PERILUS
No. XVII

Experiments in speech processes

Department of Linguistics
Stockholm University
Published in December 1993

PERILUS XVII

Experiments in speech processes

Department of Linguistics
Stockholm University
Published in December 1993

This issue of PERILUS was edited by Mats Dufberg and Olle Engstrand. **PERILUS** — **P**honetic **E**xperimental **R**esearch, **I**nstitute of **L**inguistics, **U**niversity of **S**tockholm — mainly contains reports on current experimental work carried out in the phonetics laboratory. Copies are available from Department of Linguistics, Stockholm University, S-106 91 Stockholm, Sweden.

Department of Linguistics
Stockholm University
S-106 91 Stockholm
Sweden

Telephone: 08-16 23 47
(+46 8 16 23 47, international)

Telefax: 08-15 53 89
(+46 8 15 53 89, international)

Telex/Teletex: 8105199 Univers

ISSN 0282-6690

Contents

The phonetics laboratory group

Ann-Marie Almé
Göran Aurelius[1]
Robert Bannert[2]
Jeanette Blomquist
Peter Branderud
Una Cunningham-Andersson
Hassan Djamshidpey
Mats Dufberg
Arvo Eek[3]
Susanne Eisman
Ahmed Elgendi
Olle Engstrand
Gärda Ericsson[4]
Anders Eriksson
Pétur Helgason
Eva Holmberg[5]
Bo Kassling
Diana Krull
Amalia Khachaturian[6]
Catharina Kylander
Francisco Lacerda
Ingrid Landberg
Björn Lindblom[7]
Rolf Lindgren
James Lubker[8]
Bertil Lyberg[9]
Robert McAllister
Lennart Nord[10]
Liselotte Roug-Hellichius
Johan Stark
Johan Sundberg[11]
Ulla Sundberg
Gunilla Thunberg
Hartmut Traunmüller
Karen Williams
Eva Öberg

---

1 Also S:t Görans Children's Hospital, Stockholm.
2 Also Institute of Linguistics, Department of Phonetics, University of Umeå.
3 Visiting from the Institute for Language and Litterature, Estonian Academy of Sciences, Tallinn, Estonia.
4 Also Department of Phoniatrics, University Hospital, Linköping.
5 Also Massachusetts Eye and Ear Infirmary, Boston, MA, USA.
6 Visiting from the Institute of Linguistics, Armenian Academy of Sciences, Yerevan, Armenia.
7 Also Department of Linguistics, University of Texas at Austin, Austin, Texas, USA.
8 Also Department of Communication Science and Disorders, University of Vermont, Burlington, Vermont, USA.
9 Also Swedish Telecom, Stockholm.
10 Also Department of Speech Communication and Music Acoustics, Royal Institute of Technology (KTH), Stockholm.
11 Also Department of Speech Communication and Music Acoustics, Royal Institute of Technology (KTH), Stockholm.

Current projects and grants

Articulatory-acoustic correlations in coarticulatory processes: a cross-language investigation

Supported by: Swedish National Board for Industrial and Technical Development (NUTEK), grant to Olle Engstrand; ESPRIT: Basic Research Action, AI and Cognitive Science: Speech

Project group: Peter Branderud, Olle Engstrand, Bo Kassling, and Robert McAllister

Speech transforms — an acoustic data base and computational rules for Swedish phonetics and phonology

Supported by: Swedish National Board for Industrial and Technical Development (NUTEK) and the Swedish Council for Research in the Humanities and Social Sciences (HSFR), grant to Olle Engstrand.

Project group: Susanne Eisman, Olle Engstrand, Björn Lindblom, Rolf Lindgren, and Johan Stark

APEX: Experimental and computational studies of speech production

Supported by: The Swedish Council for Research in the Humanities and Social Sciences (HSFR), grant to Björn Lindblom.

Project group: Diana Krull, Björn Lindblom, Johan Sundberg, and Johan Stark

Paralinguistic variation in speech and its treatment in speech technology

Supported by: The Swedish Council for Research in the Humanities and Social Sciences (HSFR), grant to Hartmut Traunmüller

Project group: Anders Eriksson and Hartmut Traunmüller

Typological studies of phonetic systems

Supported by: The Swedish Council for Research in the Humanities and Social Sciences (HSFR), grant to Björn Lindblom.

Project group: Olle Engstrand, Diana Krull, Björn Lindblom, and Johan Stark

Second language production and comprehension: Experimental phonetic studies

Supported by: The Swedish Council for Research in the Humanities and Social Sciences (HSFR), grant to Robert McAllister
Project group: Mats Dufberg and Robert McAllister

Sociodialectal perception from an immigrant perspective

Supported by: The Swedish Council for Research in the Humanities and Social Sciences (HSFR), grant to Olle Engstrand.
Project group: Una Cunningham-Andersson and Olle Engstrand

An ontogentic study of infants' perception of speech

Supported by: The Tercentenary Foundation of the Bank of Sweden (RJ), grant to Francisco Lacerda
Project group: Francisco Lacerda, Björn Lindblom, Ulla Sundberg, and Göran Aurelius

Early language-specific phonetic development: Experimental studies of children from 6 to 30 months

Supported by: The Swedish Council for Research in the Humanities and Social Sciences (HSFR), grant to Olle Engstrand
Project group: Jeanette Blomquist, Olle Engstrand, Bo Kassling, Johan Stark and Karen Williams

Speech after glossectomy

Supported by: The Swedish Cancer Society, grant to Olle Engstrand
Project group: Olle Engstrand and Eva Öberg

Previous issues of Perilus

PERILUS I, 1978–1979

PERILUS II, 1979–1980

PERILUS III, 1982–1983

PERILUS IV, 1984–1985

Introduction *Björn Lindblom*

Labial coarticulation in stutterers and normal speakers *Ann-Marie Almé and Robert McAllister*

Movetrack *Peter Branderud*

Some evidence on rhythmic patterns of spoken French *Danielle Dueza and Yukihoro Nishinuma*

On the relation between the acoustic properties of Swedish voiced stops and their perceptual processing *Diana Krull*

Descriptive acoustic studies for the synthesis of spoken Swedish *Francisco Lacerda*

Frequency discrimination as a function of stimulus onset characteristics *Francisco Lacerda*

Speaker–listener interaction and phonetic variation *Björn Lindblom and Rolf Lindgren*

Articulatory targeting and perceptual consistency of loud speech *Richard Schulman*

The role of the fundamental and the higher formants in the perception of speaker size, vocal effort, and vowel openness *Hartmut Traunmüller*

PERILUS V, 1986–1987

About the computer-lab *Peter Branderud*

Adaptive variability and absolute constancy in speech signals: two themes in the quest for phonetic invariance *Björn Lindblom*

Articulatory dynamics of loud and normal speech *Richard Schulman*

An experiment on the cues to the identification of fricatives *Hartmut Traunmüller and Diana Krull*

