found for the F0 contours, which although much more variable show an average declination of 4 semitones in a primary clause, and 3.5 semitones in a secondary clause. In view of the predictability of SPL from Psub and F0, Eqs 15 and 19, we may estimate a corresponding declination in SPL to be 8.5 dB for a primary clause and 7.5 dB for a secondary clause. The contributions from Psub and F0 are about the same, somewhat greater for Psub. Disregarding the influence of local stress, these predicted SPL values are of the correct order of magnitude for a neutral declarative sentence.

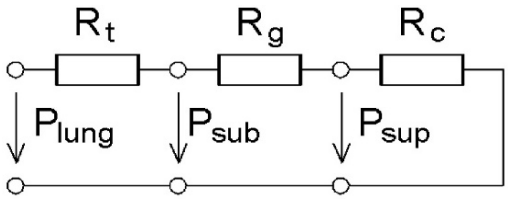


Figure 16. Pressure fall in the trachea, the glottis, and the supra-glottal pathways.

4. Aerodynamic Modelling and Data

Some of the observations above may be explained by reference to an equivalent circuit analysis of the respiratory system (Rothenberg, 1968; Stevens, 1998). A complete system tends to be complicated and there is uncertainty about details. However, as shown in Figure 16, the origin of the declination of Psub in a neutral declarative sentence can be traced to a single RC time constant, the product of the combined glottal resistance and the sub- and supra-glottal resistances, Rg+Rt+Rc, and Cℓ, the inverse of the mechanical compliance of the muscular system executing a pressure on the lungs.

We shall now predict the declination in subglottal pressure by selecting a value Cℓ = 1/17 proposed for the thorax (Liljencrants, Fant and Kruckenberg, 2000). Following Stevens (1998), the influence of the diaphragm which operates in parallel to the thorax has been neglected. The resistance of the respiratory pathways is dominated by the glottal resistance. We shall adopt a value of Rg = 46 acoustical ohms corresponding to a mean glottal area of Ag = 0.04 cm^2.

The choise of Ag deserves a comment. A peak glottal area is about three times greater than the mean value. Thus, the assumed Ag = 0.04 cm^2 would correspond to a peak value of 0.12 cm^2, which could be modelled by a glottis of elliptic shape with length 1.2 cm and a width of 1.2 mm, which seems realistic.

The mean glottal resistance Rg is derived from the aerodynamic equations

$$\Delta P = \rho v^2/2 \tag{21}$$

$$U = vA \tag{22}$$

$$R = \Delta P/U = A^{-1}(\rho \Delta P/2)^{1/2} \tag{23}$$

where the particle velocity v is in cm/s and the pressure drop ΔP in dynes/cm^2(1 cm H_2O = 980 dynes/cm^2). The flow U is in cm^3/s and the density of air ρ =1.14*10^{-3} g/cm^3.

Thus, in terms of the transglottal pressure drop, ΔP = Ptr = Psub-Psup, and assuming a mean transglottal pressure of 6 cm H_2O

$$Rg = Ag^{-1}(\rho Ptr/2)^{1/2} = 46 \text{ acoustical ohm} \tag{24}$$

A simple exponential decay with the time constant Tc = (Rg+Rt)Cℓ =2.7 sec, would model a decay of Psub from 7.5 to 4.3 cm H_2O in an interval of

T = 1.5 seconds, and 3.6 cm H_2O after 2 seconds. This is in reasonable agreement with the observed data above.

The corresponding mean rate of flow is

$$Ug = Ag(2Ptr/\rho)^{1/2} = 130 \text{ cm}^3/\text{s} \quad (25)$$

During a 2 seconds' phonation without glottal leakage the volume of the exhaled air would be 260 cm^3.

In regions of inhalations between breath groups, the transglottal pressure goes negative. As a specific example, at the onset of the sentence in Figure 20B, there is a breathing interval of 0.3 seconds effective duration with a mean value of Ptr = Psub-Psup = $-2, 5 - (-1) = -1.5$ cm H_2O. According to Eq. 25 and assuming a glottal area of 1 cm^2 the rate of inhalation would be 1600 cm^3/s, i.e. 485 cm^3 during a 0.3 second interval. It would equal the consumption of air in a preceding phonation of 3 seconds length, assuming a mean glottal opening of 0.05 cm^3 and a mean Psub of 6 cm H_2O.

The modelling above can only point at some essentials. It should be followed up by more detailed measures of actual air consumption during a breath group. The rate of air consumption in brief intervals of breathy voicing, such as at voiced/voiceless boundaries and group final vocal fold abductions, is much larger than in intervals of normal voicing. It remains to be seen to what extent the excess flow at these intervals of vocal cord abduction are compensated by intervals of complete vocal tract constriction such as in stops. Large individual and voice type variations can be anticipated.

The observed tendency of the final Psub being independent of the duration of the breathgroup can not be explained by a greater rate of air consumption in short groups. Instead, we could conceive of an underlying gesture of relaxation of the driving pressure within a breath group. Instances of a more constant Psub, or the presence of a focal rise in Psub, require a control via increasing pulmonary pressures.

The aerodynamics of overlaid perturbations in the subglottal contour due to sudden glottal and articulatory variations demand more complicated analog networks (Stevens, 1998). However, some stationary aspects of the joint effect of the glottal and the supraglottal resistances can be treated by a simplified static model (Figure 16). In particular, we can model how Psub and Psup and thus Ptr vary with different combinations of Ag and the effective area Ac of a supraglottal constriction. Calculations are carried out with respect to a constant lung pressure Plung, applied to a single branch network with a tracheal resistance Rt in series with Rg and the constriction resistance Rc. Following Rothenberg (1968) and Stevens (1998, p. 453) we assume a flow independent Rt of 1.5 acoustical ohms. With the support of Eq. 23 and the continuity of flow, vgAg = vcAc, we derive the following pressure distribution:

$$Plung = vg\ AgRt + (\rho vg/2)(1 + Ag^2/Ac^2) = Pt + Pg + Pc \quad (26)$$

The glottal particle velocity vg is solved from the known Ag and Ac and a lung pressure of Plung = 6 cm H_2O. This value represents breathgroup medial positions along the declination contour, and is somewhat lower than our average group initial value of 7.3 cm H_2O. It can be noted that the reference value for calculations of glottal

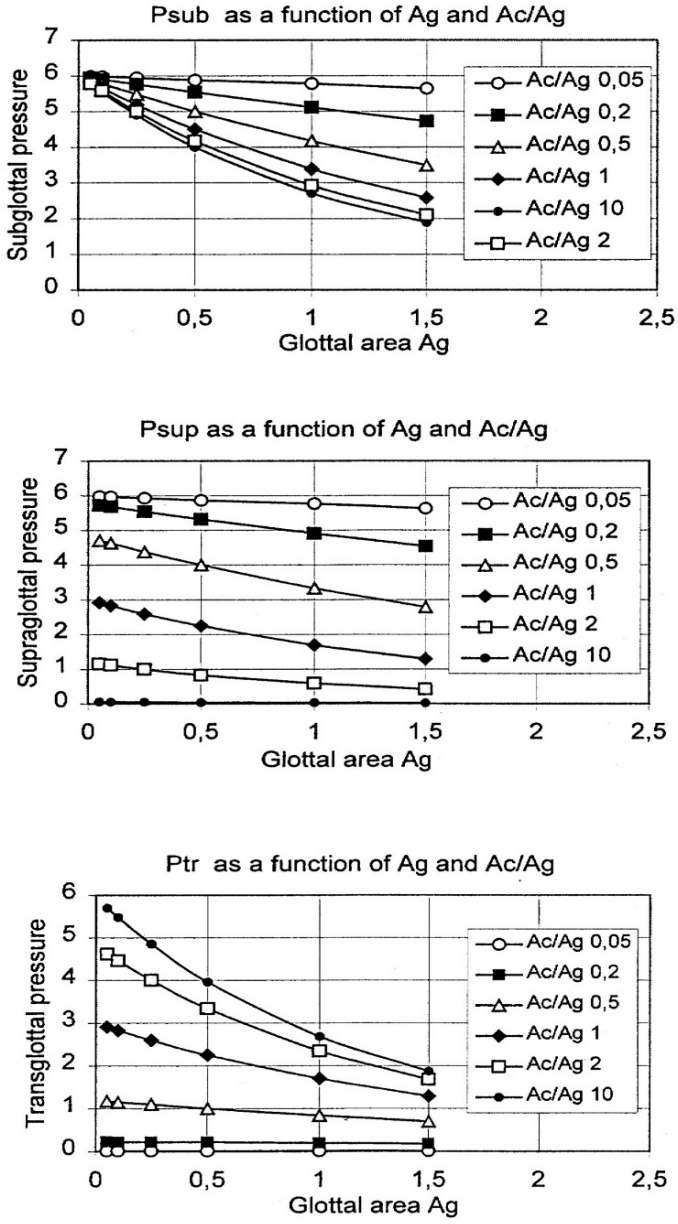


Figure 17. Psub, Psup and Ptr as a function of mean glottal area Ag and Ac. Extended Ag range.

flow adopted by Stevens (1998) is 8 cm H_2O. We now have Psub = Plung–Pt and Psup = Plung–Pt–Pg. The transglottal pressure Ptr is by definition equal to Pg. Graphs of Psub and Ptr as a function of Ag for various Ac/Ag ratios are plotted in Figure 17. Figure 18 shows Psub and Ptr within a lower range of Ag, representative of voicing.

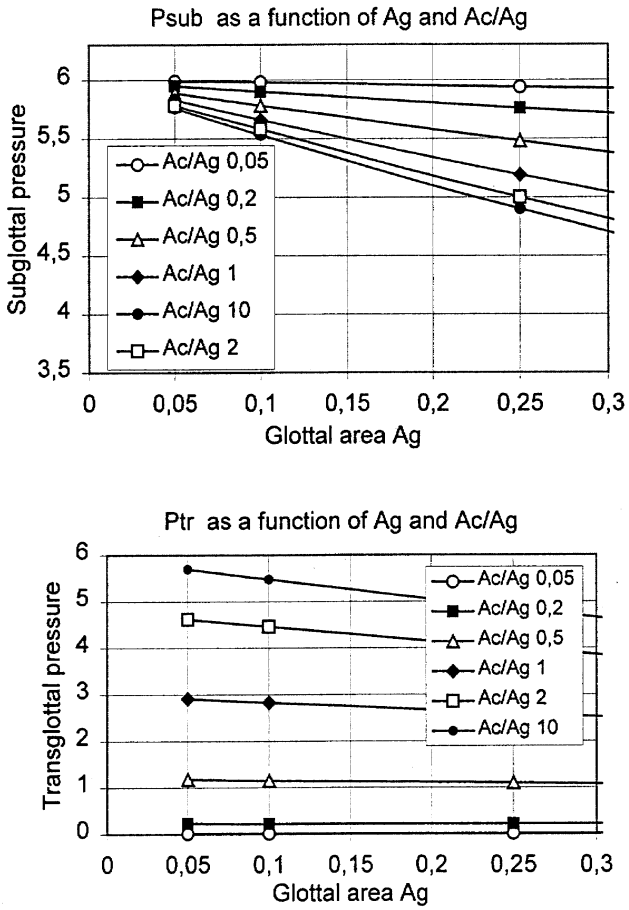


Figure 18. Psub and Ptr as a function of Ag and Ac. Limited Ag range.

The reference Psub at vocal tract complete closure is set to 6 cm H_2O. At a moderate Ag of 0.05 cm^2, and a relative open vocal tract, $Ac/Ag > 5$ typical of voicing, Psub is close to 5.75 cm H_2O and is reduced to 5.5 cm H_2O at $Ag = 0.1\ cm^2$. According to our observations, these values are typical of the decay of Psub from an unvoiced stop with complete constriction to the beginning of the following vowel. The pressure loss derives from the resistance, Rt = 1.5 ohm, in the tracheal pathways. The transglottal pressure Ptr is more dependent on the Ac/Ag ratio than on Ag alone. When Ag = Ac, it is reduced to a value close to Plung/2, i.e. just below 3 cm H_2O in our example. As a consequence, the voice source power and spectral slope are substantially reduced. This is typical of both voiced stops and voiced intervocalic /h/, the latter with an Ag of the order of 0.25 cm^2according to Stevens (1998, p. 424). One should also have in mind the aerodynamic reaction of a narrow supraglottal constriction on the voice source. As explained by Stevens (1998, p. 95), an increased

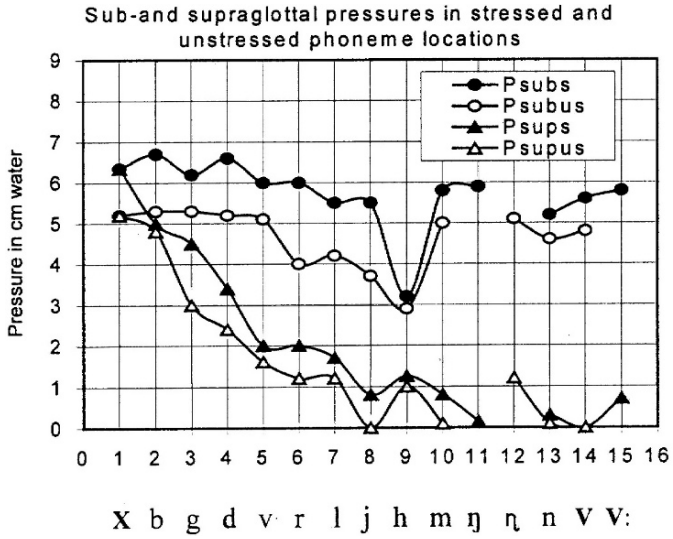


Figure 19. Psub and Psup sampled in vowels and consonants in stressed and unstressed context.

mouth pressure Psup exerts a pressure on the upper surface of the vocal folds, causing a widening of the glottal slit.

The modelling above provides a background for processing of Psub and Psup data in the extended material on prose reading of Fant, Kruckenberg, Liljencrants and Hertegård (2000). For this purpose, in Figure 19, we have tabulated Psub and Psup of major vowel and consonant classes separately in stressed and unstressed contexts.

The data in Figure 19 have been ordered in a sequence of decreasing supraglottal pressure, starting with **X**, which is the joint category of unvoiced stops and fricatives. Next follows voiced consonants, and at the right end the main vowel category, labelled **V**, and the maximally close long stressed vowels, **V**: i.e. [i:] [y:] [ʉ:] [u:].

They are produced with a diphtong like gesture towards closure which is palatal for [i:] and [y:] and labial for [ʉ:] and [u:]. The samples pertain to their maximally constricted phase. They are produced with a finite Psup of the order of 0.8 cm H_2O, which according to Figure 17 and 18 would correspond to an Ac/Ag of about 3 at Ag = 0.05 cm^2. The slightly higher Psub in category **V:** than in **V** is the consequence of a smaller Ac.

There is a finite Psup in [m] and [ɭ] but less in [ŋ] and [n], which would indicate a difference in effective mouth constriction, but this has to be verified from a larger corpus.

There is a trend of about 1.5 cm H_2O higher Psub in stressed than in unstressed consonants. In the vowels, the difference is smaller and of the order of 1 cm H_2O.

The [h] sound shows considerable variations with the extremes of a totally unvoiced stressed glottal fricative with a free supra-glottal airway and a negligible mouth pressure build-up, Psup, and at the other end, a complete lack of noise and complete segmental deletion, in which the only sign of the [h] is a slightly breathy

vowel onset, typical of unstressed inter-vocalic positions. The data included in Figure 19 pertain to the average of two mixed voice noise occurrences. The low Psub = 3,2 cm H_2O combined with Psup = 1,3 would according to Figure 17 predict a glottal area Ag of the order of 0,7 cm^2 and Ac/Ag of the order of 1.

The accuracy of mapping measured data to model data as exemplified above is limited, but acceptable for demonstrating major relations. One obvious uncertainty is the influence of varying lung pressure within a sentence. However, this technique deserves to be tried out on a greater corpus with pre-selected categorizations, and with more complete equivalent network models.

5. PROMINENCE

5.1. *The RS Parameter*

A unique feature of our system is that both F0 and duration are controlled by a continuously scaled prominence parameter which has been labelled Rs, now changed to RS. The technique for prominence rating originates from Fant and Kruckenberg (1989). A listening crew was engaged in the assessment of each syllable or word in a read text, presented over a loudspeaker in repeated chunks of the order of a sentence. A direct estimate technique, involving the setting of a pencil mark on a vertical line scaled from 0 to 35 for each syllable or word, was used. The outcome is a continuous interval scale. The standard deviation among the 15 subjects in our listening crew was of the order of 3 RS-units only, which implies an uncertainty of the means, $0.7\sigma/(N^{0.5})$, of the order of 0.4 units. These data pertain to a standard text read by a single speaker. They have been presented in some earlier publications (Fant, Kruckenberg and Liljencrants, 1999, 2000A). As a guide, average values around RS = 10 for unstressed syllables and RS = 20 for stressed syllables were suggested. We found that words received about the same RS as the dominating syllable in the word.

More representative data are now incorporated in our prosody rules. They derive from two experts judging the prose reading of five subjects. Content words averaged RS = 19 and function words RS = 11. Numerals and adjectives topped the scores with RS = 21 followed by nouns RS = 20, verbs and adverbs RS = 18.

In the class of function words the scores showed rather small variations around the mean value of RS = 11, ranging from 12,5 for pronouns to 10,5 for auxiliary verbs. Highly reduced syllables, usually articles, were graded in the range RS = 3 − 8. However, as expected, function words were occasionally raised in prominence and content words could be reduced.

5.2. *Speech Processing and Display*

A most important tool for our data collection has been the speech processing and display system *Wspect*, developed by Johan Liljencrants. Besides oscillogram and spectrogram it generates F0 on a log scale, with a standard of one semitone per 2 mm, and two intensity traces, SPL and SPLH. In some studies, as in Figure 13 records of sub- and supraglottal pressures were included. The most complete representations

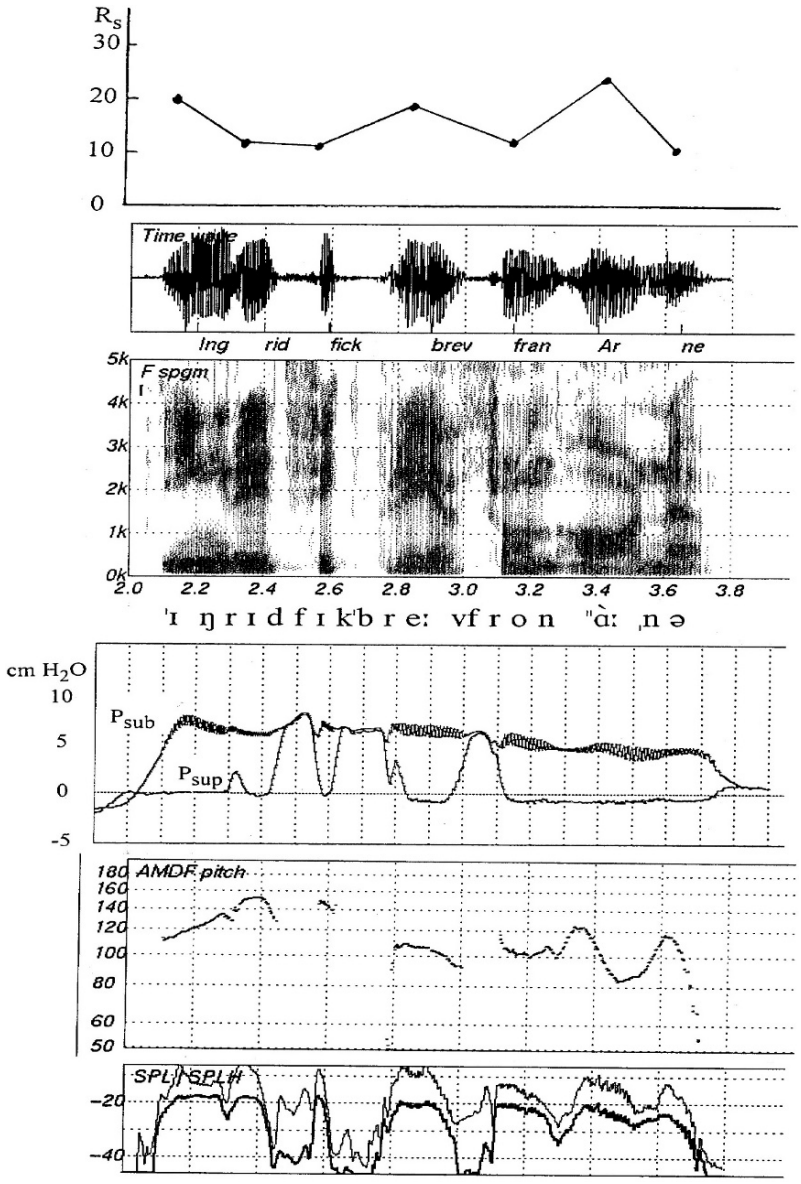


Figure 20A. Example of multi-parameter display, system Wspect. "Ingrid fick brev från Arne"

incorporate synchronized traces of syllable prominence RS, derived from listening tests. A set of 19 illustrations covering one minute of prose reading appeared in Fant, Kruckenberg, Liljencrants and Hertegård, (2000). Four of these are reproduced in Figures 20 A,B,C,D. All segmentations and data collections were performed by hand.