Second formant locus patterns as a measure of consonant–vowel coarticulation *Diana Krull*

Exploring discourse intonation in Swedish *Madeleine Wulffson*

Why two labialization strategies in Setswana? *Mats Dufberg*

Phonetic development in early infancy – a study of four Swedish children during the first 18 months of life *Liselotte Roug, Ingrid Landberg and Lars Johan Lundberg*

A simple computerized response collection system *Johan Stark and Mats Dufberg*

Experiments with technical aids in pronunciation teaching *Robert McAllister, Mats Dufberg and Maria Wallius*

PERILUS VI, Fall 1987 (Ph.D. thesis)

Effects of peripheral auditory adaptation on the discrimination of speech sounds *Francisco Lacerda*

PERILUS VII, May 1988 (Ph.D. thesis)

Acoustic properties as predictors of perceptual responses: a study of Swedish voiced stops *Diana Krull*

PERILUS VIII, December 1988

Some remarks on the origin of the "phonetic code" *Björn Lindblom*

Formant undershoot in clear and citation form speech *Björn Lindblom and Seung-Jae Moon*

On the systematicity of phonetic variation in spontaneous speech *Olle Engstrand and Diana Krull*

Discontinuous variation in spontaneous speech *Olle Engstrand and Diana Krull*

Paralinguistic variation and invariance in the characteristic frequencies of vowels *Hartmut Traunmüller*

Analytical expressions for the tonotopic sensory scale *Hartmut Traunmüller*

Attitudes to immigrant Swedish – A literature review and preparatory experiments *Una Cunningham-Andersson and Olle Engstrand*

Representing pitch accent in Swedish *Leslie M. Bailey*

PERILUS IX, February 1989

Speech after cleft palate treatment – analysis of a 10-year material *Gärda Ericsson and Birgitta Ystrom*

Some attempts to measure speech comprehension *Robert McAllister and Mats Dufberg*

Speech after glossectomy: phonetic considerations and some preliminary results *Ann-Marie Almé and Olle Engstrand*

PERILUS X, December 1989

F0 correlates of tonal word accents in spontaneous speech: range and systematicity of variation *Olle Engstrand*

Phonetic features of the acute and grave word accents: data from spontaneous speech. *Olle Engstrand*

A note on hidden factors in vowel perception experiments *Hartmut Traunmüller*

Paralinguistic speech signal transformations *Hartmut Traunmüller, Peter Branderud and Aina Bigestans*

Perceived strength and identity of foreign accent in Swedish *Una Cunningham-Andersson and Olle Engstrand*

Second formant locus patterns and consonant–vowel coarticulation in spontaneous speech *Diana Krull*

Second formant locus – nucleus patterns in spontaneous speech: some preliminary results on French *Danielle Duez*

Towards an electropalatographic specification of consonant articulation in Swedish. *Olle Engstrand*

An acoustic-perceptual study of Swedish vowels produced by a subtotally glossectomized speaker *Ann-Marie Almé, Eva Öberg and Olle Engstrand*

PERILUS XI, May 1990

In what sense is speech quantal? *Björn Lindblom and Olle Engstrand*

The status of phonetic gestures *Björn Lindblom*

On the notion of "Possible Speech Sound" *Björn Lindblom*

Models of phonetic variation and selection *Björn Lindblom*

Phonetic content in phonology *Björn Lindblom*

PERILUS XII, May 1991

On the communicative process: Speaker–listener interaction and the development of speech *Björn Lindblom*

Conversational maxims and principles of language planning *Hartmut Traunmüller*

Quantity perception in Swedish [VC]-sequences: word length and speech rate. *Hartmut Traunmüller and Aina Bigestans*

Perceptual foreign accent: L2 user's comprehension ability *Robert McAllister*

Sociolectal sensitivity in native, non-native and non speakers of Swedish – a pilot study *Una Cunningham-Andersson*

Perceptual evaluation of speech following subtotal and partial glossectomy *Ann-Marie Almé*

VOT in spontaneous speech and in citation form words *Diana Krull*

Some evidence on second formant locus-nucleus patterns in spontaneous speech in French *Daniell Duez*

Vowel production in isolated words and in connected speech: an investigation of the linguo-mandibular subsystem *Edda Farnetani and Alice Faber*

Jaw position in English and Swedish VCVs *Patricia A. Keating, Björn Lindblom, James Lubker, and Jody Kreiman*

Perception of CV-utterances by young infants: pilot study using the High-Amplitude-Sucking technique *Francisco Lacerda*

Child adjusted speech *Ulla Sundberg*

Acquisition of the Swedish tonal word accent contrast *Olle Engstrand, Karen Williams, and Sven Strömquist*

PERILUS XIII, May 1991
(Papers from the Fifth National Phonetics Conference, Stockholm, May 29–31, 1991)

Initial consonants and phonation types in Shanghai *Jan-Olof Svantesson*

Acoustic features of creaky and breathy voice in Udehe *Galina Radchenko*

Voice quality variations for female speech synthesis *Inger Karlsson*

Effects of inventory size on the distribution of vowels in the formant space: preliminary data from seven languages *Olle Engstrand and Diana Krull*

The phonetics of pronouns *Raquel Willerman and Björn Lindblom*

Perceptual aspects of an intonation model *Eva Gårding*

Tempo and stress *Gunnar Fant, Anita Kruckenberg, and Lennart Nord*

On prosodic phrasing in Swedish *Gösta Bruce, Björn Granström, Kjell Gustafson and David House*

Phonetic characteristics of professional news reading *Eva Strangert*

Studies of some phonetic characteristics of speech on stage *Gunilla Thunberg*

The prosody of Norwegian news broadcasts *Kjell Gustafson*

Accentual prominence in French: read and spontaneous speech *Paul Touati*

Stability of some Estonian duration relations *Diana Krull*

Variation of speaker and speaking style in text-to-speech systems *Björn Granström and Lennart Nord*

Child adjusted speech: remarks on the Swedish tonal word accent *Ulla Sundberg*

Motivated deictic forms in early language acquisition *Sarah Williams*

Cluster production at grammatical boundaries by Swedish children: some preliminary observations *Peter Czigler*

Infant speech perception studies *Francisco Lacerda*

Reading and writing processes in children with Down syndrome – a research project *Iréne Johansson*

Velum and epiglottis behaviour during production of Arabic pharyngeals: a fibroscopic study *Ahmed Elgendi*

Analysing gestures from X-ray motion films of speech *Sidney Wood*

Some cross language aspects of co-articulation *Robert McAllister and Olle Engstrand*

Articulation inter-timing variation in speech: modelling in a recognition system *Mats Blomberg*

The context sensitivity of the perceptual interaction between F0 and F1 *Hartmut Traunmüller*

On the relative accessibility of units and representations in speech perception *Kari Suomi*

The QAR comprehension test: a progress report on test comparisons *Mats Dufberg and Robert McAllister*

Phoneme recognition using multi-level perceptrons *Kjell Elenius och G. Takács*

Statistical inferencing of text–phonemics correspondences *Bob Damper*

Phonetic and phonological levels in the speech of the deaf *Anne-Marie Öster*

Signal analysis and speech perception in normal and hearing-impaired listeners *Annica Hovmark*

Speech perception abilities of patients using cochlear implants, vibrotactile aids and hearing aids *Eva Agelfors and Arne Risberg*

On hearing impairments, cochlear implants and the perception of mood in speech *David House*

Touching voices – a comparison between the hand, the tactilator and the vibrator as tactile aids *Gunilla Öhngren*

Acoustic analysis of dysarthria associated with multiple sclerosis – a preliminary note *Lena Hartelius and Lennart Nord*

Compensatory strategies in speech following glossectomy *Eva Öberg*

Flow and pressure registrations of alaryngeal speech *Lennart Nord, Britta Hammarberg, and Elisabet Lundström*

PERILUS XIV, December 1991
(Papers from the symposium Current phonetic research paradigm: Implications for speech motor control, Stockholm, August 13–16, 1991)

Does increasing representational complexity lead to more speech variability? *Christian Abry and Tahar Lallouache*

Some cross language aspects of co-articulation *Robert McAllister and Olle Engstrand*

Coarticulation and reduction in consonants: comparing isolated words and continuous speech *Edda Farnetani*

Trading relations between tongue-body raising and lip rounding in production of the vowel /u/ *Joseph S. Perkell, Mario A. Svirsky, Melanie L. Matthies and Michael I. Jordan*

Tongue-jaw interactions in lingual consonants *B Kühnert, C Ledl, P Hoole and H G Tillmann*

Discrete and continuos modes in speech motor control *Anders Löfqvist and Vincent L. Gracco*

Paths and trajectories in orofacial motion *D.J. Ostry, K.G. Munhall, J.R. Flanagan and A.S Bregman*

Articulatory control in stop consonant clusters *Daniel Recasens, Jordi Fontdevila and Maria Dolors Pallares*

Dynamics of intergestural timing *E. Saltzman, B. Kay, P. Rubin and J. Kinsella-Shaw*

Modelling the speaker–listener interaction in a quantitative model for speech motor control: a framework and some preliminary results *Rafael Laboissiere, Jean-Luc Schwartz and Gérard Bailly*

Neural network modelling of speech motor control using physiological data *Eric Vatikiotis-Bateson, Makoto Hirayama and Mitsuo Kawato*

Movement paths: different phonetic contexts and different speaking styles *Celia Scully, Esther Grabe-Georges and Pierre Badin*

Speech production. From acoustic tubes to the central representation *René Carré and Mohamed Mrayati*

On articulatory and acoustic variabilities: implications for speech motor control *Shinji Maeda*

Speech perception based on acoustic landmarks: implications for speech production *Kenneth N. Stevens*

An investigation of locus equations as a source of relational invariance for stop place categorization *Harvey M. Sussman*

A first report on consonant underarticulation in spontaneous speech in French *Danielle Duez*

Temporal variability and the speed of time's flow *Gerald D. Lame*

Prosodic segmentation of recorded speech *W.N. Campbell*

Rhythmical – in what sense? Some preliminary considerations *Lennart Nord*

Focus and phonological reduction *Linda Shockey*

Recovery of "deleted" schwa *Sharon Y. Manuel*

Invariant auditory patterns in speech processing: an explanation for normalization *Natalie Waterson*

Function and limits of the F1:F0 covariation in speech *Hartmut Traunmüller*

Psychoacoustic complementarity and the dynamics of speech perception and production *Keith R. Kluender*

How the listener can deduce the speaker's intended pronunciation *John J. Ohala*

Phonetic covariation as auditory enhancement: the case of the [+voice]/[–vocie] distinction *Randy L. Diehl and John Kingston*

Cognitive-auditory constraints on articulatory reduction *Klaus J. Kohler*

Words are produced in order to be perceived: the listener in the speaker's mind *Sieb G. Nooteboom*

An acoustic and perceptual study of undershoot in clear and citation-form speech *Seung-Jae Moon*

Phonetics of baby talk speech: implications for infant speech perception *Barbara Davis*

Use of the sound space in early speech *Peter F. MacNeilage*

The emergence of phonological organization *M.M. Vihman and L. Roug-Hellichius*

In defense of the Motor Theory *Ignatius G. Mattingly*

Learning to talk *Michael Studdert-Kennedy*

PERILUS XV, December 1992

Use of place and manner dimension in the SUPERB UPSID database: Some patterns of in(ter)dependence *Björn Lindblom, Diana Krull and Johan Stark*

Comparing vowel formant data cross-linguistically *Diana Krull and Björn Lindblom*

Temporal and tonal correlates to quantity in Estonian *Diana Krull*

Some evidence that perceptual factors shape assimilations *Susan Hura, Björn Lindblom and Randy Diehl*

Focus and phonological reduction *Linda Shockey, Kristyan Spelman Miller and Sarah Newson*

The Phonetics of sign language; an outline of a project (paper in Swedish, summary in English) *Catharina Kylander*

The role of the jaw in constriction adjustments during pharyngeal and pharyngealized articulation *Ahmed M. Elgendy*

Young infants prefer high/low vowel contrasts *Francisco Lacerda*

Young infant's discrimination of confusable speech signals *Francisco Lacerda*

Dependence of high-amplitude sucking discrimination results on the pre- and post-shift window duration *Francisco Lacerda*

Prototypical vowel information in baby talk *Barbara Davis and Björn Lindblom*

PERILUS XVI, May 1993 (Ph.D. thesis)

Aerodynamic measurements of normal voice *Eva Holmberg*

Phonetic Experimental Research,
Institute of Linguistics,
University of Stockholm (PERILUS),
No. XVII, 1993, pp. 1–34

F_0-excursions in speech and their perceptual evaluation as evidenced in liveliness estimations[1]

Hartmut Traunmüller and Anders Eriksson

Abstract

Published data on F_0 in speech show its range of variation to be the same for men and women if expressed in semitones. An analysis of additional production data shows that the "liveliness" of speech is related to the extent of the excursions of F_0 from its "base-value". In order to learn how listeners evaluate F_0-excursions, a set of experiments was performed in which subjects had to estimate the liveliness of utterances. The stimuli were obtained by LPC-analysis of one natural utterance that was modified by resynthesizing F_0, the formant frequencies and the time scale in order to simulate some of the natural extra- and paralinguistic variations that affect F_0 and/or liveliness: The speaker's age, sex, articulation rate, and voice register. In each case, the extent of the F_0-excursions was varied in 7 steps. The results showed that, as long as no variation in voice register was involved, listeners judged F_0-intervals to be equal if they were equal in semitones. If the voice register was shifted without adjustment in articulation, listeners appeared to judge the F_0-excursions in relation to to the spectral space available below F_1. The liveliness ratings were found to be strongly dependent on articulation rate and they were observed to be affected by the perceived age of the speaker.