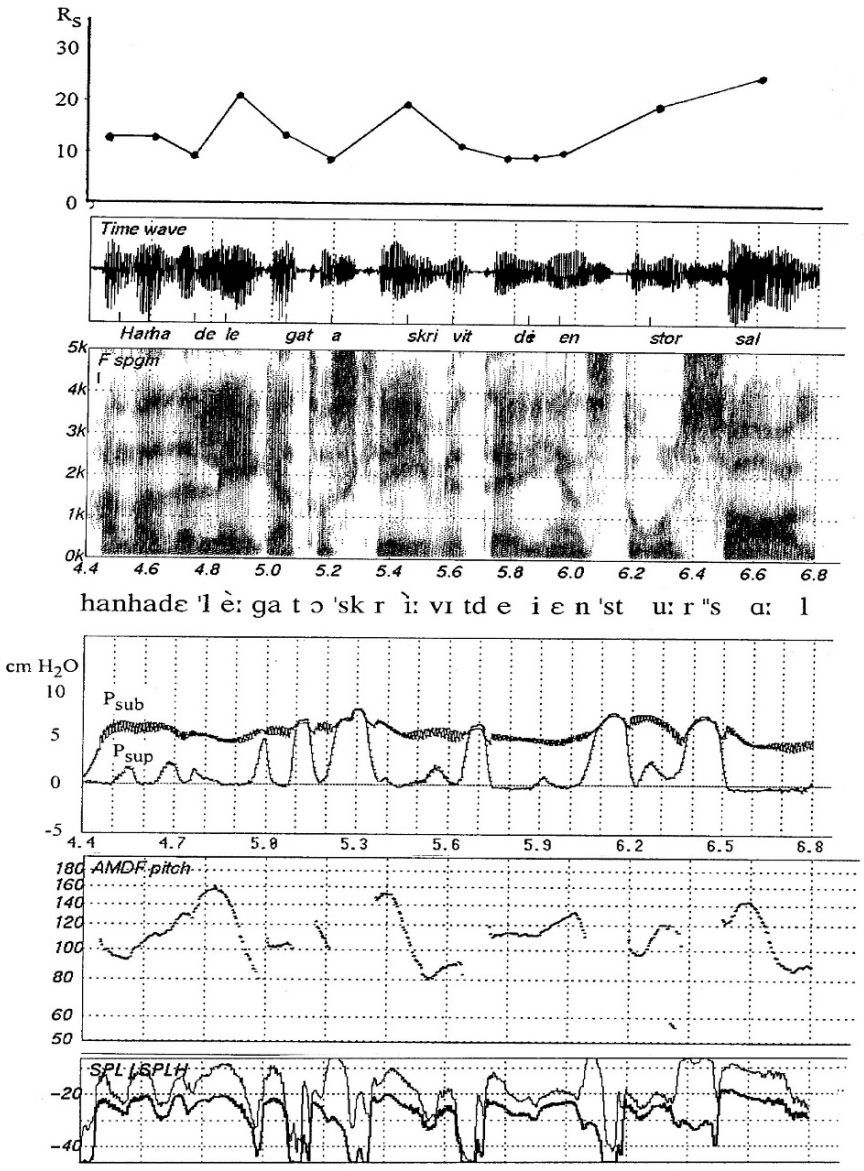


Figure 20B. Example of multi-parameter display, system Wspect. "Han hade legat och skrivit det i en stor sal—"

5.3. *Subglottal Pressure*

That subglottal pressure, Psub, is significantly correlated to prominence is well established in the literature, e.g. Ladefoged (1967); Collier (1974). Our present study aims at a more detailed view with respect to the Swedish language. We find

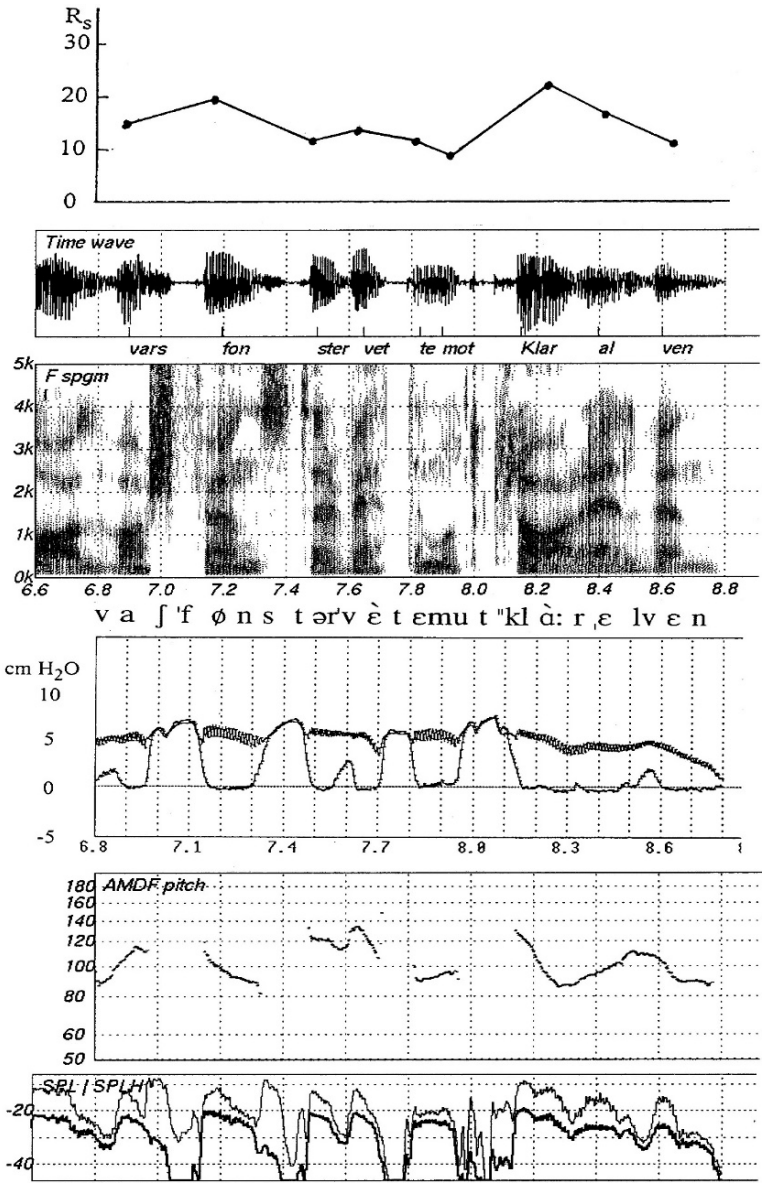


Figure 20C. Example of multi-parameter display, system Wspect. "–vars fönster vette mot Klarälven."

that subglottal pressure variations play a role not only in contrastive high degrees of stress, but also at more moderate non-focal stress levels.

The declination of subglottal pressure within a sentence showed approximately the same average falling contour as F0 and SPL An interesting finding in all recordings, see Figures 13, 14, 20, 27, is that within stressed long vowels, the subglottal

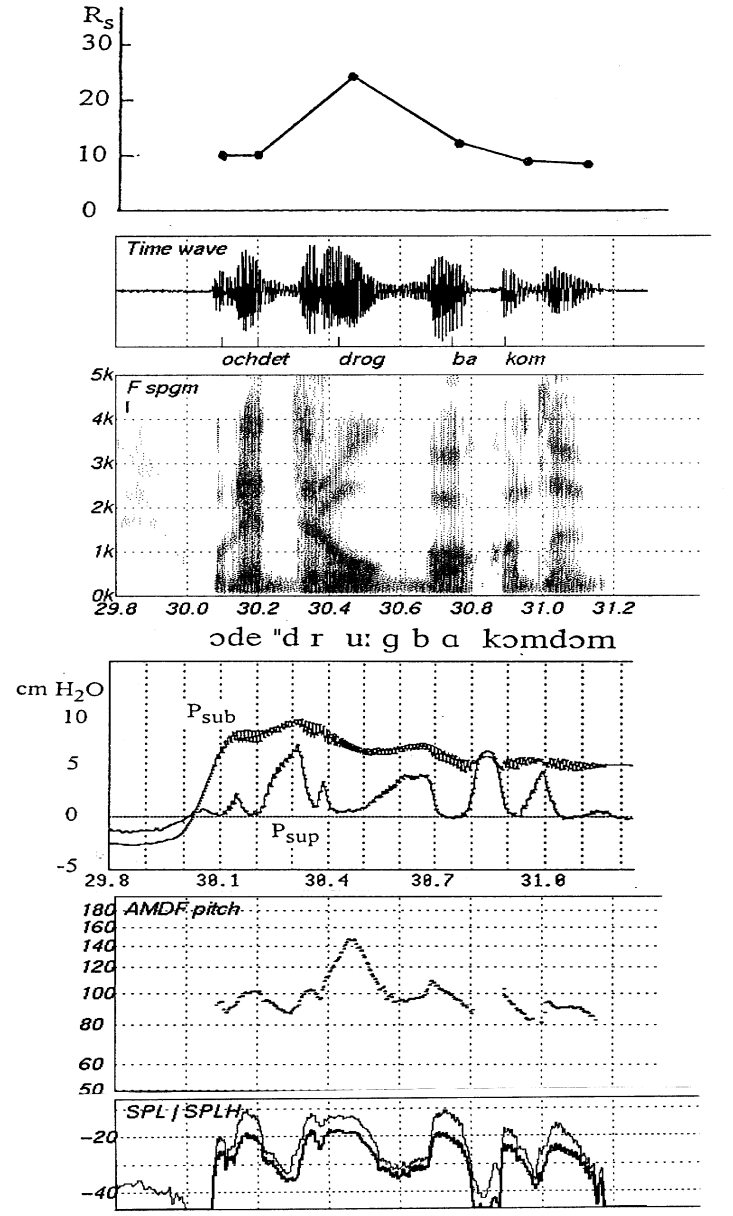


Figure 20D. Example of multi-parameter display, system Wspect. "—,och det drog bakom dom—"

pressure is elevated in advance, reaching a broad maximum at the syllable boundary and a decaying contour in the vowel, independent of the F0 contour. The rise is usually slower than the fall, see the local F0 peak in the focally stressed word "drog" in Figure 20D.

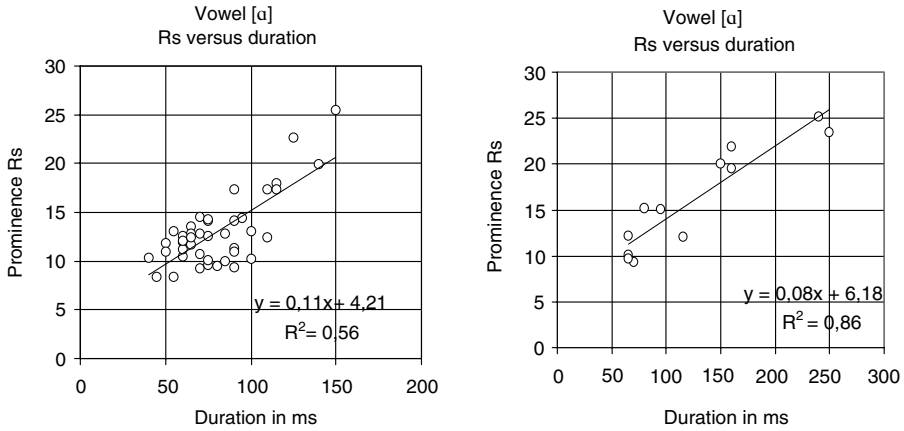


Figure 21. RS versus duration regression data from the prose reading corpus. At the left, the short vowel [a]. At the right, long vowel [ɑ:], including the lexically long but unstressed [ɑ].

From the regression data of Figure 25, we may observe that a 10 unit increase of RS, i.e. the normal span between unstressed and stressed syllables, is associated with an increase of about 1.5 cm H_2O in Psub, which is in close agreement with the findings from the consonant and vowel data of Figure 19.

5.4. *Phoneme Duration*

More extensive analysis has been made of duration as a prominence parameter.

Regression plots have been made of RS versus duration of most vowels within the entire corpus of prose reading, as exemplified by the phonemically short and long vowels [a] and [ɑ:] in Figure 21. The long vowel [ɑ:] displays a less spread regression, $R^2 = 0.86$, than that of the short [a], $R^2 = 0, 56$. The long/short phonemic distinction in Swedish is almost neutralized in unstressed context, Rs < 13, but there remain small differences in the formant pattern, and according to our data also a difference in duration, 53 ms for the short [a] and 70 ms for the phonemically long but unstressed [ɑ].

5.5. *Syllable Duration*

In several earlier studies, (Fant and Kruckenberg, 1989, 1994; Fant, Kruckenberg and Nord, 1991; Fant et al., 2000A,B; Kruckenberg and Fant, 1995) we have pointed out the systematic relation between syllable duration and stress. The data in Figure 22 reveal a close similarity between our subject SH and that of an earlier subject ÅJ reading the same text. Except for one-phoneme syllables, the stressed/unstressed difference is close to 100 ms independent of the number of phonemes in the syllable.

One phoneme syllables are made up of a single short or long vowel, in the stressed category long vowels only. On the average, each phoneme added to a syllable increases its duration by about 70 ms. Syllables in sentence and clause final position

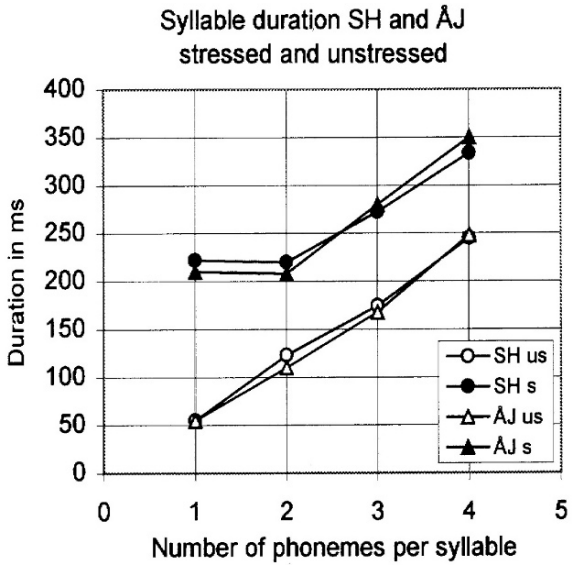


Figure 22. Stressed and unstressed syllable duration as a function of the number of phonemes in the syllable. Two subjects, SH and ÅJ. Pre-pause lengthened syllables have been excluded.

subject to a pre-juncture final lengthening, usually before a pause, have been omitted in the statistics.

Our initial labelling of syllables as stressed and unstressed was performed prior to the prominence assessments. For subject SH, an average of RS = 11 was noted for unstressed syllables and RS = 19 for stressed syllables. If we extrapolate duration data to be valid for RS = 10 and RS = 20, respectively, the resulting difference between stressed and unstressed syllables with two or more phonemes is increased from 100 ms to 125 ms. This difference is carried by 80 ms in vowels and 45 ms in consonants.

The data in Figure 22 are averages over the entire prose reading corpus, in which differences in relative complexity set by the occurrence of inherently long consonants are averaged out.

Data for these two speakers are quite similar. When they are instructed to produce a more distinct reading, the duration of stressed syllables increases, whilst unstressed syllables remain more or less the same. The stressed versus unstressed contrast is a speaker specific measure. Moreover, it is language specific. In French, which is classified as syllable timed, the stressed/unstressed contrast is considerably smaller than in Swedish or English (Fant, Kruckenberg and Nord, 1991; Kruckenberg and Fant, 1995).

5.6. *Intensity*

Intensity, usually measured in terms of sound pressure level, SPL in decibels, is a well established correlate of stress. Of special interest is to determine the relative

benefits of our new intensity parameter SPLH with high frequency pre-emphasis, Figure 11.

Because of the higher weight given to the second and higher formants, SPLH serves as a closer acoustic parameter of sonority, i.e. loudness, than SPL. Their difference, SPLH-SPL, serves as an indirect measure of voice source spectral slope in samples that have the same formant pattern. However, with increasing stress of a vowel there usually follows a shift in the formant pattern, which may raise the intensity of the second and higher formants (Fant, 1956, 1959, 1960) and thus influence the SPLH-SPL measure. Unless corrected for by a calculation, the formant pattern dependency will obscure an analysis of the voice source component. However, the combined effect raises the loudness, and thus the auditory prominence.

On the other hand, it should be recognized that extra stress on Swedish long close vowels [i:] [y:] [ʉ:] [u:] is effected by a gesture towards a partial closure, which is labial for [ʉ:] and [u:] and palatal for [i:] and [y:], and potentially lowers SPLH. The narrowing also causes a loss of efficiency of the voice source (Bickley and Stevens, 1986; Fant, 1997). In this case, prominence may be considered to follow a learnt speech motor pattern.

As a specific example we may refer to data and production simulations of the short vowel [a]. A 10-unit-increase in RS from 10 to 20 gave an increase in SPL of 5 dB and in SPLH of 8 dB. Of the observed difference of 3 dB in SPLH-SPL, 1 dB may be attributed to the voice source with the LF parameter Fa increasing from something like 500 to 1000 Hz. The remaining 2 dB derive from the difference in formant pattern, essentially the higher location of F_1 in the stressed version with $F_1 = 700$ and $F_2 = 1200$ Hz versus the unstressed pattern of $F_1 = 600$ and $F_2 = 1300$ Hz. F_3 was 2500 Hz in both cases. The unstressed formant pattern is also quite sensitive to coarticulation effects.