1. Introduction

1.1 F_0-excursions in speech production

There is a substantial amount of data on the frequency of the voice fundamental (F_0) in the speech of speakers who differ in age and sex. Such data have been published for several languages and for various types of discourse. The data reported include nearly always the average F_0, usually expressed in Hz, and less often the average period. Most studies also report on the between-speaker spread in average F_0. Somewhat smaller, but still quite large is the number of studies which, in addition, report on the F_0-range used by each speaker or by the average

---

1) Also submitted to *Journal of the Acoustic Society of America*.

speaker. Unfortunately, the statistics of F_0-values is often not very well described by a normal distribution. If F_0 is scaled linearly (in Hz), there is, typically, some skewness towards higher values and if scaled logarithmically (in semitones), the skewness is in the opposite direction. Analysis of the duration of periods reveals an even stronger skewness (Mikeev, 1971). In addition, it has been observed that some speakers show a bimodal F_0-distribution, in particular when speaking with increased vocal effort, as in a parliamentary debate (Rappaport, 1959). In order to compare the results from studies in which different ways of describing the F_0-variation have been chosen, we are forced to assume normality. We will, however, not include any reports for which this assumption appears to involve a risk of introducing a substantial error. The results of some of the remaining studies are summarized in Table I. The table includes only those investigations in which both male and female adult speakers performed the same kind of task.

The original reports summarized in Table I contain data on average F_0 and on the average standard deviation (SD) of F_0 per speaker reported in Hz, in semitones, or as a frequency modulation factor (SD/mean) in %. In some cases, the range was reported in terms of two SD in semitones. In all except one of the reports (Rose, 1991), women's average F_0 was clearly higher and F_0-range clearly wider as compared with men if expressed in Hz. The between-sex difference more or less disappears for F_0-range if it is expressed in semitones or as a modulation factor.

The very high values for average F_0 observed in male speakers of Wú dialects of Chinese (Rose, 1991) are quite remarkable. They show that even the average F_0 used in speech belongs to the set of properties that can be prescribed by social convention. Although these Chinese dialects present an extreme case, the phenomenon is not unique. An increased average F_0 can also be observed in the Swedish dialect spoken in Småland (Elert and Hammarberg, 1991). In most languages, however, the F_0-range used by speakers appears to be given by physiological factors. Speakers tend to use the lower part of their physiological F_0-range. Thus, the lowest F_0 a speaker uses in ordinary speech is approximately the same as the lowest F_0 at which he is capable of maintaining phonation. In voice range profiles (phonetograms) that show the lowest and the highest F_0 at which a speaker is capable of sustaining phonation as a function of sound pressure level (SPL), F_{0min} can often be seen to rise with SPL (Pabon and Plomp, 1988), and in unrestrained speech F_0 has also been observed to increase with an increase in vocal effort (Ladefoged, 1967). An increase in muscular tenus caused by emotional factorscan also lead to an increase in F_{0min}.

As for the extent of F_0-excursions, it is known that these are influenced by conventional linguistic factors reflected in the language and text in question and by various paralinguistic factors. In linguistic terms, the extent of the F_0-excursions

Table I. Mean value of F_0 in Hz and average F_0-variation (SD) in semitones according to ten investigations that report results from adult male and female speakers in the same setting. Under 'Type', the speech samples are classified according to their expected liveliness, as explained in text.

Investigation	Type	n	Sex	Age	F_0	SD
Rappaport (1958), German	1	190	m		129	2.3
	1	108	f		238	1.9
Chevrie-Muller *et al.* (1967), French	2	21	m	20–61	145	2.5
	2	21	f	19–72	226	2.3
Takefuta *et al.* (1972), English	4	24	m		127	3.8
	4	24	f		186	5.4
Chen (1974), Mandarin Chinese	2	2	m	30–50	108	4.1
	2	2	f	30–50	184	3.8
Bo *et al.* (1975), French	2	30	m		118	2.8
	2	30	f		207	3.0
Kitzing (1979), Swedish	2	51	m	21–70	110	3.0
	2	141	f	21–70	193	2.7
Johns-Lewis (1986), English:						
Conversation	2	5	m	24–49	101	3.4
	2	5	f	24–49	182	2.7
Reading	3	5	m	24–49	128	4.35
	3	5	f	24–49	213	4.5
Acting	4	5	m	24–49	142	4.85
	4	5	f	24–49	239	5.3
Graddol (1986), English:						
Reading passage A	3	12	m	25–40	119	3.6
	3	15	f	25–40	207	3.05
Reading passage B	3	12	m	25–40	131	4.55
	3	15	f	25–40	219	3.9
Pegoraro Krook (1988), Swedish	2	198	m	20–79	113	2.65
	2	467	f	20–89	188	2.55
Rose (1991), W	2	4	m	25–62	170	4.1
	2	3	f	30–64	187	3.8
Average per investigation[#)]		11	m		124	3.4
		11	f		211	3.4
Average per balanced speaker[#)]		471	m		119	2.8
		471	f		207	2.7

[#)] European languages only

in an utterance can be referred to as its "prosodic explicitness". In paralinguistic terms, they can be said to be reflected basically in the "degree of liveliness" or "vivacity" of the speech sample.

Locally, the explicitness of the prosody within an utterance is affected by the placement of focal and contrastive stress. More globally, the extent of F_0-excursions is affected by attitudinal and emotional factors. Emotionally depressed, sad or ashamed speakers produce speech with very little variation in F_0, while increased variation in F_0 reflects an excited emotional state in the speaker, such as surprise, interest, and joy, but also contempt and anger (Fairbanks and Pronovost, 1939; Fónagy and Magdics, 1963; Williams and Stevens, 1972; Scherer, 1974; Bezooyen, 1984). Increased F_0-excursions can also be observed in speech directed to infants (Garnica, 1977). In this case, the increased F_0-excursions appear to serve the purpose of evoking and maintaining a positively excited emotional state in the *listener*.

As for the linguistic factor, we would expect F_0-excursions to be more frequent and probably also larger in tone languages than in languages that do not use tone for segmental distinctions. This has been confirmed in a comparison of Northern Chinese and English (Chen, 1974), where it is also shown that speakers of English with Chinese as a second language use more extensive F_0-excursions in their Chinese than in their English, but that native speakers of Chinese use still more extensive F_0-excursions.

As for the contribution of the type of text, when reading aloud, it has been shown that it does influence the SD of F_0 to a significant degree (Graddol, 1986), but the effects on SD of variations in the type of discourse such as "conversation" compared with "acting" are larger (Johns-Lewis, 1986).