In connected speech the SPLH-SPL of voiced sounds show considerable variations, from 0 dB when the spectrum is totally dominated by the fundamental and up to 18 dB in maximally open sounds.

In focal accentuation, the prominence measure RS is greater than 20 and usually in the range $RS = 23 - 28$ in the prose reading.

5.7. *Spectral Tilt*

The concept of spectral tilt pertains to the slope of the voice source which could be quantified by the LF parameter FA, which is the frequency at which the rate of fall of the source spectrum increases by an additional 6 dB per octave. However, to determine FA experimentally is a rather intricate and difficult process (Fant, 1995). In our study we have instead adopted the SPLH-SPL parameter as a measure of spectral tilt, which combines the overall spectral slope of the source and the spectrum shaping of the filter function. According to the parameter evaluations in Figure 26, SPLH-SPL gave the closest correlation to perceived prominence, RS.

The transformation of changes in FA data in Hz to equivalent shifts of spectral levels in dB at a frequency f, relies on the following expression

$$\Delta L = 10\log_{10}\{[1 + (f/FA1)^2]/[1 + (f/FA2)^2]\} \quad \text{dB} \tag{27}$$

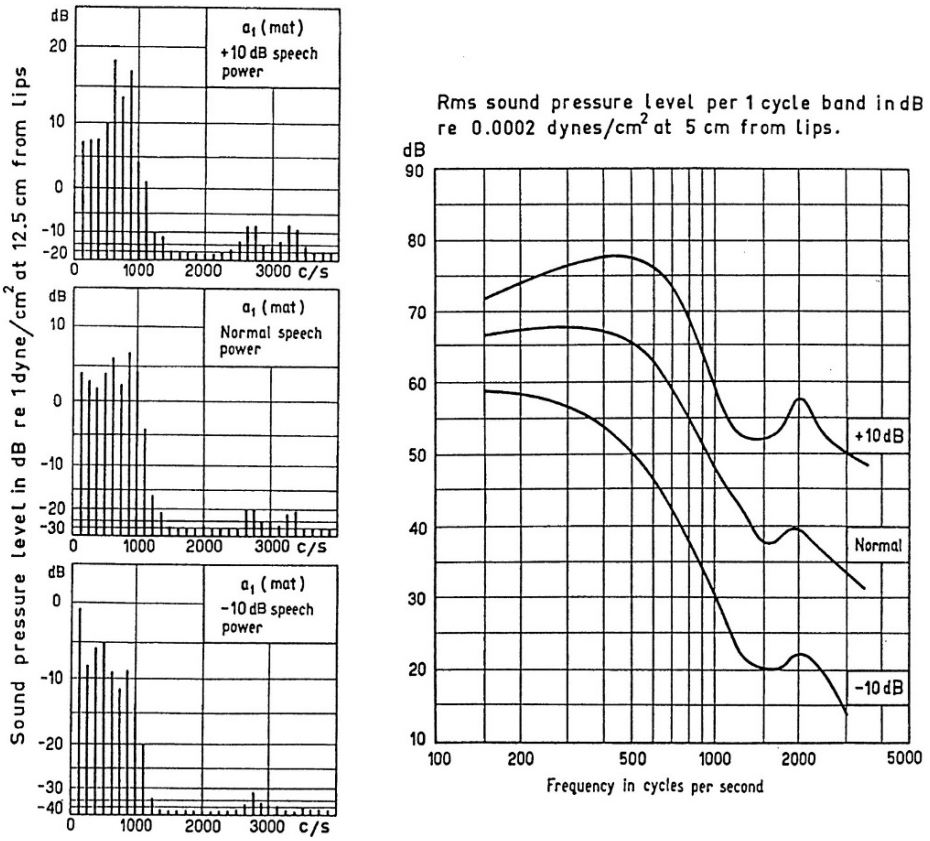


Figure 23. Spectrum changes comparing weak, normal, and loud voice. Vowel [ɑ:] and spectra averaged over the reading of a piece of connected speech. From Fant (1959, 1995).

From RS = 10 to RS = 20 the LF slope parameter FA increases from about 750 Hz to 1500 Hz, which according to Eq. 27 would add 5 dB to the spectral level at 2500 Hz. As discussed above, the FA shift accounts for only 1 dB of the 3 dB overall stressed/unstressed difference in SPLH-SPL of the vowel [a].

The non-linear spectrum shift associated with increasing voice effort as reported by Fant (1959), originating from his earlier work at the Ericsson Telephone Company in 1967 (Figure 23), still constitutes a representative reference material. An increase of the amplitude of the first formant, F1, by 10 dB was accompanied by 4 dB in the voice fundamental and 14 dB in the level of F2, 16 dB in F3 and 14 dB in F4. These data from long time average spectra of a sentence are supplemented by spectral sections of an [ɑ:] vowel, sustained at three corresponding loudness levels. A unique feature of this graph is the absolute calibration of sound pressure level at a fixed microphone distance. It has enabled an indirect derivation of the underlying glottal flow (Fant and Lin, 1988).

As outlined in Fant (1995), the total range from weak to loud voice in Figure 23, which amounts to 10 dB at the average voice fundamental F0 = 125 Hz, 27.5 dB

at 500 Hz, 32 dB at 1500 Hz, and 33 dB at 2500 Hz, may be quantified within the frame of the LF model as a decrease of the Rd parameter from Rd = 1.6 to Rd = 0.5, corresponding to opening quotients OQ of 75% and 47% and FA values of 275 Hz, respectively 1420 Hz. The voice flow amplitude, Uo, would increase by 10 dB and the Ee parameter by 20 dB.

An implication of the +20 dB step in Ee but only +10 dB in Uo is the relative stability of glottal flow which should be considered in voice source rules. According to Eq. 27 the difference in FA, and thus in spectral tilt, adds a high frequency boost of 13 dB. The total spectral gain at 2500 Hz is thus 20+13 = 33 dB in exact agreement with the observed data above. However, at the loud voice effort we might anticipate an additional vocal fold adduction which could bring up the FA to a higher value than 1420 Hz. Thus, with FA2 = 2500 Hz an additional 3 dB gain would be anticipated.

5.8. *The F0 Pattern*

A non-specific sampling of F0 in vowels, as in Figure 26, is not an adequate base for prominence correlation. Stressed/unstressed comparisons have to be made with respect to the specific Swedish word accent 1 or 2, and with due consideration to contextual constraints within a sentence.

Figure 24 illustrates F0 contours of Swedish accent 1 and accent 2 in sentence medial position for Rs = 15, 20, 25 and 30. These patterns were derived from the prose reading of three males and two females (Fant and Kruckenberg 2000A,B). The major effect of increasing prominence is the growth of F0 in the main syllable of accent 1 and in the secondary syllable of accent 2. The accent 2 H*L fall also increases with prominence, but saturates at about RS = 22. In the interval RS = 15 to 25 the F0 increase in the dominant syllable is 8 semitones for both accent 1 and 2.

What is said above about the accent 2 secondary syllable also applies to the primary syllable of accent 1. A common observation is that the duration of the focal F0 peak measured at its base is about 250 ms.

5.9. *An Overview of Acoustical Correlates to Prominence*

A major part of our work has been devoted to studies of the acoustical correlates of perceived prominence. The stressed/unstressed contrast from RS = 10 to RS = 20 is associated with an increase of syllable duration of the order of 125 ms of which about 85 ms in the vowel part. The increase in SPL is of the order of 4–6 dB and in SPLH 6–9 dB, the higher values for vowels with a substantially higher F_1 locations in the stressed than in the unstressed version. The glottal source excitation amplitude Ee shows an increase of the order of 3–5 dB, i.e. somewhat lower than SPL.

Figure 25 shows regression lines for increase of RS with respect to Psub, SPL, duration and SPLH of a vowel [a] sampled at an early part of a new sentence. We have noted regression coefficients R^2 of 0,71 for Psub, 0,76 for SPL and 0,82 for

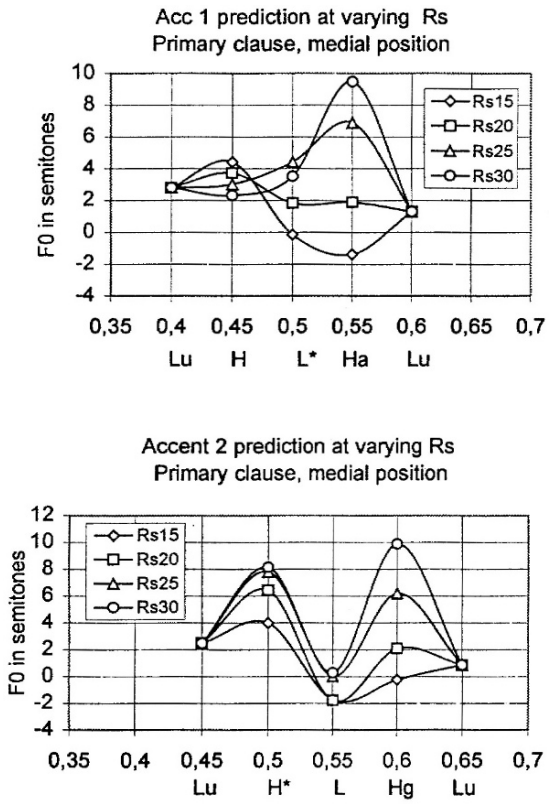


Figure 24. Accent 1 and accent 2 F0 parameters at Rs = 15, 20, 25, and 30 representing the mean of five subjects' prose reading. (Fant and Kruckenberg, 2000a,b).

SPLH. Because of the restricted sampling range the spread in duration of the [a] in this study, $R^2 = 0,8$, was less than the $R^2 = 0,56$ for the unlimited sampling in Figure 21. This is a reminder of the need to relate descriptive data to specific contexts, the "ceteris paribus" principle of Roman Jakobson (1952).

As shown in Figure 26, we found R^2 values of 0,87 for the parameter SPLH-SPL, which is related to relative sonority and spectral tilt, see section 5.6. The highest correlation score was obtained by combining data on duration and SPLH-SPL, which gave an $R^2$of 0,90. For the voice source scale factor Ee we found an R^2 value of 0,6, which is less than the value 0,8 for SPL. This may be expected, since intensity depends on both source and filter, i.e. formant locations.

Data on F0 samples gave a rather low score, $R^2 = 0,45$. This could be expected, since the perceptual impact of F0 is related to accent modulations, which can be quantified only by a prosodic decomposition of a sentence.

These and other data at our disposal allow an estimate of expected parameter variations in the range from RS = 11 for unstressed syllables to RS = 20 for stressed syllables.We find an increase of SPL of the order of 3–5 dB and 6–8 dB in SPLH.

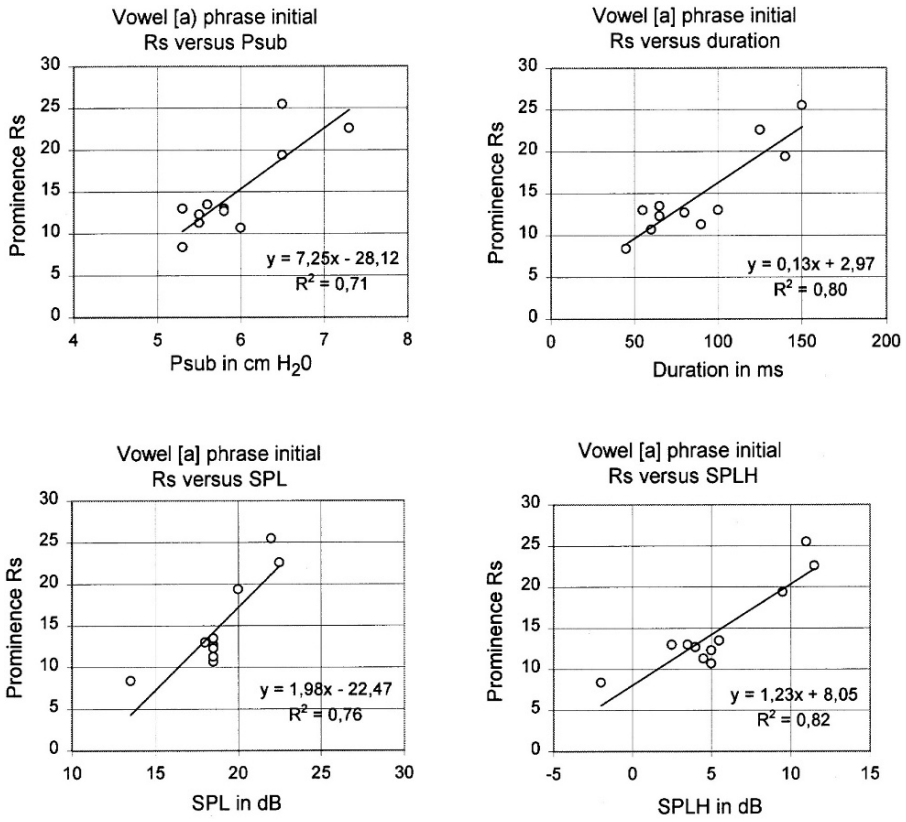


Figure 25. Regression graphs of Rs versus Psub, duration, SPL, and SPLH. Vowel [a] sampled in early positions within a major prosodic group.

Vowel duration increased by 70–100 ms, the higher values for open vowels. Syllable duration increased by 120 ms. Accent modulation depth in F0 was of the order of 3–7 semitones, the higher values for accent 2.

Similar results were obtained in an earlier study (Fant, Kruckenberg and Liljencrants, 2000A) directed to accented syllables in the range of RS>15 where we noted an increase from RS = 15 to RS = 25 of 6 dB in SPL, 9 dB in SPLH, and thus 3 dB in SPLH-SPL. Corresponding F0 shifts of 4–8 semitones were noted.

In Figures 27A,B,C, section 6, illustrating neutral versus focal accentuation, we observed F0 shifts from 6 to 9 semitones, i.e. of the same magnitude as in the data above. Associated shifts of the order of 6 dB in SPL and 8 dB in SPLH were noted. An interesting observation is that rather small differences in duration were found within contrasting pairs. This finding may in part be speaker specific, but supports a general conclusion of our work, that duration is not a necessary component of focal prominence. However, in the range of Rs < 20, duration appears to be of main importance.

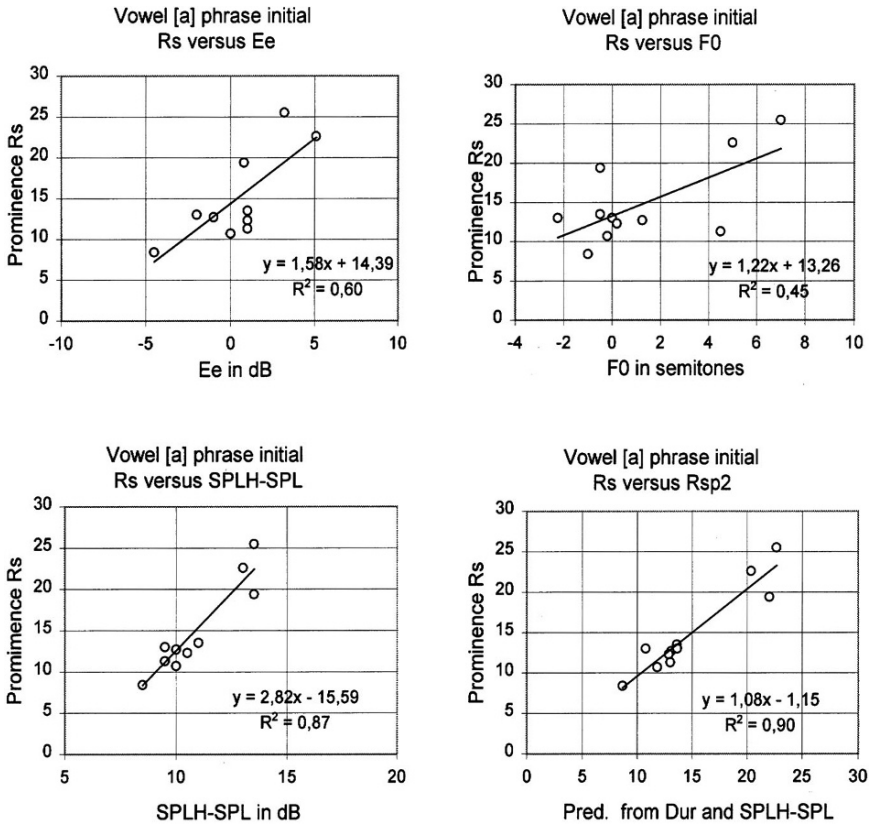


Figure 26. Regression graphs of Rs versus Ee, F0, SPLH–SPL and Rs versus a joint prediction from SPLH-SPL and duration. Same set of vowels as in Figure 25.

A change in formant pattern towards that of a more open articulation enhances SPL and even more so SPLH, which can be regarded as a measure of sonority, i.e. loudness. However, Swedish long narrow vowels [i:] [y:] [ʉ:] [u:], when highly stressed, are articulated with a closing gesture reducing SPL and SPLH, which is perceptually relevant.

A glottal adduction at constant Psub contributes to raising the efficiency of the voice source with increase in both SPL and SPLH-SPL, through FA (Fant, 1997); (Fant and Kruckenberg, 1996). As pointed out by Hanson (1997B), glottal adduction is potentially a means of executing stress without an increase of subglottal pressure, whilst focal accentuation usually requires an active Psub increase.

The role of spectral tilt as a parameter contributing to signalling emphasis in focal accent, is by now well established, but opinions differ whether it also contributes to the stressed-unstressed contrast at lower prominence levels. Sluijter and van Heuven (1996) report positive observations while Campbell and Beckman (1997) state that spectral tilt requires an accent contrast. In our data we have frequent examples

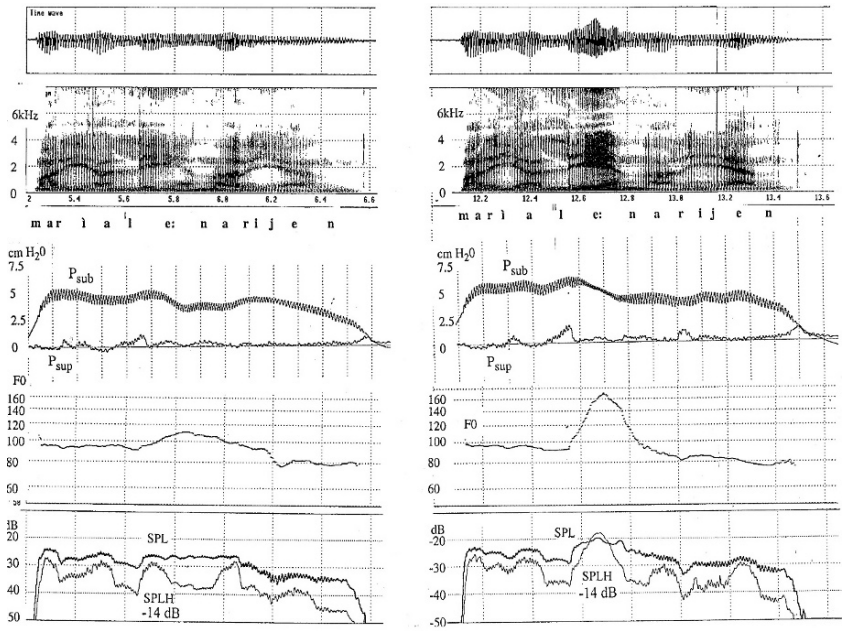


Figure 27A. Normal and high prominence of the accent 1 word Lenar in "Maria Lenar igen" [maria'le: narijɛn].