Based on the descriptions of the various types of speech material which resulted in the data summarized in Table I, we have estimated the degree of liveliness that might be expected in the type of discourse used in each case. This has been done by assigning one of four liveliness classes to each type of discourse. The business conversations by telephone, analysed by Rappaport (1958), we have put into the lowest liveliness class. The second class contains somewhat more personal conversations and such tasks as reading a text for the purpose of clinical investigation of one's voice. The third class contains cases where texts have been read aloud in such a way that it can be assumed that the subjects attempted to read in a pleasant way. Into the highest class we have put Johns-Lewis' "acting" and the investigation by Takefuta (1972), who had asked his subjects to vary their intonation pattern as much as they could when repeatedly producing a set of given sentences of the kind that can easily be loaded with various paralinguistic meanings.

For each liveliness class we have calculated the average SD (in semitones) keeping the tone languages apart from the rest. The result is shown in Table II. Although the liveliness classification is somewhat arbitrary, the table can be said to illustrate the following three points:

1) The SD of F_0 increases with increasing "liveliness" of the discourse.

2) The SD of F_0 is larger in tone languages than in non-tone languages.

3) In the most lively types of context, women show a larger SD of F_0 than men, while their SD tends to be lower than that of men in the least lively types of context. This conclusion presupposes that the SD is scaled in semitones or as a modulation factor.

If it is the case that the lowest F_0 frequency speakers use in an utterance is given by the floor of their physiological F_0-range and they increase the extent of their F_0-excursions with increasing liveliness of the discourse, then the average F_0 will increase with increasing SD. This is confirmed by the data of Johns-Lewis (1986) and Graddol (1986), listed in Table I.

In the present investigation we wanted to simulate variations in liveliness. In order to do this without affecting other paralinguistic variables, we needed to know how the expansion of the F_0-excursions is performed when a speaker increases his liveliness, *ceteris paribus*. While the data by Johns-Lewis (1986) and Graddol (1986) are suggestive of an answer, they must be interpreted with some caution since the texts used in the different types of discourse were not the same. There is, however, an investigation by Bruce (1982) in which an actress was asked to produce sentences first with a detached and then with an involved attitude. In this study, the F_0-values of the local minima and maxima of the F_0-contour were reported. Fig. 1 shows, for each minimum and maximum, the excess of the F_0-value in the involved

Table II. Average F_0-variation (SD in semitones) as a function of the type of speech as classified in Table I, sexes pooled. For each investigation in which the SD was higher for women than for men, a "+" sign is shown. In contrary cases, a "–" sign has been entered.

	European lang.		Chinese lang.	
Liveliness class	**SD**	**N**	**SD**	**N**
(4) Very high	4.8	+ +		
(3) High	4.0	+ – –		
(2) Moderate	2.8	– + – – –	4.0	– –
(1) Low	2.1	–		

version over that of the corresponding point on the F_0-contour of the detached version (in semitones) as a function of the F_0-value in the detached version. The regression line in Fig. 1 describes these data fairly well. The F_0-value corresponding to the point where the regression line crosses the horizontal zero-line is the invariant we are looking for. We are going to refer to it as F_b, the "base-value" of F_0. If the F_0-distribution is normal, the frequency position of the base-value F_b can be calculated as

$$F_b = F_{mean} - k \cdot \sigma(F) \quad (1)$$

Since this is valid for any value of σ, it is possible to obtain an estimate of F_b even on the basis of one single utterance, given that k is known. Although in Fig. 1 a logarithmic scaling of pitch has been chosen, the choice of scale is actually not very crucial in this case. Linear regression lines fit the data equally well if a linear (Hz), tonotopic (bark), equivalent rectangular bandwidth (ERB), or logarithmic (semitones) scale of pitch is used.

Fig. 2 shows the F_0-data for each of 5 male and 5 female speakers in three types of discourse: Conversation, reading aloud, and acting. These are the data obtained by Johns-Lewis (1986). The majority of the speakers, 3 male and 4 female, showed a uniform behaviour: Average F_0 and F_0-range (SD) have the smallest values in conversation; both values are higher in reading aloud, and highest in acting. Except for the between-speaker differences in mean F_0, none of these speakers deviated much from the average shown by the dashed line. The remaining 3 speakers, 2 male and 1 female, showed, at some point, a change in F_0 without change in F_0-range. This is likely to be due to a change in vocal effort instead of F_0-variation. The other 7 speakers appear to have adapted only their F_0-variation to the type of discourse. As distinct from the case shown in Fig. 1, the choice of scaling is crucial here. Due to the between-speaker variation in average F_0, Fig. 2 would look different if F_0 had not been scaled in semitones and our conclusion that the majority of speakers behaved in a uniform way would retain its validity only in a qualitative sense.

On the basis of the line that shows the average of the 7 uniformly behaving speakers in Fig. 2 it is possible to calculate the value of k in Equ. 1. We obtain $k = 1.5$ for this case. The data shown in Fig. 1 do not allow a precise calculation of k since σ is not known precisely, but a reasonable estimate would be $1.6 < k < 2.0$. A value of k can also be calculated on the basis of Graddol's data (1986), which include a comparatively large number of speakers, 12 male and 15 female, but the difference in the extent of the F_0-excursions between the two types of discourse is not so large, and therefore the data are somewhat obscured by statistical noise. We obtain $k = 1.7$ for male and $k = 1.1$ for female speakers. Although the variation in

Graddol's data is not primarily due to variation in liveliness, it is not unreasonable to assume that speakers manipulate their F_0-range approximately in the same way as long as no change in vocal effort, voice register, or emotional tension is involved. Given these restrictions, the F_b of a speaker can, as a rule of thumb, be expected to be about 1.5 σ below his average F_0 in any type of discourse. If F_0-values have a normal distribution, F_0 will be higher than F_b 93 % of the time.

1.2 The perception of F_0-excursions.

Although a lot of research has been done on the psychoacoustics of pitch perception, pitch perception in music, and on the linguistic functions of F_0, so far we know very little about the perceptual evaluation of F_0-excursions in speech. Brown *et al.* (1974) investigated the effect of F_0-manipulations on perceived personality features, the main components being the "benevolence" and "competence" attributed