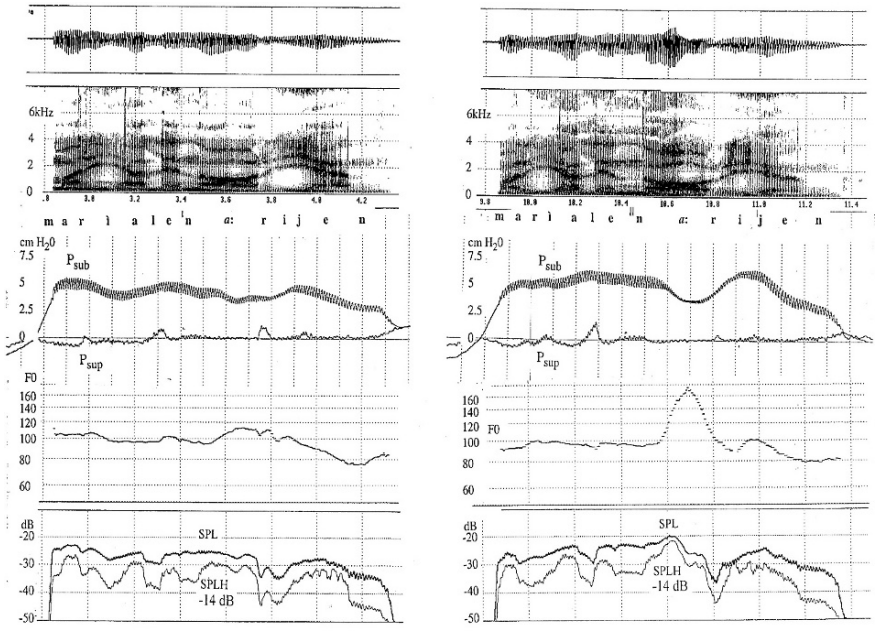


Figure 27B. Normal and high prominence of the accent 1 word Lenar in "Maria Lenar igen" [maríale'n a:rijɛn].

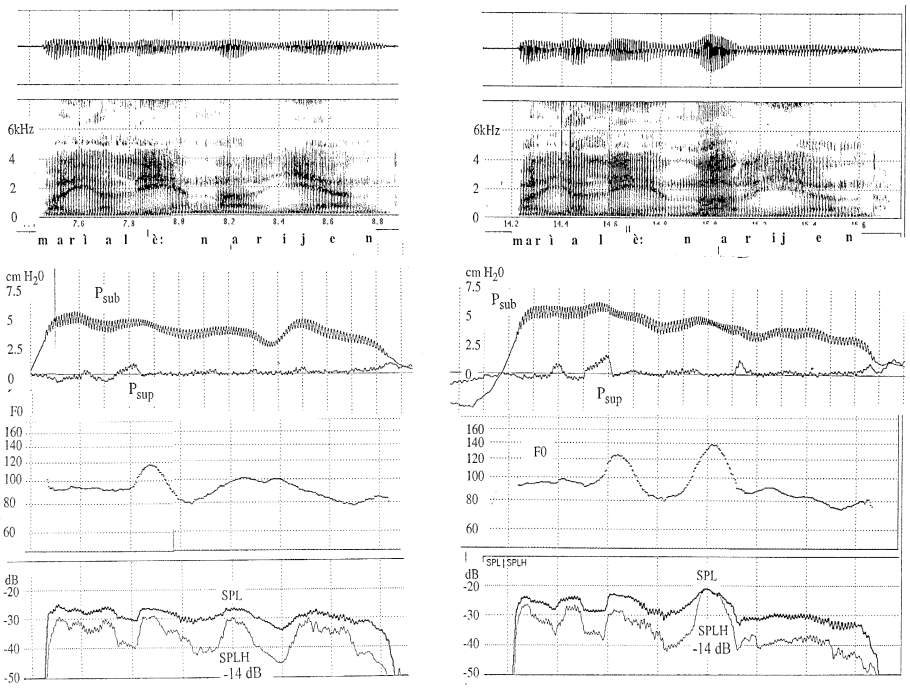


Figure 27C. Normal and high prominence of the accent 2 word "lenar" in "Maria lenar igen" [maria'lè:ˌnarijɛn].

of stressed/unstressed contrasts at lower prominence levels without a spectral tilt component. On the other hand, our data in Figure 26 indicate that the SPLH-SPL parameter is effective even at moderate RS levels.

Synthesis experiments can add to an insight in the relative salience of prominence related acoustical parameters. However, to establish a percentage contribution from each of the acoustical parameters is not a straight forward statistical process. Their role is context dependent, and free variations exist.

Duration carries a substantial load at low prominence levels, and is of minor importance at higher prominence levels. The accent 2 H*L fall shows a similar pattern. Its contribution to RS saturates at about RS = 22, where the F0 peak in the secondary syllable takes over. The intensity parameters SPL and SPLH have a more uniform function over a larger RS domain. The importance of the spectral tilt component is more apparent at high than at low prominence levels.

We need further insight in the relative perceptual salience of multiple cues, but with due regard to inherent constraints in the production mechanism, in order to avoid unnatural combinations. An important issue is to gain experience concerning individual variations and speaking style.

5.10. *Secondary Prominence Correlates*

Prominence is signalled not only by the basic prosodic parameters, F0, duration and intensity as determined by underlying lung pressure gestures and voice source properties, but also by articulatory patterns in a hyper/hypo domain. We have already noted that, with increasing prominence, open vowels are produced more open, and the close long Swedish vowels more closed. The basic principle is the approach towards extreme versus neutral articulatory targets. An associated tendency is that consonants lose noise component in unstressed contexts. An example is voiced [h] which has lost the glottal articulation component and assimilates vocal cord adduction. This is very much a matter of coarticulation. A far going reduction eliminates the [h] segment, and there remains but a slightly breathy onset of the following vowel.

Stress increases the spectral contrast between adjacent consonants and vowels, but also the temporal intensity contrast in boundary regions, as a result of a higher articulatory force. Thus, in Figures 27A,B,C the focal versions show larger supra-glottal pressure peaks in voiced consonants indicating more complete closure.

Glottalization, i.e. a brief interval of vocal fold closure interrupting the voice source at a vowel onset, can have the function of a stress prompter. The typical situation is at a voiced juncture between two words. An example can be seen in Figure 20A at the time 3.3, where there is a distinct intensity minimum preceding the last word “Arne”.

A stressed syllable may also be prompted by a following juncture. In normal accent 1 the rise in F0 is usually continued into the next syllable, see the first word in Figure 20A. At a higher contrastive stress level, the accentuation is realized as a single F0 peak and fall in an interval of apparent sub-glottal pressure decay, which contributes to signalling prominence and a post-focal boundary.

6. INTONATION ANALYSIS AND MODELLING

6.1. *The Swedish Word Accents*

In Swedish we have two distinct tonal accents referred to as accent 1 and 2 (Bruce 1977). There exist a rather limited number of word pairs in which the tonal pattern distinguishes meaning. A classical example with traditional accent notations is “ánden” (the duck) indicating a rise in the accented syllable, whereas in the accent 2 word “ànden” (the spirit) there is a fall.

However, the main importance of the accent distinction is to preserve a correct pronunciation. In connected speech more than half of the content words carry accent 2, which dominates di-and polysyllabic words.

With our notations, essentially derived from the canonical work of Bruce (1977), the accents carry modulation contours, which in disyllabic words are transcribed

Accent 1	(H)	L* Ha Lu
Accent 2		H* L Hg

L* and Ha define two sample points in the voiced part of an accent 1 primary syllable. H pertains to a preceding unstressed syllable, which may be absent in sentence initial position. It is of secondary importance only. When present it acts as a possible reference point for connecting to the following low point L*. Unaccented syllables are denoted Lu.

In accent 2 the sample points H* and L in the primary syllable are followed by a high point Hg in the next or a later syllable. In compound words Hg is located in the final constituent, but may not be strictly syllable bound.

All monosyllables carry accent 1 in their lexical form as pronounced in isolation. Increasing prominence of accent 2 words is only in part related to the size of the H*L fall. It saturates at a moderate stress level, at which Hg takes over as the major stress correlate. The major role of the H*L fall is to signal the identity of accent 2.

6.2. *Data from Lab Sentences*

The following sentences are exemplified in Figures 27 A, B and C

"Maria Lenar igen" [maria'le:narijɛn] (accent 1)
"Maria Lenar igen" [mariale'nɑ:rijɛn] (accent 1)
"Maria lenar igen" [marialè:ˌnarijɛn] (accent 2)

In each Figure, the test words are produced with normal and high prominence, something like RS = 20 and RS = 26 respectively.

With high prominence, the dominating F0 peak is located on the vowels [e:] and [ɑ:], corresponding to Ha of accent 1 and Hg of the secondary syllable in accent 2 of the vowel [a]. All three peaks have almost the same shape and height of about 10 semitones and a duration of 300 ms, which indicates the same physiological origin, i.e. the same muscular gesture. This is a useful prototype. Peaks are centred in the vowel, and the falling branches aim at a terminal juncture. In contrast, at the low RS level, there is a smoothed-out F0 contour continuing to the right.

Figures 27A, B and C also illustrate the typical subglottal pressure build-up in advance of major stresses, and the role of SPL and especially SPLH as stress correlates. In these examples the duration patterns at normal and high prominence are rather similar. Duration differences tend to saturate at high prominence levels.

7. PROSE READING

7.1. *Normalization*

Our main corpus for intonation analysis and modelling derives from the readings of 3 males and 2 females of a two-minute long passage from a novel (Fant, Kruckenberg, Gustafson and Liljencrants, 2002). In order to derive representative average intonation contours, we have employed a system of frequency and time normalization.

A basic requirement is the semitone scale. We have introduced a fixed semitone scale with the unit St defined by

$$\mathrm{St} = 12[\ln(\mathrm{Hz}/100)/\ln 2] \tag{1}$$

which attains the reference value of St = 0 semitones at 100 Hz, St = 12 at 200 Hz and St = −12 at 50 Hz.

The conversion from semitones to frequency is accordingly

$$\mathrm{Hz} = 2^{\mathrm{St}/12}100 \tag{2}$$

The semitone scale preserves the main shape of male and female intonation contours. The first stage of the normalization is to subtract a subject's average F0 in St units from his or her St contour, which provides a common base for males and females. In our study we found long-time average St values of +9,5 and +7 for the two females, and −1, 0 and +1 respectively for the three males.

The next step in the normalization pertains to the time scale. Individual differences in utterance length are removed. This is accomplished by a sampling of F0 data limited to one or two measures per syllable. Accented syllables receive two measures, L* and Ha for accent 1, and H* and L for accent 2. All other syllables receive one measure only, which applies to the secondary peak Hg of accent 2, the initial H of accent 1 and all unstressed syllables, denoted Lu.

In a final stage of synthesis we include predicted durations and the exact timing of the F0 data points within a segmental frame.

We now have an efficient tool for comparing individual speakers within the same frequency and time frame. A general observation is the small inter-subject spread in accent modulation depth, in specific of the H*L fall in the primary syllable of an accent 2 word, which is of the order of one semitone only. In addition, normalized H* values constitute rather stable anchor points for the upper bound of an intonation contour, with an inter-subject standard deviation of somewhat less than two semitones. This is illustrated in Figure 28.

The relatively large spread in the early part of the sentence, to be seen in the upper part of the figure, is explained by some individuals inserting a juncture before a long preposition phrase.

The lower part contains the mean curve of the five subjects and our reference female subject. Except for a higher initial value and a lower final value, she has a higher starting point and a lower final value. However, except for individual global gestures, the main trends within a sentence are the same for males and females.

Our normalization procedure has been quite successful. Observe the similarities in overall declination rates as well as in absolute F0 levels. Subjects are normalized with respect to long time averages and not to their reading of the specific sentence.

7.2. *Prosodic Grouping*

A complete sentence of moderate size may be produced as a single prosodic group with certain rise and decay characteristics in F0, or as a succession of groups each carrying an intonation module, i.e. a base curve upon which accent modulations are superimposed. At present we employ four different modules depending on position. In sentence final positions before full stops they receive an extra F0 lowering.

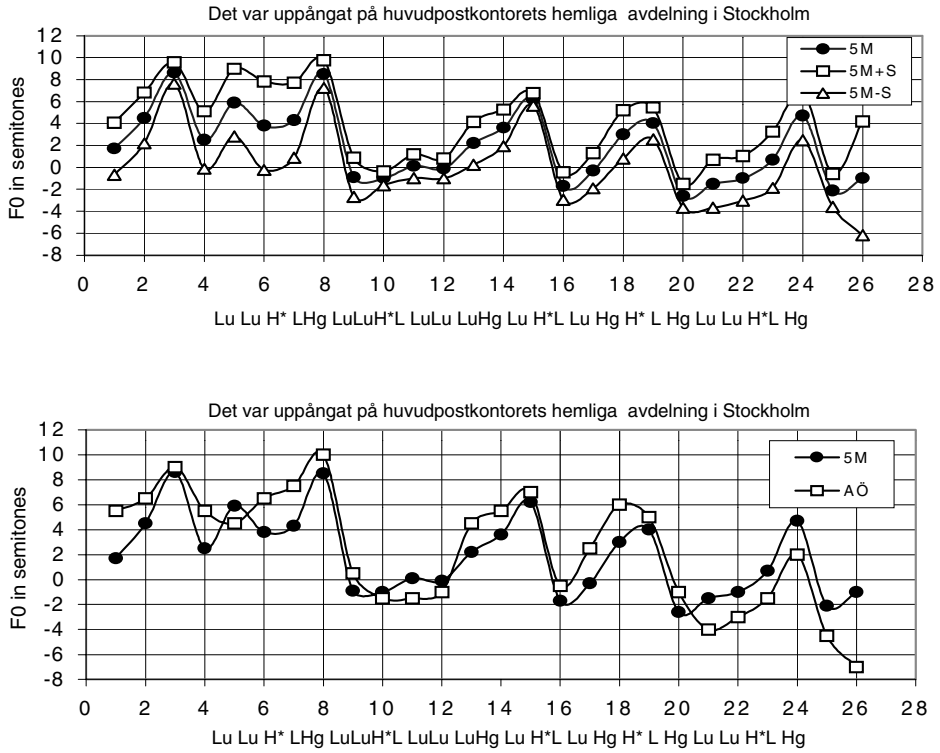


Figure 28. Above, mean of five subjects' sampled intonation contours and the mean plus and minus a standard deviation. Below, a comparison of our reference female subject AÖ and the mean of the five subjects.

A possible sequence of three base curves is shown in Figure 29.

In boundary regions, i.e. at junctures, there is always a final lengthening with or without a proper pause. Individual variations have been studied by Fant, Nord and Kruckenberg (1986); Fant and Kruckenberg (1989) and by Fant, Kruckenberg and Barbosa Ferreira (2003). Pause duration tends to be proportional to the F0 reset. Typical of the juncture at a sentence boundary, defined by a full stop in the text, is an F0 reset of 7 semitones and a pause close to 1000 ms. At the juncture between two main clauses, the F0 reset is of the order of 3,5 semitones combined with a pause of about 400 ms. At lower syntactical levels, there exist a large number of possible combinations of syntactical constituents, where smaller values of F0 reset and pause duration may be expected. However, in our experience, individual variations in reading are excessively large. We have also noted a systematic difference between our prose reading and news reading over the radio, in which sentence pauses were of the order of 500 ms (Fant, Kruckenberg and Barbosa Ferreira, 2003).

A syntactic parser exploited to its full capacity to define junctures may generate unnatural breaks, which should be avoided. Work along these lines has led us to

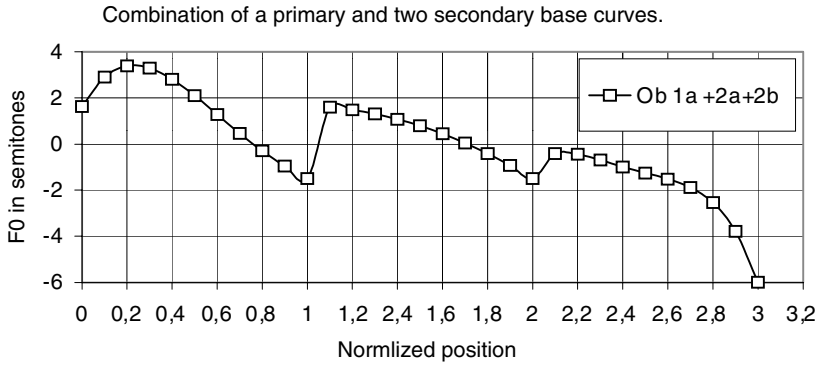


Figure 29. Example of successive base contours in a long sentence divided into three major prosodic groups.

predict junctures and pauses on a probability basis, and with respect to the size of constituents. Thus, in our collected data, the boundary before a subordinate clause was only to 30% realized by a pause, and to 70% by terminal lengthening only. Similar values were observed for noun phrases. Before and after a preposition phrase, 92% of all junctures were realized by final lengthening only. On the other hand, we encountered relatively large pauses between major constituents of a complex noun phrase.

Our five speakers had quite similar pause durations close to 1000 ms at a full stop, but showed a large spread in the number and total time allotted to pausing within a complete sentence. This is illustrated in Figure 30.

7.3. *Duration*

Our modelling of temporal structures is based on a separate databank of measured duration of phonemes and syllables, structured by sequential constraints and

Figure 30. Average values of sentence internal total pause duration, and pauses between sentences. Five subjects.

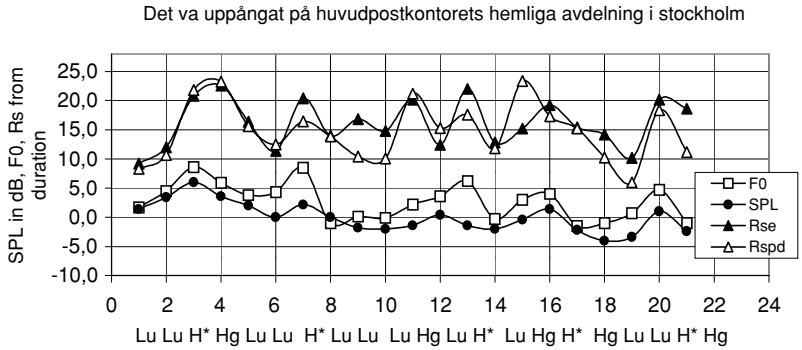


Figure 31. Syllable prominence predicted from duration, Rspd, and assessed from listening, Rse. The two lower curves show F0 and SPL.

prominence levels. Specific values are derived from the RS parameter, by linear regression within a frame of reference values for stressed and unstressed positions. All phonemes within a syllable share the same RS. The definition of syllable boundaries is thus of some importance.

The reverse process, i.e. predicting RS values of a syllable from the duration of its constituent phonemes in text reading, provides a means of estimating the perceptual salience of duration as a cue to perceived prominence. We have found a high correlation between a sequence of RS estimates from listening and a sequence of RS estimates from duration alone, see Figure 31.

8. SYNTHESIS

8.1. *The Rule System*

Our prosody rules have been tested in an Mbrola diphone environment, with access to the Infovox database for conversion from text to phonetic form, with word class and accent tagging included. In addition, we use their database of one or several reference speakers for phoneme to sound conversion on a diphone basis. These have been spoken with monotone pitch, which is a requirement for insertion of specific intonation patterns.