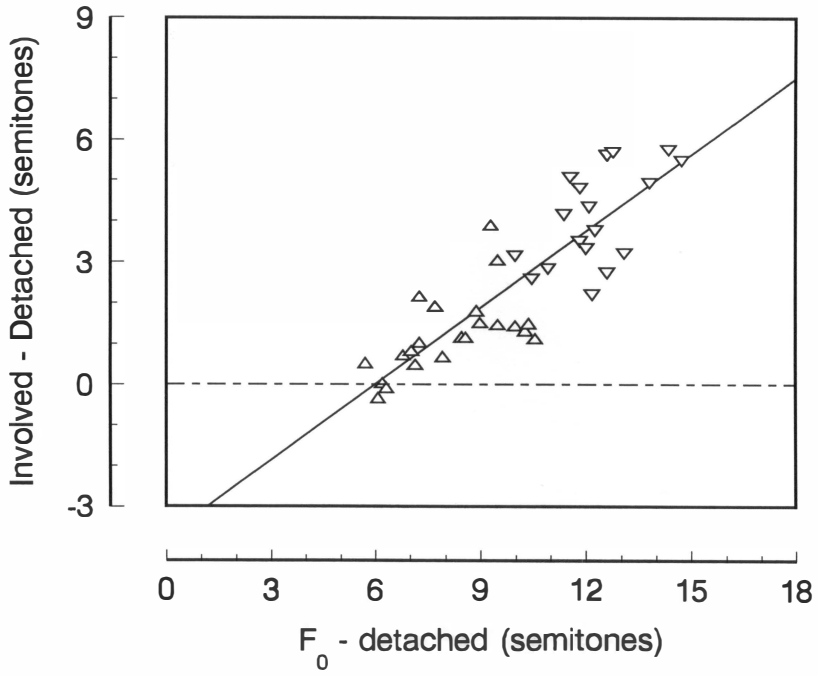


Figure 1. Local maxima and minima in the F_0-contour of four utterances produced with a detached and an involved attitude by a female speaker of Swedish. Mean values from six repetitions. F_0-excess in involved version plotted against F_0-values in detached version. Regression line also shown (r = 0.86). Data from Bruce (1982).

to the speaker, but we are only aware of one previous study, by Hermes and van Gestel (1991), in which the perceptual equivalence of F_0-excursions in speech was investigated by means of well-controlled experiments. Hermes and van Gestel (1991) let their subjects adjust the size of F_0-excursions in resynthesized speech signals. The subjects had to match the perceptual prominence of the syllable marked by the excursion with that of the corresponding syllable in a fixed comparison stimulus produced in a different register with a similar F_0-contour. The results showed that the listeners judged the F_0-excursions to be approximately equivalent when they had the same size expressed in ERB, considering the lowest harmonic alone.

If the result obtained by Hermes and van Gestel (1991) were to hold in general, and given the data listed in Table I, the speech of women should be heard as more

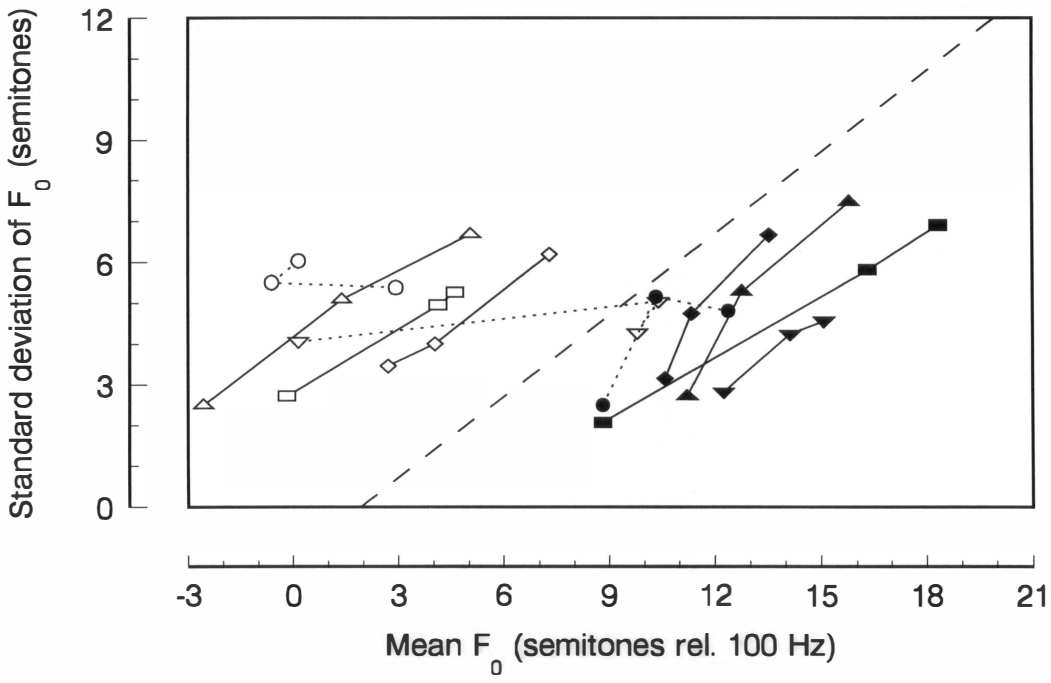


Figure 2. F_0 data of 5 male and 5 female speakers (open and filled symbols) in three types of discourse: Conversation, reading aloud, and acting, connected by lines in this order. Data from Johns-Lewis (1986). Regression line (dashed) fitted to the average of the 7 subjects who behaved in a similar way.

lively than that of men. Although the impressionistic view that this might be the case has been expressed by some observers, this impression is not shared by all (Henton, 1989). If, instead, F0-excursions are judged to be equivalent if their size is the same in semitones, then the data in Table I tell us that in a conversation about any topic whose intrinsic liveliness is low, women should be judged to speak slightly less lively than men, while they should be judged to speeak more lively than men in more lively types of discourse.

In this context, it should be noted that a logarithmic scaling of pitch relieves us — both as listeners and as researchers — from the problem of deciding which partial we should consider. If expressed in semitones or as a modulation factor, the excursions of all the partials are the same — if expressed in mel, ERB, bark, or Hz, they are all different.

2 Methods

2.1 Stimuli

All the stimuli used in the three perceptual experiments to be reported were transformations of the same original sentence. A similar method was used by Brown *et al.* (1974). The transformations served to modify the extent of the excursions of F_0 from F_b. In addition, the speaker's virtual age, sex, articulation rate, and voice register were modified. As for these additional types of variation, Exp. 1 was mainly concerned with age and sex, Exp. 2 with speech rate, and Exp. 3 with voice register.

The original sentence had been recorded previously for the purpose of developing the technique of simulating extra- and paralinguistic variations by means of LPC-analysis and resynthesis after recalculation of the parameter values describing the speech signal (Traunmüller *et al.*, 1989). The sentence "Det finns folkstammar som äter både kattkött och hundkött", perhaps to be translated as 'There are ethnic groups who eat both *chat* and *chien*' or 'There are tribes who eat both cat and dog', was produced by a female speaker, 28 years of age, sitting in a booth with sound-absorbing walls. The utterance was recorded using a Sennheiser MD221U microphone and a Revox PR99 tape recorder, running at 7 1/2 ips. The recorded speech signal was low-pass filtered at 6.3 kHz and digitized with a sampling frequency of 16 kHz and 16 bit/sample. The digitized speech signal was fed into a computer, an Apollo workstation, and subjected to LPC-analysis. Before analysis, the speech file was high-pass filtered in order to remove some low-frequency background noise. The limiting frequency was 140 Hz, which was lower than the lowest observed F_0-value. The LPC analysis was done using a preemphasis coefficient of 0.92 and a Hamming window with a total length of 20 ms, moving forward in steps of 5 ms. The analysis was performed with 15 reflection coeffi-

cients, assuming 7 formant peaks. The description of the speech signal thus obtained was then used as the basis of various transformations.

The parameter values descriptive of the speech signal were recalculated to simulate four different types of speaker; two adults, one male and one female, and two children with an intended age of approximately 5 and 9 years. The parameters affected by the recalculations were F_0, the formant frequencies, and speech rate. The Q-values of the formants were kept at their original values.
The values of F_0 were recalculated according to the equation

$$f' = k_b\,[160 + k_e\,(f - 160)] \tag{2}$$

where f' is the recalculated value of F_0 for a given analysis frame, f is its original value, k_e is the 'excursion factor' by which the deviation of F_0 from F_b was multiplied (k_e = 1.00 for the versions in which the F_0-modulation factor was the same as that in the original version), and k_b is the 'base-value factor' that describes the relation between the values of F_b in the stimuli that differ as to virtual age, sex, and voice register (k_b = 1.00 for the adult female version in the modal register and also for the adult male falsetto version).

The mean F_0 of the original utterance was 215 Hz with an SD of 38.4 Hz (3.0 semitones). After inspection of the F_0-contour of the original utterance, shown in Fig. 3, and based on the analysis of the data obtained by Bruce (1982), Johns-Lewis (1986) and Graddol (1986), as detailed in the Introduction, we assumed a base-value of 160 Hz (the numerical constant in Equ. 1), which is 1.43 σ below the mean, calculated in Hz.

The values of the excursion factor k_e were chosen to cover a large range of variation in liveliness, from completely monotonous up to the upper limit of naturalness. The degrees of variation were distributed between those two extremes in 7 steps, as listed in Table III. The values chosen for k_b are listed in Table IV. The latter table also contains the mean values of F_0 in the mean liveliness (k_e = 1.00) versions of the utterance for the different types of speaker and register.

In order to simulate the adult male speaker and the two children, the formant frequencies were transformed in accordance with the power-function approach described in Traunmüller (1988). Following this approach, the modified formant frequencies F_n' are obtained in accordance with the general equation

$$F_n' = k\,F_n^{\,p} \tag{3}$$

where F_n is the original frequency position of any formant (index n), while k and p are constants descriptive of the transformation in question. Since k and p in Equ.

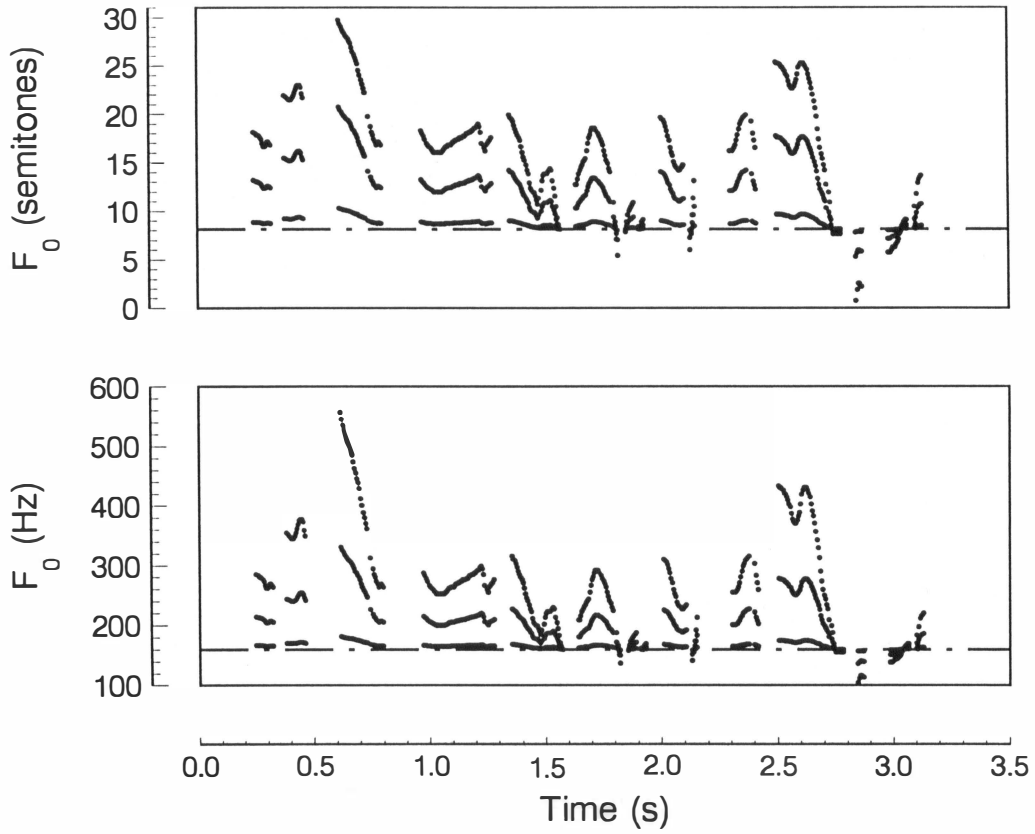


Figure 3. F_0-contours of the utterances with k_b = 1.00 and with the k_e = 0.125, 1.000, and 2.315. The base line at F_b = 160 Hz is also shown.

Table III. Mean and SD of F_0 in the adult female modal register versions shown for each of the 8 different values of the F_0-excursion factor k_e that were used in the experiments (k_e = 0.00 occured only exceptionally).

k_e	0.000	0.125	0.354	0.650	1.000	1.398	1.837	2.315
Mean F_0	160.0	166.8	179.4	195.6	214.8	236.6	260.6	286.8
Std Dev (Hz)	0.0	4.8	13.6	25.0	38.4	53.7	70.6	88.9
Std Dev (st)	0.00	0.44	1.29	2.15	3.02	3.85	4.63	5.37

2 are rather abstract quantities, the computer program written for the purpose of parameter recalculation has been formulated in such a way that it does not require the specification of k and p. Instead, it requires two transformation factors k_{300} and k_{3000} to be specified. These factors are descriptive of the frequency modification to be effected at 300 Hz and at 3000 Hz. While the meaning of these factors is immediately clear, the abstractness is moved into the corresponding reformulation of Equ. 2:

$$F_n' = 300\ k_{300}\ (F_n/300)^p \qquad (4)$$

with

$$p = 1 + \log\ (k_{3000} / k_{300})$$

The factors k_{300} and k_{3000} are also listed in Table IV. These values were based on data on the formant frequencies of Japanese vowels produced by kindergarten children (age 4 to 5 years), girls 12 to 14 years of age, adult women, and adult men (Fujisaki *et al.*, 1970). The factors chosen for the 9 year old child were obtained by interpolation between the data on kindergarten children and those on 12 to 14 year old girls. The previous experimentation with speech signal transformations (Traun-