As outlined in the previous sections, the programming employs the following steps:

1. A prosodic grouping in terms of a sequence of intonation modules, with associated pauses and related boundary conditions, is carried out.
2. Each syllable is assigned an RS. Default values are lexically derived.
3. Syllables carrying primary stress are given two F0 data points, and other syllables one point, see section 6.1.
4. Syllable positions within a prosodic group are converted to relative positions in a scale from 0 to 1. This normalization is motivated by the tendency of the total F0 decay within a prosodic group to be independent of its length.

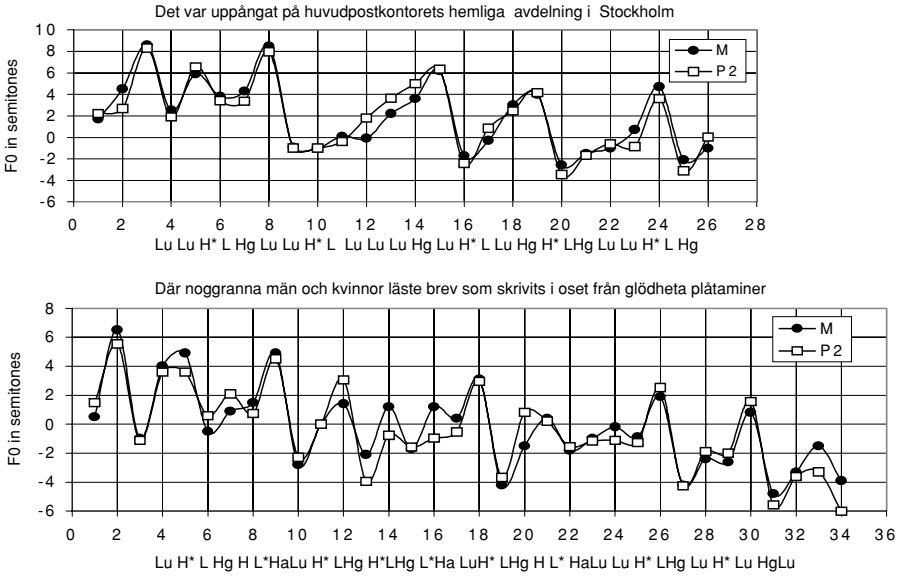


Figure 32. Predicted F0, P2, and measured F0, M.

5. F0 of unstressed syllables are placed at appropriate positions within a prosodic groups. Accented syllables attain F0 modulations superimposed on an intonation module.
6. F0 of accented syllables are derived from RS values and relative position within an intonation module, and to some extent also from the position of the module within a complete sentence.
7. Durations of sound segments are calculated.
8. Special rules apply to the positioning of F0 points within the temporal frame.

Figure 32 shows an example of a normalized F0 contour of the average of our five speakers' spoken data, and the corresponding contour predicted from our general rules. There is a close agreement. The average departure is of the order of 1,5 semitones only, which covers accent modulation as well as more global features related to intonation modules.

Our synthesis by rule for Swedish is judged to have a prosody superior to other systems demonstrated up till now. The segmental concatenation inherent in the Mbrola system accounts for some audible degradation, which is unavoidable, but this is a minor problem.

8.2. *A Note on Auditory Integration of Pitch*

In Mbrola synthesis, intonation contours and accent modulations are straight line approximations of the true continuous F0 curves of real speech. It is remarkable that the percept all the same is quite convincing. We have made a test comparing two realizations of the Swedish accent 2 word [ànna], in which the second syllable,

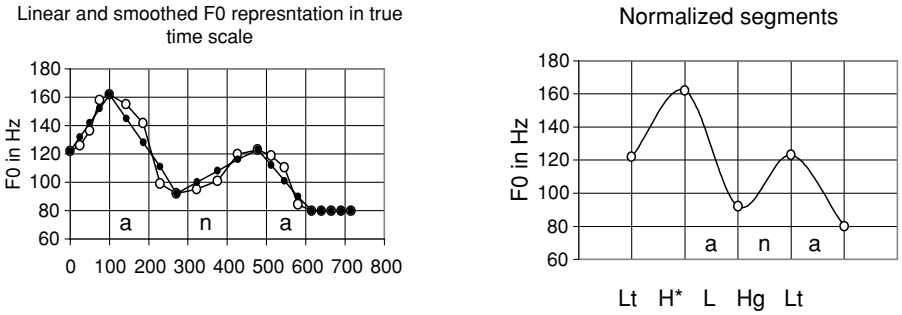


Figure 33. Straight line and cosine approximation of F0 in the accent 2 word "Anna". To the right the normalized intonation contour.

the vowel [a], attains a typical F0 peak. One is programmed from our rules with straight line F0 components, the other with a cosine approximation of each line, see Figure 33, which also shows the normalized representration.

No major differences could be detected between the two versions, but a careful listening revealed, that the linear approximation had a pitch of about one half semitone below that of the more gradually shaped contour.

The tonal percept thus appears to follow rules of auditory integration, which could be modelled in analogy to linear or maybe nonlinear systems. A guess would be a time constant of the order of 50–100 ms.

These findings can throw some light on alternative means of modelling F0 contours. The Fujisaki method of modelling accentual modulations by means of second order linear filter responses to a combination of step functions has its root in the early work of Öhman (1967). Applications to Swedish tone accents appear in Fujisaki, Ljungqvist and Murata (1993). A recent study with respect to Chinese is Fujisaki et al. (2000).

In the light of the auditory smoothing of F0 contours, it now appears to be sufficient to use linear approximations to the rising and falling parts of local accentual modulations. In addition, as in Figure 33, one may add an automatic smoothing to create intonation contours of a more natural appearence. The filtering part of the F0 encoding, inherent in the Fujisaki model, may thus be excluded, and the synchrony within the segmental structure is facilitated by a direct choice of high and low insertion points. In our system, the prosodic modules related to underlying phrase or sentence intonation, are modelled by a higher order polynomial curve fitting to the average of a representative population of speakers, as was exemplified in Figure 29. The pro and contra aspects of the two different approaches need to be discussed.

8.3. *Multi-Language Applications*

We have recently attempted to transfer experience from our Swedish rule system to synthesis of English and French. Our preliminary results are quite promising, especially with respect to French prosody, which is illustrated in Figure 34.

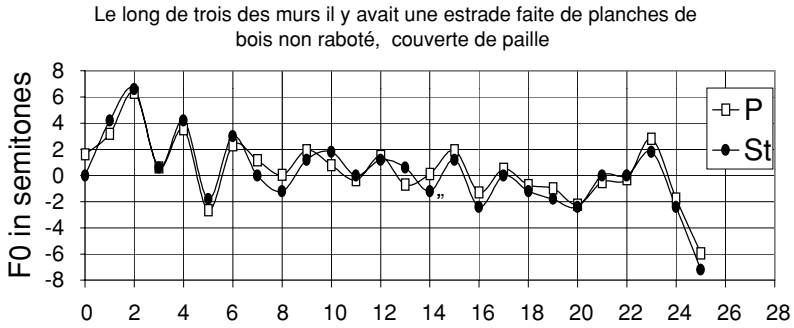


Figure 34. Measured F0, St, and predicted F0, P. The text is: "Le long de trois des murs il y avait une estrade faite de planches de bois non raboté, couverte de paille".

This is a normalized graph of a spoken sentence and a prediction from tentative rules. The close match is to some extent influenced by the analysis-by-synthesis performed on the training material, but our tentative rules have functioned well also in other sentences.

For French we have introduced a modified version of our Swedish accent 1, which accounts for the typical iambic pattern of word intonation within a prosodic group. The final rise, typical of sentence internal prosodic groups, can generally be introduced without a specific intonation module by a high RS value in the last content word. Sentence final groups have the same or larger declination towards a low F0 than in Swedish, and the pre-pause lengthening is more apparent.

In British English a substantial fall of the F0 contour is frequently found also in non-final prosodic groups. Special rules apply to compound stress.

9. CONCLUSIONS

Prosodic categories have their roots in speech production theory. Unique for our study is the incorporation of subglottal pressure and its co-variation with F0 and intensity. We have employed a multi-parameter recording and display system, with access to auditory assessment of word and syllable prominence, which has been applied to the perceptual grading of acoustic parameters.

Studies of parameter covariation within a speaker's available F0 range have shown systematic differences below and above a mid frequency F0r. A major part of an intonation contour is found in the lower part, whilst the upper part conveys prominence peaks.

Our system of frequency and time normalization could have applications not only in synthesis but also, in the tradition of Gårding (1989), in descriptive studies to sort out individual intonation patterns from a norm. A unique feature of our intonation analysis is the consistent use of a semitone scale with an absolute reference to 100 Hz. It has allowed us to calculate representative average contours for a group of mixed male and female speakers. The sampling system also enables a substantial amount of data reduction.

The FK text-to-speech prosody rules for Swedish have functioned remarkably well. They are more detailed and more complete than the earlier proposed schemes of Carlson and Granström (1973) and of Bruce et al. (2000). We have, in a relatively short time, performed tentative transfers to French and English, but these have to be followed up by more detailed studies.

Our modular tools appear to have a language universal significance, and can be adjusted for language specific needs. This applies to prosodic grouping in terms of modular base curves, as well as of superimposed F0 accent modulation.

Our data are in fair agreement with the language universal rules suggested by Collier (1991). One example is the dimension of "hat patterns", i.e. of the duration and peak height of single prominent F0 peaks, modelled as triangles, to be compared with our smoothly shaped peaks from original F0 recordings, see Figures 27, A,B,C. These have a symmetrical bell shape of about 250 ms length and a height of 5–10 semitones depending on the degree of prominence. This is an example of a basic muscular gesture. Time constants for F0 decrease may be faster than for F0 rise (Fujisaki et al. 2000).

Prosodic grouping and declination patterns also rely on physiological constraints. The tendency we have observed, of the total F0 declination within a major prosodic group to be approximately independent of the duration, agrees with the data of Collier (1991). To the inventory of universal constraints we may add auditory integration, which explains why straight line approximations of F0 contours are sufficient in synthesis.

As an outcome of our synthesis rules we have been able to demonstrate what happens in a switching of language codes. One example is a simulation of a Frenchman reading an English text without a proper insight in the sound inventory and prosody, and also the converse, simulating an Englishman reading a French text. Systematic manipulation of synthesis rules of this type could have applications in second language teaching.

Gunnar Fant and Anita Kruckenberg

REFERENCES

Bickley, C. and Stevens, K. N. (1986). Effects of a vocal tract constriction on the glottal source: Experimental and modelling studies. *Journal of Phonetics* 14, 373–382.

Bruce, G. (1977). *Swedish Word Accents in Sentence Perspective.* Lund: Gleerup.

Bruce, G., Filipsson, M., Frid, J., Granström, B., Gustafson, K., Horne, M. and House, D. (2000). Modelling of Swedish Text and Discourse Intonation in a Speech Synthesis Framework. In Botinis A. (ed.), *Intonation. Analysis Modelling and Technology*. Kluwer Academic Publishers, 291–320.

Carlson, R. and Granström, B. (1973). Word accent, emphatic stress, and syntax in a synthesis-by-rule scheme for Swedish. *STL-QPSR,* 2–3/1973, 31–35.

Campbell, N. and Beckman, M. (1997). Stress, prominence and spectral tilt. ESCA Workshop on Intonation. In: Botinis A, Kouroupetroglou G and Carayannis G (eds.). *Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications, Athens,* 18–20.

Collier, R. (1974). Laryngeal muscle activity, subglottal air pressure, and the control of pitch in speech. *Haskins Laboratories Status Report*, SR-39/40, 137–169.

Collier, R. (1991). Multi-language intonation synthesis. *Journal of Phonetics* 19, 61–73.
Fant, G. (1956). *On the predictability of formant levels and spectrum envelopes from formant frequencies.* For Roman Jakobson, Mouton and Co., 's-Gravenhage, 109–120.
Fant, G. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. *Ericsson Technics* 15/1, 1–106.
Fant, G. (1960). *Acoustic theory of speech production.* Mouton, the Hague.
Fant, G. (1982). Preliminaries to analysis of the human voice source. *STL-QPSR, KTH,* 4/1982, 1–27.
Fant, G. (1993). Some problems in voice source analysis. *Speech Communication* 13, 7–22.
Fant, G. (1995). The LF-model revisited. Transformations and frequency domain analysis. *STL-QPSR, KTH,* 2–3/1995, 119–155.
Fant, G. (1997). The voice source in connected speech. *Speech Communication* 22, 125–139.
Fant, G. (1998). Durationsdata. *Internal report TMH.*
Fant, G. and Ananthapadmanabha, T.V. (1982). Truncation and superposition. *STL-QPSR, KTH*, 2–3/1982, 1–17.
Fant, G. and Gustafson, K. (1995). Röstkällan i GLOVE. Systemanpassning och frekvensdomänmatchning. *TMH-KTH internal report*.
Fant, G., Hertegård, S., Kruckenberg, A. and Liljencrants, J. (1997A). Covariation of subglottal pressure, F0 and glottal parameters. *Eurospeech* 97, 453–456.
Fant, G. and Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR, KTH,* 2/1989, 1–83.
Fant, G. and Kruckenberg A (1994). Notes on stress and word accent in Swedish. *Proceedings of the International Symposium on Prosody, Sept 18 1994, Yokohama.* Also published in *STL-QPSR, KTH,* 2–3/1994, 125–144.
Fant, G. and Kruckenberg, A. (1995). The voice source in prosody. *Proc. ICPhS 95*;2, 622–625.
Fant, G. and Kruckenberg, A. (1996). Voice source properties of the speech code. *TMH-QPSR*, *KTH,* 4/1996, 45–56.
Fant, G. and Kruckenberg, A. (1999A). F0-patterns in text reading. In: Allwood J (ed). *Proc of Fonetik 99, Gothenburg papers in theoretical linguistics*. Gothenburg University, 53–56.
Fant, G. and Kruckenberg, A. (1999B). Syllable and word prominence in Swedish. In: Allwood J (ed). *Proc. of Fonetik 99, Gothenburg papers in theoretical linguistics*. Göteborg University, 57–61.
Fant, G. and Kruckenberg, A. (2000). F0 analysis and prediction in Swedish prose reading. Festschrift for Eli Fischer-Jørgensen. In Grønnum N and Rischel J (eds). *Acta Linguistica Hafniensia*).
Fant, G. and Kruckenberg A. (2000). A Prominence based model of Swedish Intonation. Proc. of ICSLP-2000, Beijing.
Fant, G., Kruckenberg, A. and Nord, L. (1991). Durational correlates of stress in Swedish, French and English. *Journal of Phonetics* 19, 351–365.
Fant, G., Hertegård, S. and Kruckenberg A. (1996). Focal accent and subglottal pressure. *TMH-QPSR, KTH,* 2/1996, 29–32.
Fant, G., Kruckenberg, A. and Liljencrants, J. (1999). Prominence correlates in Swedish prosody. *Proc. of International Conference of Phonetic Sciences, San Francisco* 3:1749–1752.
Fant, G., Kruckenberg, A. and Liljencrants, J. (2000). Acoustic-phonetic analysis of prosody in Swedish. In: Botinis A. (ed). *Intonation. Analysis, Modelling and Technology,* Kluwer, Academic Publishers, 55–86.
Fant, G., Kruckenberg, A. and Liljencrants J. (2000). The source-filter frame of prominence. *Phonetica* 57, 113–127.
Fant, G., Kruckenberg, A., Liljencrants, J., and Hertegård, S. (2000). Acoustic phonetic studies of prominence in Swedish. *TMH-QPSR* 2/3 2000, 1–52.
Fant, G., Kruckenberg, A., Gustafson K. and Liljencrants, J. (2002). A new approach to intonation analysis and synthesis of Swedish, *Speech Prosody 2002, Aix en Provence.* Also in *Fonetik 2002, TMH-QPSR* 2002.
Fant, G., Kruckenberg, A. and Nord, L. (1990). Acoustic correlates of rhythmical structures in text reading. *Nordic Prosody V*, Turku, 1990, 70–86.

Fant, G., Kruckenberg, A. and Nord, L. (1991). Durational correlates of stress in Swedish, French and English. *Journal of Phonetics* 19, 351–365.

Fant, G., Kruckenberg, A. and Barbosa Ferreira, J. (2003).Individual variations in pausing. A study of read speech. *Proc. Fonetik 2003*, *Phonum* 9, Umeå University, 193–196.

Fant, G., Liljencrants, J. and Lin, Q. (1985). A four-parameter model of glottal flow, *STL-QPSR, KTH,* 4/1985, 1–13.

Fant, G. and Lin, Q. (1988). Frequency domain interpretation and derivation of glottal flow parameters, *STL-QPSR, KTH,* 2–3/1988, 1–21.

Fant, G., Hertegård, S., Kruckenberg, A. and Liljencrants, J. (1997). Covariation of subglottal pressure, F0 and glottal prameters. *Eurospeech* 97, 453–456.

Fant, G., Kruckenberg, A., Hertegård, S. and Liljencrants, J. (1997A). Sub- and supraglottal pressures in speech. *Proc. Fonetik 1997*, Umeå University, 25–28.

Fant, G, Kruckenberg, A., Hertegård, S. and Liljencrants, J. (1997B). Accentuation and subglottal pressure in Swedish. In: Botinis A, Kouroupetroglou G and Carayannis G (eds). *Proc of ESCA Workshop on Intonation: Theory, Models and applications, Athens,* 111–114.

Fant, G., Nord, L. and Kruckenberg, A. (1986). Individual Variations in Text Reading. A Data-Bank Pilot Study. *STL-QPSR* 4/1986, 1–17.

Fry, D. (1955). Duration and intensity as physical correlates of linguistic stress. *J. Acoust. Soc. Am.* 27, 765–768.

Fujisaki, H., Ljungqvist, M. and Murata, H. (1993). Analysis and modeling of word accent and sentence intonation in Swedish. *Proc. 1993 Intern. Conf. Acoust. Speech and Signal Processing,* vol. **2**, 211–214.

Fujisaki, H., Tomana, R., Narusawa, S., Ohno, S. and Wang, C. (2000). Physiological mechanisms for fundamental frequency control in standard Chinese. *ICSLP 2000,* 1–4.SS (01), 1–4.

Gårding, E. (1989). Intonation in Swedish. *Working papers*. Lund University Linguistics Department **35**, 63–88. Also in Hirst, Daniel and Alberto Di Cristo (eds) *Intonation Systems. Cambridge University Press* (1998), 112–130.

Hanson, H. (1997A). Glottal characteristics of female speakers. Acoustic correlates, *J Acoust Soc Am*, 101/1, 466–481.

Hanson, H. (1997B). Vowel amplitude variation during sentence production. *Proc of ICASSP-97*, III, 1627–1630.

Hanson, H. and Chuang, E. (1999). Glottal characteristics of male speakers. Acoustic correlates and comparison with female data. *J Acoust Soc Am* 106/2, 1064–1077.

Kruckenberg, A. and Fant, G. (1995). Notes on syllable duration in French and Swedish. *Proc. of ICPhS 95,* II, 158–161.

Jakobson, R., Fant, G. and Halle, M. (1952). *Preliminaries to speech analysis. The distinctive features and their correlates.* Acoustics Laboratory, Massachusetts Inst. of Technology, Technical Report No. 13 (58 pages). Published by MIT press, seventh edition, 1967.

Ladefoged, P. (1967). *Three areas of experimental phonetics*. Oxford University Press.

Liljencrants, J. (1999). Judges of prominence. In: Andersson, R., Abelin, Å., Allwood, J. and Lindblad, P. (eds). *Proc of Fonetik 99, 12th Swedish Phonetics Conference,* June 2–4, 1999, Göteborg, Sweden, 101–107.

Liljencrants, J. Fant, G. and Kruckenberg A. (2000). Subglottal pressure and prosody in Swedish. *Proc. of ICSLP-2000,* Beijing.

Rothenberg, M. (1968). The breath stream dynamics of simple-released plosive production. *Bibliotheca Phonetica* No. 6, Basel, S Karger.

Sluijter, A. and van Heuven, V. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustic Society of America*, 100/4: 2471–2484.

Strik, H. and Boves, L. (1992). On the relation between voice source parameters and prosodic features in connected speech. *Speech Communication* 11/2–3, 167–174.

Stevens, K. N. (1998): *Acoustic Phonetics*. MIT Press, 1998.

Stevens, K. N. and Hanson, M. (1994). Classification of glottal vibrations from acoustic measurements. In: Fujimura O and Hirano M (eds). *Vocal Fold Physiology, Singular Publ Group*, 147–170.

Sundberg. J., Andersson, M. and Hultqvist, C. (1999). Effects of subglottal pressure variations on professional baritone singers' voice sources. *J Acoust Soc Am* 105/3, 1965–1971.

Titze, I. (1989). On the relation between subglottal pressure and fundamental frequency in phonation. *J Acoust Soc Am* 85/2, 901–906.

Öhman, S. (1967). Word and sentence intonation: a quantitative model. *KTH STL-QPSR* 2–3/1967, 20–54.

TMH-QPSR reports are distributed by the department of Speech Music and Hearing, KTH in Stockholm. They are a continuation of the earlier report series, STL-QPSR, issued by the Speech Transmission Laboratory.

*) More references on the subject are to be found in http://www.speech.kth.se/~gunnar/

PUBLICATION LIST 1950–2004

1945

1. Fant, G. (1945). Beräkning av uppfattbarhetsstörningar i ett givet transmissionssystem vid inskränkningar i det överförda frekvensbandet, ett problem vid impulsering med tonfrekvens på interurbanlinje. *Examensarbete i Telegrafi och Telefoni, KTH,* inlämnat 1945-06-13.

1948

1. Fant, G. (1948). Undersökning av 10 sekunders standardfras. *L.M. Ericsson protokoll* H/P 1051.
2. Fant, G. (1948). Analys av de svenska vokalljuden. *L.M. Ericsson protokoll* H/P 1035 (52 pages)

1949

1. Fant, G. (1949). Analys av de svenska konsonantljuden. *L.M. Ericsson protokoll* H/P 1064 (139 pages)
2. Wedenberg, E. and Fant, G. (1949). Auditory training of deaf children. *Acta Oto-Lar.* XXXVII, Fasc. 5, 462–469.

1950

1. Fant, G. (1950). A continuously variable filter. *J. Acoust. Soc. Amer.* **22**, 449–453.
2. Fant, G. (1950). Transmission properties of the vocal tract. Part I, Acoustics Laboratory, *Massachusetts Inst. of Technology, Quarterly Progress Report*, July–September 1950, 20–23.
3. Fant, G. (1950). Transmission properties of the vocal tract. Part II, *Acoustics Laboratory, Massachusetts Inst. of Technology, Quarterly Progress Report*, October–December 1950, 14–19.

1951

1. Fant, G. (1951). Rapport över studier i U.S.A. 1949–1951. *KTH Taltransmissionslaboratoriet* (46 pages).

1952

1. Fant, G. (1952). The heterodyne filter. *Kungl. Tekniska Högskolans Handlingar* Nr 55 (78 pages).

2. Fant, G. (1952). Transmission properties of the vocal tract with application to the acoustic specification of phonemes, *Acoustics Laboratory, Massachusetts Inst. of Technology, Quarterly Progress Report* No. 12
3. Jakobson, R., Fant, G. and Halle, M. (1952). *Preliminaries to speech analysis. The distinctive features and their correlates.* Acoustics Laboratory, Massachusetts Inst. of Technology, Technical Report No. 13 (58 pages). Published by MIT press, seventh edition, 1967.

1953

1. Fant, G. (1953). Speech communication research, *IVA (Royal Swedish Academy of Engineering Sciences, Stockholm)* 2, 331–337.
2. Stevens, K.N., Kasowski, S. and Fant, G. (1953). An electrical analog of the vocal tract. *J. Acoust. Soc. Amer*. 25, 734–742.

1954

1. Fant, G. (1954). Phonetic and phonemic basis for the transcription of Swedish word material. *Acta Oto.Lar.*, Suppl. 116, 24–29.
2. Fant, G. (1954). Fyrpolteori för talorganen. Föredrag vid *RVK-1954*.
3. Fant, G. (1954). Talorganens egenfrekvenser. Föredrag vid *RVK-1954*.
4. Fant, G. (1954). Relativa förekomsten av ord och talljud i svenska språket. Föredrag vid *RVK*-1954.
5. Lidén, G. and Fant, G. (1954). Swedish word material for speech audiometry and articulation tests. *Acta Oto-Lar.* Suppl. 116, 189–210.

1955

1. Fant, G. (1955). Några allmänna synpunkter på system för telefoni över linjer med låg kanalkapacitet. *KTH, Inst. för Telegrafi-Telefoni,* Rapport nr 1, Taltransmissionslaboratoriet (21 pages).
2. Fant, G. (1955). Aktuella system för telefoni på linjer med reducerad kanalkapacitet. *KTH, Inst. för Telegrafi-Telefoni*, Rapport nr 2, Taltransmissionslaboratoriet (12 pages).

1956

1. Fant, G. (1956). On the predictability of formant levels and spectrum envelopes from formant frequencies. *For Roman Jakobson*, Mouton and Co., _s-Gravenhage, 109–120.

1957

1. Fant, G. (1957). *Den akustiska fonetikens grunder*. KTH, Inst. för Telegrafi-Telefoni, Rapport nr 7, Taltransmissionslaboratoriet (61 pages).

1958

1. Fant, G. (1958), *Modern instruments and methods for acoustic studies of speech*, Acta Polytechnica Scandinavica, No 1, 1–81.
2. Fant, G. (1958). On the acoustics of speech. Summary of thesis publications. *KTH* (13 pages).
3. Fant, G. and Möller, A. (1958). Taltransmissionsforskning i England och Holland. *KTH, Inst. för Telegrafi-Telefoni*, Rapport nr 12, Taltransmissionslaboratoriet. (21 pages).
4. Fant, G. and Richter, M. (1958). Some notes on the relative occurrence of letters, phonemes, and words in Swedish. *Proc. of the VIIIth Intl. Congress of Linguists*, Oslo 1958, 815–816.
5. Fant, G. (1958). *Acoustic theory of speech production.* KTH, Inst. för Telegrafi-Telefoni, Rapport nr 10, 1958 Taltransmissionslaboratoriet (319 pages); (Contains a chapter on information theory not published in the Mouton book.

1959

1. Fant, G. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. *Ericsson Technics* **1**–1959, 1–106.
2. Fant, G. (1959). Reserapport USA 21/11–19/12 1958. *KTH, Inst. för Telegrafi-Telefoni, Taltransmissionslaboratoriet* Rapport nr 14, (20 pages)
3. Fant, G. (1959). Speech research at the Royal Institute of Technology, Stockholm. *Seminar on Speech Compression and Processing*, Bedford, Mass., 28–30 Sept. 1959.

1960

1. Fant, G. (1960). The acoustics of speech. In Sir Alexander Ewing, (ed.) *The Modern Educational Treatment of Deafness*. Manchester Univ. Press.
2. Fant, G. (1960). *Acoustic theory of speech production.* The Hague, Netherlands: Mouton, 2nd edition. 1970, (Translated into Russian, *Nauka, Moskva*, 1964).
3. Fant, G. (1960). Third order active filters. *KTH, Inst. för Telegrafi-Telefoni*, P.M. av den 8/10 1960.
4. Fant, G. (1960). Erfarenheter från framställning av syntetiskt tal med resonansanalog. Föredrag vid *RVK-1960.*
5. Fant, G. (1960). Aktuellt om vokodertekniken. *KTH*, Vokoder PM nr 1/1960.
6. Fant, G. (1960). 51-channel analyzer for spectrum sampling. *STL-QPSR* 1/1960, 17–18.
7. Fant, G. (1960). Structural classification of Swedish phonemes. *STL-QPSR* 2/1960, 10–15.
8. Fant, G. and Liljencrants, J. (1960). Enkla filter med aktiva RC och LRC-nät. Föredrag vid *RVK-1960.*
9. Fant, G. and Mártony, J. (1960). Pole-zero matching techniques. *STL-QPSR* 1/1960, 14–16.

10. Fant, G. and Möller, A. (1960). Bryggstabiliserade LRC oscillatorer. Föredrag vid *RVK*-1960.
11. Fant, G. and Stevens, K.N. (1960). Systems for speech compression. In, Fortschritte der Hochfrequenztechnik, Vol. 5. *Akademische Verlagsgesellschaft M.b.H.,* Frankfurt am Main, 229–262.

1961

1. Fant, G. (1961). The acoustics of speech. In, *Proc. of the Third Intl. Congress on Acoustics*, Stuttgart 1959. (L. Cremer, ed.). Amsterdam 1961, 188–201.
2. Fant, G. (1961). Akustisk analys av mänskligt tal. Del I. Mätmetoder. In, *KOSMOS*, Svenska Fysikersamfundets årsbok 1961, 87–89.
3. Fant, G. (1961). Optimum coding in speech transmission links and Human and synthetic speech production. Lectures at *Univ. of Michigan, Ann Arbor, Mich*., USA, May17–18 1961 in World Space Communication.
4. Fant, G. (1961). Formantvokodersystem STL-1, *KTH*, Vokoder PM nr 4/1961.
5. Fant, G. (1961). Formantvokoder av serietyp med frekvensuppdelad nivåstyrning. Patentansökan 2199/61.
6. Fant, G. (1961). Bryggstabilisering av RC- och LC-oscillatorer. *Elektronik*, No. **2**, 58–61.
7. Fant, G. (1961). A new anti-resonance circuit for inverse filtering. *STL-QPSR* 4/1961, 1–6.
8. Fant, G. (1961). Nya metoder för talkommunikation. Teknisk Tidskrift 91, 653–654.
9. Fant, G. and Lindblom, B. (1961). Studies of minimal speech sound units. *STL-QPSR* 2/1961, 1–11.
10. Fant, G. and Mártony, J. (1961). Quantization of formant coded synthetic speech. *STL-QPSR* 2/1962, 16–18.
11. Mártony, J. and Fant, G. (1961). Pole-zero matching of spectra of [l]. *STL-QPSR* 1/1961, 1–2.

1962

1. Fant, G. (1962). Descriptive analysis of the acoustic aspects of speech. Logos (U.S.A.) 5, 3–17.
2. Fant, G. (1962). Plan för formantvokoder system OVE III. *Taltransmissionslaboratoriet, KTH*, intern rapport
3. Fant, G. (1962). Studier av mänskligt tal och syntetiskt tal. *KOSMOS*, Svenska Fysikersamfundets årsbok 1962, 83–96.
4. Fant, G. (1962). Speech analysis and synthesis. *KTH, Inst. för Telegrafi-Telefoni*, Rapport nr 26, Taltransmissionslaboratoriet, 1962, 1–63. (Selected STL-QPSR articles reporting work under grants from US Airforce and US Army).
5. Fant, G. (1962). Sound spectrography. *Proc. of the Fourth Intl. Congress of Phonetic Sciences,* Helsinki 1961, s-Gravenhage, 14–33.

6. Fant, G. (1962). Informationsöverföring inom människan. Föredrag vid extra kurs i bioteknologi vid *KTH*, 1962.
7. Fant, G. (1962). Summary of Speech Communication Seminar in Stockholm 1962 Translated to Russian for the *Acoustical Journal of the Sovjet Academy of Science* part IX, 151–152
8. Fant, G. (1962). Formant bandwidth data. *STL-QPSR* 1/1962, 1–2.
9. Fant, G. and Liljencrants, J. (1962). How to define formant level. A study of the mathematical model of voiced sounds. *STL-QPSR* 2/1962, 1–9.
10. Fant, G. and Mártony, J. (1962). Instrumentation for parametric synthesis (OVE II), synthesis strategy, and quantization of synthesis parameters. *STL-QPSR* 2/1962, 18–24.
11. Fant, G. and Sonesson, B. (1962). Indirect studies of glottal cycles by synchronous inverse filtering and photo-electrical glottography. *STL-QPSR* 4/1962, 1–3.
12. Briess, B. and Fant, G. (1962). Studies on voice pathology by means of inverse filtering. *STL-QPSR* 1/1962, 6.

1963

1. Fant, G. (1963). Vokodersystem. *Patentansökan*.
2. Fant, G. (1963). Kopplingsanordning för grundtonsadaptiv bandbreddsvariatiion av filter vid spektrumanalys. *Patentansökan*.
3. Fant, G. (1963). Antiresonansfilter med passivt RC-nät. Föredrag vid *RVK-1963*.
4. Fant, G. (1963). Research proposal. *KTH, Taltransmissionslaboratoriet.*
5. Fant, G. (1963). The International Speech Communication Seminar. Review. *Soviet Physics Acoustics* (a translation of 'Akusticheskii Zhurnal') 9/2, 113–123.
6. Fant, G. (1963). Abstracts of Papers on Speech Analysis, Stockholm Speech Communication Seminar, 1962, Royal Inst. of Technology, Stockholm, Sweden. *J. Acoust. Soc. Amer*. **35**, 1112–1117.
7. Fant, G., Fintoft, K., Liljencrants, J. Lindblom, B. and Mártony, J. (1963). Formant amplitude measurements, Paper C2, *Proc. of Speech Communication Seminar, Stockholm 1962.*
8. Fant, G., Fintoft, K., Liljencrants, J., Lindblom, B. and Martony, J. (1963). Formant-amplitude measurements. *J. Acoust. Soc. Amer* 35, 1753–1761.
9. Fant, G., Lindblom, B. and Mártony, J. (1963). Spectrograms of Swedish stops. *STL-QPSR* 3/1963, 1.
10. Fant, G. and Mártony, J. (1963). Formant amplitude measurements. *STL-QPSR* 1/1963,1–5.
11. Fant, G., Mártony, J., Rengman, U. and Risberg, A. (1963). OVE II synthesis strategy. Paper F5, *Proc. of Speech Communication Seminar, Stockholm 1962,* Stockholm 1963.
12. Fant, G. and Risberg, A. (1963). Evaluation of speech compression systems. *STL-QPSR* 2/1963, 15–21.
13. Fant, G., Risberg, A. and Mártony, J. (1963). Nuvarande status av tekniken för talkompression. Föredrag vid *RVK-1963*, *5*.

14. Fant, G. and Risberg, A. (1963). Auditory matching of vowels with two formant synthetic sounds. *STL-QPSR* 4/1963, 7–11.
15. Mörner, M., Fransson, F. and Fant, G. (1963). Voice register terminology and standard pitch. *STL-QPSR* 4/1963, 17–23.
16. Tappert, C.C., Mártony, J. and Fant, G. (1963). Spectrum envelopes for synthetic vowels. *STL-QPSR* 3/1963, 2–6.

1964

1. Fant. G. (1964). Formants and cavities. In, *Proc. of the Fifth Intl. Congr. of Phonetic Sciences*, Munster (E. Zwirner and W. Bethge, eds.), Basel, S Karger., 120–141.
2. Fant, G. (1964). Phonetics and speech research. In, *Research Potentials in Voice Physiology*, State Univ. of New York, N.Y. 1964, 199–239.
3. Fant, G. (1964). Transformteori och enkla nät. Kompendium i Teletransmissionsteori, del I, 1962–1963, *KTH* (92 pages).
4. Fant, G. (1964). Signalteori. Kompendium i teletransmissionsteori, 1963–1964, *KTH*(101 pages).
5. Fant, G. (1964). Comments to Professor Mol_s 'The Relation between Phonetics and Phonemics'. *Linguistics* **9**, 29–31.
6. Fant, G. (1964). Auditory patterns of speech. *Models for the Perception of Speech and Visual Form*, Boston, Mass., Nov. 11–14.
7. Fant, G. (1964). Auditory patterns of speech. *STL-QPSR* 3/1964, 16–20.
8. Fant, G. and Liljencrants, J. (1964). Enkla filter med aktiva RC- och LRC-nät. *Elektronik* No. 1, 76–81 och 104.
9. Fant, G. and Mártony, J. (1964). Information bearing aspects of formant amplitude. *Proc. Vth Intl. Congress of Phonetic Sciences,* Münster, Aug. 16–23, 1964.
10. Fant, G. and Sonesson, B. (1964). Speech at high ambient air-pressure. *STL-QPSR* 2/1964, 9–21.
11. Bjuggren, G. and Fant, G. (1964). The nasal cavity structures. *STL-QPSR* 4/1964, 5–7.
12. Tarnóczy, T. and Fant, G. (1964). Some remarks on the average speech spectrum. *STL-QPSR* 4/1964, 13–14.

1966

1. Fant, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. *STL-QPSR* 4/1966, 22–30.
2. Fant, G. (1966). The nature of distinctive features. *STL-QPSR* 4/1966, 1–14.
3. Chistovich, L., Fant, G., de Serpa-Leitao, A. and Tjernlund, P. (1966). Mimicking and perception of synthetic vowels. *STL-QPSR* 2/1966, 1–18, and part II in *STL-QPSR* 3/1966, 1–3.
4. Fant, G., Lindblom, B. and de Serpa-Leitao, A. (1966). Consonant confusions in English and Swedish. A pilot study. *STL-QPSR* 4/1966, 31–34.
5. Fant, G., Ondrackova, J., Lindqvist-Gauffin, J. and Sonesson, B. (1966). Electrical glottography. *STL-QPSR* 4/1966, 15–21.

1967

1. Fant, G. (1967). Sound, features and perception. *STL-QPSR* 2–3/1967, 1–14.
2. Fant, G. (1967). Kompendium i talöverföring, del 1. *KTH, inst. för talöverföring* 1967.

1968

1. Fant, G. (1968). Analysis and synthesis of speech processes. In, *Manual of Phonetics*, Chapt. 8, 173–276 (B. Malmberg, ed.). Amsterdam, North-Holland Publ. Co.
2. Fant, G. and Lindqvist-Gauffin, J. (1968). Pressure and gas mixture effects on divers_ speech. *STL-QPSR* 1/1968, 7–17.
3. Shuplakov, V., Fant, G. and de Serpa-Leitao, A. (1968). Acoustical features of hard and soft Russian consonants in connected speech, A spectrographic study. *STL-QPSR* 4/1968, 1–6.

1969

1. Fant, G. (1969). Stops in CV-syllables. *STL-QPSR* 4/1969, 1–25.
2. Fant, G., Henningsson, G. and Stålhammar, U. (1969). Formant frequencies of Swedish vowels. *STL-QPSR* 4/1969, 26–31.

1970

1. Fant, G. (1970). Automatic recognition and speech research. *STL-QPSR* 1/1970, 16–31.
2. Fant, G., Liljencrants, J., Malá, V. and Borovicková, B. Perceptual evaluation of coarticulation effects. *STL-QPSR* 1/1970. 10–13.
3. Carlson, R., Granström. B. and Fant, G. (1970). Some studies concerning perception of isolated vowels. *STL-QPSR* 2–3 1970, 19–35.
4. Derkach, M., Fant, G. and de Serpa-Leitao, A. (1970). Phoneme coarticulation in Russian hard and soft VCV-utterances with voiceless fricatives. *STL-QPSR* 2–3/1970, 1–7.

1971

1. Fant, G. (1971). Distinctive features and phonetic dimensions. In, *Applications of Linguistics.* (G.E. Perren and J.L.M. Trim, eds.). (Selected papers of the Second Intl. Congr. of Applied Linguistics, Cambridge 1969). Cambridge University Press, Great Britain.
2. Metas, O., Fant, G. and Stålhammar, U. (1971). The A vowels of Parisan French. *STL-QPSR* 4/1971, 1–18.

1972

1. Fant, G. (1972). Vocal tract wall effects, losses, and resonance bandwidths. *STL-QPSR* 2–3/1972, 28–52.
2. Fant, G. (1972). Q-codes. In, *Intl. Symp. on Speech Communication Ability and Profound Deafness, Stockholm 197*0 (A.G. Bell Ass. for the Deaf, Washington, DC., eds.). 261–268.
3. Fant, G., Ishizaka, K. and Lindqvist-Gauffin, J. (1972). Subglottal formants. *STL-QPSR* 1/1972, 1–12.
4. Zemlin, W. and Fant, G. (1972). The effect of a velo-påharyngeal shunt upon vocal tract damping times, An analogue study. *STL-QPSR* 4/1972, 6–10.

1973

1. Fant, G. (1973). *Speech Sounds and Features.* The MIT Press. Cambridge, MA, USA, (contains a selected number of articles).
2. Ericsson, G. Fant, G. and de Serpa-Leitao (1973). Acoustical and perceptual evaluation of speech training in post-operative cleft palate patients. *STL-QPSR* 1/1973, 25–28.
3. Stålhammar, U., Karlsson, I. and Fant, G. (1973). Contextual effects on vowel nuclei. *STL-QPSR* 4/1993, 1–18.

1975

1. Fant, G. Non-uniform vowel normalization. *STL-QPSR* 2–3/1975, 1–19.
2. Fant, G. (1975). Vocal-tract area and length perturbations. *STL-QPSR* 4/1975, 1–14.
3. Fant, G. (1975). Key-note address. In, *Speech Recognition* (R. Reddy, ed.). New York, *Academic Press Inc.*
4. Liljencants, J. and Fant, G. (1975). Computer program for VT-resonance frequency calculations. *STL-QPSR* 4/1975, 15–20.
5. Fant, G. and Pauli, S. (1975). Spatial characteristics of vocal tract resonance modes. In G. Fant, (ed.) *Speech Communication,* Proc. Speech Communication Seminar, Stockholm 1974, Stockholm, Almqvist and Wiksell, Vol 2.
6. Fant, G., Stålhammar, U. and Karlsson, I. (1975). Swedish vowels in speech materials of various complexity. In G. Fant, (ed.) *Speech Communication,* Proc. Speech Communication Seminar, Stockholm 1974, Stockholm, Almqvist and Wiksell. Vol **3**. 139–148.
7. Fant, G., Carlson, R. and Granström, B. (1975). The [e]-[ø] ambiguity. In G. Fant, (ed.) *Speech Communication,* Proc. Speech Communication Seminar, Stockholm 1974, Stockholm, Almqvist and Wiksell. Vol **3**.
8. Carlson, R., Fant, G. and Granström, B. (1975). Two-formant models, pitch, and vowel perception. *Auditory Analysis and Perception of Speech* (G. Fant and M.A.A. Tatham, eds.). London, Academic Press Inc., 55–82.

1976

1. Fant, G. (1976). Key-note speech. *US-Japan Joint Seminar on Dynamic Aspects of Speech Production*, Dec. 7–10, 1976, Tokyo; also publ. in *STL-Q*PSR 4/1976, 21–27.
2. Fant, G., Nord, L. and Branderud, P. (1976). A note on the vocal tract wall impedance. *STL-QPSR* 4/1976, 13–20.
3. Fant, G. (1976). Vocal tract energy functions and non-uniform scaling. *J. Acoust. Soc. Japan 11*, 1976, 1–18.

1977

1. Fant, G. (1977). Introduction to the Symposium on Articulartory Modeling. In *Articulatory Modeling and Phonetics* (R. Carre, R. Descout and M. Wajskop, eds.). Grenoble, GALF. 11–12.
2. Mettas, O. and Fant, G. (1977). Front vowels in Parisian sociolects. *STL-QPSR* 2–3/1977, 1–7.

1978

1. Fant, G. (1978). Vowel perception and specification. *Rivista Italiana di Acustica* **II**, 1978, 69–87.
2. Bladon, A. and Fant, G. (1978). A two-formant model and the cardinal vowels. *STL-QPSR* 1/1978, 1–8.
3. Wakita, H. and Fant, G. (1978). Towards a better vocal tract model. *STL-QPSR* 1/1978, 9–29.

1979

1. Fant, G. (1979). Glottal source and excitation analysis. *STL-QPSR* 1/1979, 85–107.
2. Fant, G. (1979). Vocal source analysis, a progress report. *STL-QPSR* 3–4/1979, 31–53.
3. Fant, G. and Liljencrants, J. (1979). Perception of vowels with truncated intraperiod decay envelopes. *STL-QPSR* 1/1979, 79–84.

1980

1. Fant, G. (1980). The relation between area functions and the acoustical signal. *Phonetica* **37**, 55–86.
2. Fant, G. (1980). Voice source dynamics. *STL-QPSR* 3/1980, 17–37.
3. Fant, G. (1980). Perspectives in speech research. *STL-QPSR* 2–3/1980, 1–16.
4. Fant, G. (1981). Talforskning och handikap. *Kommunikation trots handikapp*, (eds. Kerstin Stigmark och Karin Wengelin), Riksbankens Jubileumsfond. 1980.

1981

1. Fant, G. (1981). The source filter concept in voice production, *STL-QPSR* 1/1981, 21–37.

1982

1. Fant, G. and Ananthapadmanabha, T.V. (1982). Truncation and superposition. *STL-QPSR* 2–3/1982, 1–17.
2. Ananthapadmanabha, T.V. and Fant, G. (1982). Calculation of true glottal flow and its components. *STL-QPSR* 1/198, 1–30; also in *Speech Communication* **1**, (1982), 167–184.
3. Fant, G. (1982). Preliminaries to analysis of the human voice source. *STL-QPSR* 4/1982 1–27.
4. Fant, G. (1982). The voice source-acoustic modeling. *STL-QPSR* 4/1982, 28–48; also in *Abstracts of the Tenth Intl. Congress of Phonetic Sciences*. Dordrecht, Foris Publ., 151–177.

1983

1. Fant, G. (1983). Feature analysis of Swedish vowels–a revisit. *STL-QPSR* 2–3/1983, 1–19.
2. Fant, G. (1983). Foreword. In, *Electronic Speech Synthesis* (B. Watson, ed.). London, Granada Publ. Ltd.
3. Fant, G. (1983). Phonetics and speech technology. *STL-QPSR* 2–3/1983, 20–35; also publ. as a keynote address in *Proc. of the Tenth Intl. Congr. of Phonetic Sciences*, Vol. IIB (M.P.R. v.d. Broecke and A. Cohen, eds.). Dordrecht, Foris Publ., 13–24.

1984

1. Fant, G. (1984). Human speech and communication aids. In, *Proc. of the II Intl. Conf. on Applications of Physics to Medicine and Biology.* Singapore, World Scientific Publ. Co., 3–19.
2. Badin, P. and Fant, G. (1984). Notes on vocal tract computation. *STL-QPSR* 2–3/1984, 53–108.
3. Nord, L., Ananthapadmanabha, T.V. and Fant, G. (1984). Signal analysis and perceptual tests of vowel responses with an interactive source filter model. *STL-QPSR* 2–3/1984, 25–52.

1985

1. Fant, G. (1985). Speech technology – research and development. Presentation at the occasion of the Ericsson Prise 1985, *Ericsson Review*.
2. Fant, G. (1985). The vocal tract in your pocket calculator. In V Fromkin (ed.) *Phonetic Linguistics.* London, Academic Press. 55–77.

3. Fant, G. (1985). Talforskning—Teknik och Vetenskap. I Tal Ljud och Hörsel (eds. Elisabeth Ahlsén. Jens Allwood och Erland Hjelmqvist). Föredrag och abstracts från det andra TLH-symposiet, 14–15 mars 1985 vid *Göteborgs Universitet, Guling* 13, 49–58.
4. Fant, G., Liljencrants, J. and Lin, Q. (1985). A four-parameter model of glottal flow. Paper presented at the *French-Swedish Seminar*, Grenoble, France, April 22–24, 1985.
5. Fant, G., Liljencrants, J., and Lin, Q. (1985). A four-parameter model of glottal flow. *STL-QPSR* 4/1985, 1–13.
6. Fant, G., Lin, Q. and Gobl, C. (1985). Notes on glottal flow interaction. *STL-QPSR* 2–3/1985, 21–45.

1986

1. Fant, G. (1986). Features—fiction and facts. In J. Perkell and D. Klatt, (eds,) *Invariance and Variability of Speech Processes*. Lawrence Erlbaum Ass. Publ. 482–491.
2. Fant, G. (1986). Glottal flow, models and interaction. *Journal of Phonetics*, **4** (3/4) Theme issue, Voice Acoustics and Dysphonia. Gotland, Sweden, August 1985, 393–399.
3. Fant, G., Kruckenberg, A. (1986). Projekt god svenska—databas. *Internrapport Inst. f. Talöverföring och Musikakustik, KTH,* Stockholm.
4. Fant, G., Kruckenberg, A. (1986). Rytm och meter eller rytm i meter. Föredrag vid minikonferens 'Rytm i Tal och Musik', vid *Inst för Talöverföring och Musikakustik, KTH*, 21 april, 1986 (manuscript only)
5. Fant, G., Nord, L. and Kruckenberg, A. (1986). Individual Variations in Text Reading. A Data-Bank Pilot Study. *STL-QPSR* 4/1986, 1–17.

1987

1. Fant, G. (1987). Interactive phenomena in speech production. *Proc XIth Intl. Congr. of Phonetic Sciences (ICPhS),* Tallinn, USSR, Vol. **3**, 376–381.
2. Fant, G. and Lin, Q. (1987). Glottal source—vocal tract acoustic interaction. *STL-QPSR* 1/1987, 13–27.
3. Fant, G., Nord, L. and Kruckenberg, A. (1987). Segmental and Prosodic Variabilities in Connected Speech. An Applied Data-Bank Study. *Proc. XIth Intl. Congr. of Phonetic Sciences, (ICPhS),* Tallinn, USSR, Vol. **6**, 1987, 102–105.

1988

1. Fant, G., Lin, Q. (1988). Frequency domain interpretation and derivation of glottal flow parameters. *STL-QPSR* 2–3/1988, 1–21.
2. Fant, G. and Kruckenberg, A. (1988). Contributions to temporal analysis of read Swedish. *Working Papers 34,* Dept. of Linguistics, Lund University, 1988, 37–41.
3. Fant, G. and Kruckenberg, A. (1988). Some durational correlates of Swedish prosody. *Proc. Seventh FASE Symp*., Vol. 2, *SPEECH-88,* Edinburgh, 495–503.

4. Fant, G., Kruckenberg, A., and Nord, L. (1988). Data-bank analysis of speech prosody. *The 2nd symp. on advanced man-machine interface through spoken language,* Hawaii, November 1988.

1989

1. Fant, G. (1989). Quantal theory and features. *Journal of Phonetics* **17**, 1989, 79–86.*6*
2. Fant, G. (1989). Speech research in perspective. *STL-QPSR* 4/1989, 1–7
3. Fant, G. (1989). The speech code. In C. von Euler, I. Lundberg, G. Lennerstrand. (eds.) *Brain and Reading,* MacMillan, London, 171–182, 1989. (A revised version, "On the speech code" included in *TMH-QPSR* 2/2001, 61–67.)
4. Fant, G. and Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR* 2/1989, 1–83.
5. Fant, G., Kruckenberg, A. and Nord, L. (1989a). Stress patterns, pauses and timing in prose reading. *STL-QPSR* 1/1989, 7–12.
6. Fant, G., Kruckenberg, A. and Nord, L. (1989). Rhythmical structures in text reading. A language contrasting study. *Eurospeech 89*, Vol. 1, 498–501.
7. Fant, G., Kruckenberg, A. and Nord, L. (1989). Acoustic correlates of rhythmical structures in text reading. *Nordic Prosody V*, 23–25 Åbo.
8. Nord, L., Kruckenberg, A. and Fant, G. (1989), Some timing studies of prose, poetry and music, Proc. *Eurospeech 89*, Vol. II, 1989, 690–693.
9. Nord, L., Kruckenberg, A. and Fant, G. (1980), Timing studies of read prose and poetry with parallels in music—a research proposal, *STL-QPSR* 1/1989, 1–6.
10. Badin, P. and Fant, G. (1989). Fricative modeling, Some essentials. *Proc*. *Eurospeech 8*, Paris (J. Tubach and J.J. Mariani, eds.), Vol. II, 23–26.
11. Carlson, R., Fant, G., Gobl, C., Granström, B., Karlsson, I. and Lin, Q. (1989). Voice source rules for text-to-speech synthesis. *ICASSP 1989*, Vol. **1**, 223–226.
12. Lin, Q. and Fant, G. (1989). Vocal tract area function parameters from formant frequencies, In, Proc. of *Eurospeech 89* in Paris (J. Tubach and J.J. Mariani, eds.), Vol. **2**, 673–676.

1990

1. Fant, G. (1990). Speech research in perspective. *Speech Communication* **9**, 1990, 171–176. (Revised 2001 for *Selected Writings*))
2. Fant, G. (1990). The role of speech research in the advance of speech technology. *Speech Workshop*, Tata Institute of Fundamental Research, Bombay, December 10–12, 1990.
3. Fant, G. (1990). The speech code. Segmental and prosodic features. Proc. ICSLP 90, Kobe, Japan, Vol. 2, *J. Acoust.Soc. of Japan*, Tokyo, 1990, 1389–1397.
4. Fant, G. (1990), Människans röst och tal i Forskning i ett föränderligt samhälle, Stiftelsen Riksbankens Jubileumsfond 1965–1990, 328–348. English version, Man's Voice and Speech, in *Swedish Research in a Changing Society*, The Bank of Sweden Tercentenary Foundation 1965–1990, 339–360.

5. Fant, G., Kruckenberg, A. and Nord, L. (1990a). Acoustic correlates of rhythmical structures in text reading. *Nordic Prosody V*, Turku, 1990, 70–86.
6. Fant, G., Kruckenberg, A. and Nord, L. (1990). Segmental durations within a prosodic frame in Swedish, French and English. *Phonum* 1, Contributions to Fonetik 90, the Fourth Swedish Phonetics Conference, Umeå/Lövånger, 42–45.
7. Fant, G., Kruckenberg, A. and Nord, L. (1990). Studies of prosody and segmentals in text reading. *ATR Workshop on Speech Perception and Production, Kyoto, Japan*, November 15–16, 1990.
8. Fant, G., Kruckenberg, A. and Nord, L. (1990). Prosodic and segmental speaker variations. *Proceedings of the Tutorial and Research Workshop on Speaker Characterization in Speech Technology,* Edinburgh, June 26–28, 106–120.
9. Lin, Q., and Fant, G. (1990). A new algorithm for speech synthesis based on vocal tract modeling. *STL-QPSR* 2–3/1990, 45–52.
10. Nord, L., Kruckenberg, A. and Fant, G. (1990). Some aspects of rhythm in prose, poetry and music. *Nordic Prosody V*, Turku, 1990, 256–265.
11. Nord, L., Kruckenberg, A. and Fant, G. (1990), Some timing studies of prose, poetry and music, *Speech Communication* 9, 1990, 477–483.

1991

1. Fant, G. (1991). What can basic research contribute to speech synthesis? *Journal of Phonetics*, **19**, 1991, 75–90.
2. Fant, G. (1991). Units of temporal organization. Stress groups versus syllables and words. *Proc. XIIth ICPhS*, Aix-en-Provence, 1991, 247–250.
3. Fant, G., Kruckenberg, A. and Nord, L. (1991). Prosodic and segmental speaker variations. *Speech Communication* **10**, 1991, 521–531.
4. Fant, G., Kruckenberg, A., and Nord, L. (1991). Durational correlates of stress in Swedish, French and English. *Journal of Phonetics* **19**, 1991, 351–365.
5. Fant, G., Kruckenberg, A. and Nord, L. (1991). Stress patterns and rhythm in the reading of prose and poetry with analogies to music performance In J. Sundberg and L. Nord, R. Carlson, (eds.) *Music, Language, Speech, and Brain*, Wenner-Gren International Symposium (), Series Vol. 59, 1991, 380–407.
6. Fant, G., Kruckenberg, A. and Nord, L. (1991). Temporal organization and rhythm in Swedish. *Proc. of the XIIth ICPhS*, Aix-en-Provence, 1991, 251–256.
7. Fant, G., Kruckenberg, A. and Nord, L. (1991). Some observations on tempo and speaking style in Swedish text reading. *ESCA Workshop on the phonetics and phonology of speaking styles*, Barcelona, September 30–October 2 ,1991.
8. Fant, G., Kruckenberg, A. and Nord, L. (1991). Tempo and stress. In (eds) O. Engstrand, C. Kylander and M. Dufberg, *Perilus XIII, Fifth National Phonetics Conference, Stockholm* 31–34.
9. Fant, G., Kruckenberg, A. and Nord, L. (1991). Language specific patterns of prosodic and segmental structures in Swedish, French and English. *Proceedings of the XIIth ICPhS*, Aix-en-Provence, 1991, 118–121.
10. Fant, G. and Lin, Q. (1991). Comments on glottal flow modelling and analysis. In (J. Gauffin and B. Hammarberg (eds) *Vocal Fold Physiology. Acoustic, Perceptual and*

Physiological Aspects of Voice Mechanisms, San Diego, Singular Publishing Group, Inc., 1991, 47–56.
11. Kruckenberg, A., Fant, G. and Nord, L. (1991a). Från prosa till poesiens rytm och meter-In (eds.) E. Lilja, J. Swedenmark and K. Wåhlin, *Vers-mått. Studier framlagda vid Andra Nordiska Metrikkonferensen*, Uppsala, 1989, 147–162.
12. Kruckenberg, A., Fant, G. and Nord, L. (1991). Rhythmical structures in poetry reading, *Proc. of the XIIth ICPhS*, Aix-en-Provence, 1991, 242–245

1992

1. Fant, G. (1992). Vocal tract area functions of Swedish vowels and a new three-parameter model. *Proc. ICSLP-92,* Vol. 1, 807–810.
2. Fant, G. och Kruckenberg, A (1992a), Prediktion av stavelselängder i svenska. Preliminär intern rapport om DS-projektet
3. Fant, G., Kruckenberg, A. and Nord, L. (1992b). Prediction of syllable duration, speech rate and tempo, *Proc. ICSLP 92,* Banff, Vol **1**. 667–670.
4. Fant, G., Kruckenberg, A. and Nord, L. (1992). Prediction of syllable duration and tempo. *Sjätte svenska fonetiksymposiet*, Chalmers Tekniska Högskola, Göteborg, 20–22 maj 1992.
5. Galyas, K., Fant, G. and Hunnicutt, S. (1992). *Voice output communication aids*. A study sponsored by the International Project on Communication Aids for the Speech Impaired, IPCAS. The Swedish Handicap Institute, Stockholm (86 pages).
6. Lin, Q. and Fant, G. (1992). An articulatory speech synthesizer based on a frequency domain simulation of the vocal tract, *IEEE-ICASSP*, paper 173, San Francisco, March 23–26, 1992.

1993

1. Fant, G., (1993). Some problems in voice source analysis. *Speech Communication* **13**, 7–22.
2. Fant, G. and Kruckenberg, A. (1993). Towards an integrated view of stress correlates, in *Working Papers* **41**, Dept. of Linguistics, Lund University, 1993, 42–46.
3. Fant, G. and Kruckenberg, A. (1993). Towards an integrated view of stress correlates, *KTH-TMH manuscript* for oral presentation at ESCA workshop on Speech Prosody Lund, 1993
4. Fant, G. and Kruckenberg, A. (1993). Akustisk-fonetiska modeller av uppläst prosa. Ansökan till Riksbankens Jubileumsfond 1993.
5. Kruckenberg, A. and Fant, G. (1993). Iambic versus trochaic patterns in poetry reading. *Nordic Prosody* **VI**, Stockholm, 1993, 123–135.

1994

1. Fant, G. (1994). Lectures in China 1985. Translated into Chinese. *Academia Sinica.* (172 pages).
2. Fant, G. and Kruckenberg, A. (1994). Notes on Stress and Word Accent in Swedish, *STL-QPSR* 2–3/1944, 125–144. Also published in *Proc. Int. Symp. on Prosody, 18 Sept 1994, Yokohama*, 19–36.

3. Fant, G. and Kruckenberg, A. (1994). Voice source parameters in connected speech. A progress report, *Working Papers* **43**, Lund University, Dept. of Linguistics, 58–61.
4. Fant, G., Kruckenberg, A., Liljencrants, J. and Båvegård, M. (1994). Voice source parameters in continuous speech. Transformation of LF-parameters, *Proc. ICSLP-94*, Yokohama, Vol. 3, 1451–1454.
5. Fant, G. and Liljencrants, J. (1994). Data reduction of LF voice source parameters, Working Papers 43, Lund University, Dept. of Linguistics, 62–65.
6. Båvegård, M. and Fant, G. (1994). Notes on glottal source interaction ripple. *STL-QPSR* 4/1994, 63–78.
7. Båvegård, M., Fant, G., Gauffin, J. and Liljencrants, J. (1994). Vocal tract sweeptone data and model simulations of vowels, laterals and nasals. *STL-QPSR* 4/1993 43–76.

1995

1. Fant G. (1995). The LF-model revisited. Transformations and frequency domain analysis. *STL-QPSR* 2–3/1995 119–155.
2. Fant, G. (1995). Röstkällan i text-tal-syntes. Underlag för programmering. *Telia Promotor Infovox AB, internal report*.
3. Fant, G. (1995). Speech related to pure tone audiograms., In (eds.) G. Plant and K.E. Spens. *Profound deafness and speech communication.* Whurr Publ. Ltd, London, 299–305.
4. Fant, G. and Båvegård, M. (1995), Parametric model of the vocal tract area function, Vowels and consonants, *ESPRIT/BR SPEECHMAPS (6975)*. Delivery 28, WP2.2, 1–30. Also published in *TMH-QPSR* 1/1997, 1–20.
5. Fant, G. and Gustafson K (1995). Röstkällan i GLOVE. Systemanpassning och frekvensdomänmatchning. *TMH internal report*.
6. Fant G. and Kruckenberg A. (1995). The voice source in prosody. *Proc. ICPhS 95* Vol. **2**, 622–625.
7. Båvegård, M. and Fant, G (1995), From formant frequencies to vocal tract area function parameters, *ESPRIT/BR SPEECHMAPS (6975).* Delivery 29, WP2.3, 1–12.
8. Kruckenberg A. and Fant G. (1995). Notes on syllable duration in French and Swedish. *Proc. ICPhS 95*, Vol II, 158–161.

1996

1. Fant, G. (1996). Text-till-tal regler för källfunktioner och talartyp i formantbaserad talsyntes. *KTH, TMH internal report*.
2. Fant, G. (1996). Historical notes. Response to interview questions posed by Louis-Jean Boe and Pierre Badin. *KTH-TMH manuscript*, 15 pages.
3. Fant G., Hertegård S. and Kruckenberg, A. (1996). Focal accent and subglottal pressure. *TMH-QPSR* 2/1996, 29–32.
4. Fant G. and Kruckenberg A. (1996). Voice source properties of the speech code. *TMH-QPSR* 4/1996, 45–46.
5. Fant G. and Kruckenberg A. (1996). On the quantal nature of speech timing. *Proc. ICSLP-1996*, 2044–2047.

6. Fant G., Kruckenberg A. and Hertegård S. (1996). Focal accent and subglottal pressure. Fonetik 96, *TMH-QPSR* 2/1996, 29–32.

1997

1. Fant G. (1997). The voice source in connected speech. *Speech Communication* **22**, 125–139.
2. Fant G. (1997). Acoustical Analysis of Speech. In (ed.) M.J. Crocker, Encyclopedia of Acoustics, John Wiley, Vol. 4, 1589–1597.
3. Fant G., Hertegård S, Kruckenberg A. and Liljencrants J. (1997).Covariation of subglottal pressure, F0 and glottal parameters. *Eurospeech 97*, 453–456.
4. Fant G., Kruckenberg A., Hertegård S., and Liljencrants J. (1997). Sub- and supraglottal pressures in speech. *Proc. Fonetik 1997*, Umeå University, 25–28.
5. Fant G., Kruckenberg A., Hertegård S. and Liljencrants J. (1997). Accentuation and subglottal pressure in Swedish. In, Botinis A, Kouroupetroglou G and Carayannis G, (eds.) *Proc of ESCA Workshop on Intonation, Theory, Models and applications*, Athens, 111–114.
6. Fant G., Kruckenberg A. and Nord L. (1997). Some studies of accentuation and juncture in Swedish, Fonetik-97, PHONUM, Reports from the department of Phonetics, Umeå University, 157–160.

1998

1. Fant G. (1998) Durationsdata. Internal report KTH, TMH.
2. Fant G. and Kruckenberg A. (1998). Prominence and accentuation. Acoustical correlates. In P. Branderud and Traunmüller H. (eds). Proc of Fonetik -98, The Swedish Phonetics Conference, Stockholm University, 1998, 142–145.

1999

1. Fant G. and Kruckenberg A. (1999a). F0-patterns in text reading. In (ed.), J. Allwood, Proc. *Fonetik 99, Gothenburg papers in theoretical linguistics*, University of Gothenburg, 53–56.
2. Fant G., and Kruckenberg A. (1999b). Syllable and word prominence in Swedish., In (ed,), J. Allwood. Proc. *Fonetik 99, Gothenburg papers in theoretical linguistics*, University of Gothenburg, 57–61.
3. Fant G., and Kruckenberg A. (1999). Prominence correlates in Swedish prosody. *Proc. of International Conference of Phonetic Sciences, 1999*, San Francisco 3, 1749–1752.

2000

1. Fant, G. (2000). Half a century in phonetics and speech research. Swedish phonetics meeting in Skövde, May 24–26, 2000. (Revised version).
2. Fant G. and Kruckenberg A. (2000b) A Prominence based model of Swedish Intonation. Proc. of ICSLP-2000, Beijing.

3. Fant, G. Kruckenberg A. and Liljencrants J. (2000). Acoustic-phonetic analysis of prosody in Swedish. In (ed.) A. Botinis, *Intonation. Analysis, Modelling and Technology*, Kluwer, Academic Publishers, 55–86.
4. Fant, G. Kruckenberg A. and Liljencrants J. (2000). The Source-Filter Frame of Prominence. *Phonetica* **57**, 113–127.
5. Fant, G., Kruckenberg, A., Liljencrants, J., and Hertegård, S. (2000). Acoustic phonetic studies of prominence in Swedish. *TMH-QPSR* 2/3 2000, 1–52
6. Liljencrants, J. Fant, G. and Kruckenberg A. (2000). Subglottal pressure and prosody in Swedish. *Proc. of ICSLP-2000*, Beijing.

2001

1. Fant, G. (2001). Swedish vowels and a new three-parameter model. *TMH-QPSR* 1/2001.
2. Fant, G. (2001). On the Speech Code. TMH-QPSR 2–3 2001, 61–67. (Revised and updated version of an article, The Speech Code, In C. von Euler, I. Lundberg and G. Lennerstrand (eds.) *Brain and Reading*. MacMillan, London, 1982, 171–182.)
3. Fant G. and Kruckenberg A. (2001). F0 analysis and prediction in Swedish prose reading. In N. Grønnum and J. Rischel (eds.), *To honour Eli Fischer-Jørgensen.* Travaux du Circle Linguistique de Copenhague.Copenhagen; Reitzel, 124–147.
4. Fant, G. and Kruckenberg, A. (2001). A novel system for F0 analysis and prediction. *Papers from Fonetik 2001*, Lund University, Department of Linguistics, Working papers 49, 38–41.
5. Fant, G. Kruckenberg, A. Liljencrants J. and Botinis, A. (2001). Prominence correlates. A study of Swedish. *Eurospeech 2001*, Aalborg.

2002

1. Fant, G., Kruckenberg, A., Gustafson K. and Liljencrants, J. (2002). A new approach to intonation analysis and synthesis of Swedish, *Speech Prosody 2002, Aix en Provence*, 283–286. Also in *Fonetik 2002, TMH-QPSR* 2002.

2003

1. Fant, G., Kruckenberg, A. and Barbosa Ferreira, J. (2003). Individual variations in Pausing. A study of read speech. Proceedings from Fonetik 2003, Phonum 9, Umeå University, 193–196.

2004

1. Fant, G. (2004). More than half a century in phonetics and speech research. In Gunnar Fant, *Speech Acoustics and Phonetics*, Kluwer Academic Publishers, pp. 2–14. (Revised version of a presentation at the Swedish phonetics meeting in Skövde, May 24–26, 2000).
2. Fant, G. (2004). Speech research in a historical perspective. In J. Slifka, S. Manuel and M. Matthies (Eds.), From Sound to Sense: 50+ Years of Discoveries in

Speech Communication, Research Laboratory of Electronics MIT, June 11–13, 2004, pp. 20–40.
3. Fant, G. and Kruckenberg, A. (2004). Prosody by rule in Swedish with Language Universal Implications. *Proceedings Prosody 2004*, *Nara*, pp. 405–408.
4. Fant, G. and Kruckenberg, A. (2004). Intonation analysis and synthesis with reference to Swedish. International Symposium on Tonal Aspects of Language, TAL 2004, Beijing, pp. 57–60.
5. Fant, G. and Kruckenberg, A. (2004). Analysis and synthesis of Swedish prosody with outlooks on production and perception. In G. Fant, H. Fujisaki, J. Chao and Y. Xu (Eds.), Festschrift Wu Zongji, From tradional phonology to modern speech processing, pp. 73–95, Foreign Language Teaching and Research Press, Beijing.
6. Fant, G. and Kruckenberg, A. (2004). An integrated view of Swedish prosody. Voice production, perception and synthesis. In Gunnar Fant, *Speech Acoustics and Phonetics*. Kluwer Academic Publishers, pp. 249–300.

REFERENCE CATEGORIES

A. Speech Research Overviews

1959:3, 1961:8, 1962:3,7,8, 1963:4–6, 1964:2, 1965:2, 1968:1, 1970:1, 1973:1, 1976:1, 1977:1, 1980:3, 1983:2,3, 1985:1,3, 1989:2,3, 1990:1,2, 1990:2,4, 1991:1, 1994:1, 1996:2, 2000:1, 2004:1.

B. Speech Production, Models, Synthesis, Data

1950:2,3, 1952:2, 1953:1,2, 1954:2,3, 1956:1, 1958:2,5, 1959:1, 1960:1,2,9, 1961:1, 10,11, 1962:3,4,8,9,10,11, 1963:7,8,10,11,15,16, 1964:1,11, 1966:1, 1967:2, 1972:1,3.4 1973:1, 1975:1,2,4,5, 1976:2,3, 1978:3, 1979:3, 1980:1, 1981:1–3, 1984:2, 1985:2, 1987:1,2, 1989:1,10,12, 1990:9, 1991:1, 1992:1,6, 1994:1,7, 1995:1,4,7, 1996:3,6, 1997:2–5, 2000:4–6, 2001:2.

C. The Voice Source

1962:11,12, 1963:15, 1966:5, 1979:1–3, 1980:2, 1981:1, 1982:1–4, 1984:3, 1985:4–6, 1986:2, 1987:2, 1988:1, 1989:11, 1991:10, 1993:1. 1994,3–6: 1995,1:2,5,6, 1996:1,4, 1997:1,3, 2000:4,5.

D. Speech Perception, Audiology and Speech Aids

1949:2, 1954:1,5, 1963:14 1964:6,7, 1966:3,4, 1970:2,3, 1972:2, 1973:2, 1975:7,8, 1978:1,2, 1979:3, 1980:4, 1984:1,3, 1992:5, 1995:3.

E. Experimental Phonetics, Descriptive Analysis

1948:1,2, 1949:1, 1954:4, 1956:1, 1957:1, 1958:1,4, 1959:1, 1960:1,10, 1961:1,2,9,11, 1962:3–6,8–10, 1963:7,8,9,10, 1964:6,7,9,12, 1966:1, 1967:1, 1968:1,3, 1969:1,2, 1970:4, 1971:2,1973:1,3, 1975:1, 1977:2, 1983:1, 1986:3–5, 1987:3, 1988:2.4, 1989:4–9, 1990:3,5–8,10,11, 1991:1–12, 1992:2–4, 1993:22–5, 1994:2, 1995:8: 1996:5,6, 1997:2–6, 1998:1–2, 1999:1–3, 2000:2–6, 2001:1–3.

F. Feature Theory and the Speech Code

1952:2, 1964:5, 1966:2, 1967:1, 1971:1, 1973:1, 1983:1, 1986:1, 1989:1,3, 1990:3.

G. PROSODY

1986:3–5, 1987:3, 1988:2–4, 1989:4–9, 1990:5–8,10,11, 1991:2–12, 1992:2–4, 1993:2–5, 1994:2, 1995:6,8, 1996:3,5,6, 1997:4–6, 1998:1,2, 1999:1–3, 2000:2–6, 2001:1–5, 2002;1, 2003:1, 2004:3–6.

H. SPEECH CODING AND SYSTEM DESIGN. INSTRUMENTATION

1945:1, 1950:1, 1952:1, 1955:1,2, 1960:3–6,8–11, 1961:1, 3–7,10, 1962:1,2,11, 1963:1–3,12–13, 1964:3,4,8.

I. DIVERS SPEECH

1964:10, 1968:2.

J. LANGUAGE STATISTICS

1958:4, 1967:2, 1972:2.

K. TRAVEL REPORTS

1951:1, 1958:3, 1959:2.

Text, Speech and Language Technology

1. H. Bunt and M. Tomita (eds.): *Recent Advances in Parsing Technology.* 1996 ISBN 0-7923-4152-X
2. S. Young and G. Bloothooft (eds.): *Corpus-Based Methods in Language and Speech Processing.* 1997 ISBN 0-7923-4463-4
3. T. Dutoit: *An Introduction to Text-to-Speech Synthesis.* 1997 ISBN 0-7923-4498-7
4. L. Lebart, A. Salem and L. Berry: *Exploring Textual Data*. 1998 ISBN 0-7923-4840-0
5. J. Carson-Berndsen, *Time Map Phonology*. 1998 ISBN 0-7923-4883-4
6. P. Saint-Dizier (ed.): *Predicative Forms in Natural Language and in Lexical Knowledge Bases.* 1999 ISBN 0-7923-5499-0
7. T. Strzalkowski (ed.): *Natural Language Information Retrieval.* 1999 ISBN 0-7923-5685-3
8. J. Harrington and S. Cassiday: *Techniques in Speech Acoustics.* 1999 ISBN 0-7923-5731-0
9. H. van Halteren (ed.): *Syntactic Wordclass Tagging.* 1999 ISBN 0-7923-5896-1
10. E. Viegas (ed.): *Breadth and Depth of Semantic Lexicons.* 1999 ISBN 0-7923-6039-7
11. S. Armstrong, K. Church, P. Isabelle, S. Nanzi, E. Tzoukermann and D. Yarowsky (eds.): *Natural Language Processing Using Very Large Corpora.* 1999 ISBN 0-7923-6055-9
12. F. Van Eynde and D. Gibbon (eds.): *Lexicon Development for Speech and Language Processing*. 2000 ISBN 0-7923-6368-X; Pb: 07923-6369-8
13. J. Véronis (ed.): *Parallel Text Processing*. Alignment and Use of Translation Corpora. 2000 ISBN 0-7923-6546-1
14. M. Horne (ed.): *Prosody: Theory and Experiment*. Studies Presented to Gösta Bruce. 2000 ISBN 0-7923-6579-8
15. A. Botinis (ed.): *Intonation*. Analysis, Modelling and Technology. 2000 ISBN 0-7923-6605-0
16. H. Bunt and A. Nijholt (eds.): *Advances in Probabilistic and Other Parsing Technologies*. 2000 ISBN 0-7923-6616-6
17. J.-C. Junqua and G. van Noord (eds.): *Robustness in Languages and Speech Technology*. 2001 ISBN 0-7923-6790-1
18. R.H. Baayen: *Word Frequency Distributions*. 2001 ISBN 0-7923-7017-1
19. B. Granström, D. House and. I. Karlsson (eds.): *Multimodality in Language and Speech Systems*. 2002 ISBN 1-4020-0635-7
20. M. Carl and A. Way (eds.): *Recent Advances in Example-Based Machine Translation*. 2003 ISBN 1-4020-1400-7; Pb 1-4020-1401-5
21. A. Abeillé: *Treebanks.* Building and Using Parsed Corpora. 2003 ISBN 1-4020-1334-5; Pb 1-4020-1335-3
22. J. van Kuppevelt and R.W. Smith (ed.): *Current and New Directions in Discourse and Dialogue.* 2003 ISBN 1-4020-1614-X; Pb 1-4020-1615-8
23. H. Bunt, J. Carroll and G. Satta (eds.): *New Developments in Parsing Technology.* 2004 ISBN 1-4020-2293-X; Pb 1-4020-2294-8

24. G. Fant: *Speech Acoustics and Phonetics.* Selected Writings. 2004
ISBN 1-4020-2373-1; Pb 1-4020-2789-3
25. W.J. Barry and W.A. Van Dommelen (eds.): *The Integration of Phonetic Knowledge in Speech Technology.* 2004 ISBN 1-4020-2635-8; Pb 1-4020-2636-6
26. D. Dahl (ed.): *Practical Spoken Dialog Systems.* 2004
ISBN 1-4020-2674-9; Pb 1-4020-2675-7
27. O. Stock and M. Zancanaro (eds.): *Multimodal Intelligent Information Presentation.* 2004 ISBN 1-4020-3049-5; Pb 1-4020-3050-9
28. W. Minker, D. Bühler and L. Dybkjaer (eds.): *Spoken Multimodal Human-Computer Dialogue in Mobile Environments.* 2004 ISBN 1-4020-3073-8; Pb 1-4020-3074-6

KLUWER ACADEMIC PUBLISHERS – DORDRECHT / BOSTON / LONDON