Table IV. The factors used to recalculate F_0 and the formant frequencies in order to simulate speakers who differed in sex, age, speech rate, and voice register.

	k_b	F_0	k_{300}	k_{3000}	k_r	SF
Female, normal	1.00	215	1.00	1.00	1.000	16,000
Female, slow	1.00	215	1.00	1.00	0.820	16,000
Female, low register	0.56	120	1.00	1.00	1.000	16,000
Female, high register	1.44	309	1.00	1.00	1.000	16,000
Male, normal	0.56	120	0.85	0.80	1.000	12,474
Male, slow	0.56	120	0.85	0.80	0.820	12,474
Male, fast	0.56	120	0.85	0.80	1.220	12,474
Male, high register	1.00	215	0.85	0.80	1.000	12,474
9-year old	1.17	251	1.42	1.09	0.935	15,962
5-year old	1.32	283	1.75	1.18	0.820	15,582
5-year old, slow	1.32	283	1.75	1.18	0.672	15,582
5-year old, fast	1.32	283	1.75	1.18	1.000	15,582

müller *et al.*, 1989) had shown that speech signals transformed using these factors not only for vowels but for the whole utterance possess a fairly high degree of naturalness and the phonetic quality of both vowels and consonants appears to be conserved, given that F_0 is also transformed in an appropriate way.

For the two simulated children, the speech rate was reduced by a factor k_r, also listed in Table IV. The values chosen were based on results obtained by Haselager *et al.* (1991) with Dutch children in the age groups 5, 7, 9, and 11 years and on the additional assumption that at 12 years speech rate attains the value that is typical for adults.

If the transformation is to be performed in one step, the method used requires that the folding frequency (half of the sampling frequency) be transformed according to the same rule as applied for the formant frequencies. Therefore, the resynthesized versions have a sampling frequency that may be different from 16 kHz, as listed in Table IV. The modification of the formant frequencies affects the overall slope of the spectrum of the speech signal. Since we did not have data on the slope of the spectrum in children's speech, it was kept the same as in the original utterance. As for the male versions of the utterance, the slope of their spectrum, integrated over the whole utterance, deviated only marginally from that of the female original, so that no correction was required. The spectral slope of the uncorrected child versions showed an emphasis of the higher frequencies. This was corrected by low-pass filtering. For the 9-year old, a first order low-pass filter with a limiting frequency of 700 Hz was used, while for the 5-year old, this was achieved with two first order low-pass filters with limiting frequencies of 4000 and 315 Hz. Further, the average value of the rms-amplitude of all the stimuli was equalized before recording them on tape.

The transformations in voice register, used in Exp. 3, were not primarily intended to be simulations of a natural variation. The aim with these stimuli was to investigate what happens perceptually if F_0 is changed without adjustment in articulation, thus when the formant frequencies are left unchanged, as in the experiments by Hermes and van Gestel (1991). This is, then, similar to a change in voice register, although in natural shifts in register, we have reason to believe that speakers are also likely to readjust their articulation to obtain a higher F_1 when F_0 is increased, as observed by Maurer *et al.* (1991).

2.2 Subjects

Altogether 55 adults with no known hearing impairment served as subjects in the three perceptual experiments. The subjects were undergraduate students at the University of Stockholm and staff members at the department of linguistics.

Participation was voluntary and unpaid. No subject participated in more than one experiment.

2.3 Procedure

The experiments were run in a quiet lecture room and the stimuli were presented via headphones (AKG K 25) at a comfortable loudness level. The subjects had to note their responses on answer sheets. It was not possible, for practical reasons, to run all subjects in each experiment on one occasion. In order to ensure that the instructions given were identical for all subjects, the instructions were recorded and played as the first item on a tape that also contained all the stimuli. The instruction was immediately followed by an exercise consisting of 8 stimulus pairs. The ratings of those stimuli have not been used in the analyses. After the exercise, the tape was stopped to give the subjects an opportunity to ask for further clarifications.

The main part of all three experiments consisted in a set of magnitude estimation tasks using pairwise comparison. In each pair the standard was presented before the comparison, with a gap of 500 ms in between. A pause with a duration of 5 seconds was inserted between successive pairs to allow time for written responses.

The subjects were asked to assign a number to the comparison stimulus expressing its perceived liveliness. They were instructed to use the number 100 for stimuli whose liveliness they perceived to be equal to that of the standard and to use 50 and 200 for stimuli perceived as 'half as lively' and 'twice as lively', respectively. The subjects were further encouraged to use any more precise number they considered suitable to express the liveliness of a stimulus. The concept of 'liveliness' was not further explained. If asked for, it was only pointed out that an utterance heard as monotonous is likely to receive a very low liveliness rating.

A less copious final part of Exp. 1 and 2 consisted of presentations of single stimuli, representing the neutral stimuli (k_e = 1.00) for each of the speakers simulated in the main part of the experiments. In this part, the subjects had to judge the sex and to rate the age of the speakers.

3. Experiment 1: The effect of virtual sex and age on the perception of liveliness

3.1 Subjects

Eighteen listeners, 7 male and 11 female, served as subjects in this experiment.

3.2 Stimuli and procedure.

The types of speech used in this experiment were the following: Adult female, adult male, 5-year old child, and 9-year old child, with characteristics as listed in Table IV. To test for a possible effect of speech rate, a fifth set of stimuli was included.

These stimuli were identical to the female versions except for speech rate which was the same as that of the 5-year old child ($k_r = 0.82$). For each type of speaker, there were seven versions with different extent of the F_0-excursions (k_e).
The stimuli were presented in four groups, separated by pauses. Each group was introduced by an alerting signal, a soft sounding 'bell'. Within each group, the stimulus pairs were presented in random order.

Group 1 consisted of 8 stimulus pairs. In this group, the female version with $k_e = 1.00$ was used as the standard and all comparison stimuli were also female.

Group 2, also consisting of 8 pairs, had a male standard with $k_e = 1.00$ and male comparisons. (These groups each included one stimulus with a constant F_0. The responses to that monotonous stimulus have, however, been excluded from the following evaluation.)

Group 3 consisted of 35 stimulus pairs with the female standard and 7 stimuli with different k_e for each of the five above mentioned types of speech.

Group 4 consisted of the five versions with $k_e = 1.00$, each presented alone for the purpose of judging the sex and the age of the speakers. In addition, the two child versions were also presented as they were prior to the adjustment of their overall spectral slope.

3.3 Results and discussion

Before pooling the results, the responses of the individual listeners were subjected to multiple regression analysis. This analysis showed the answers from one of the subjects to lack a significant correlation with any of the variables k_e, k_b, k_b k_e and k_r, which distinguish the different stimuli. The responses of this subject were excluded from further analysis since they would have added nothing but noise.

The pooled results from the remaining 17 subjects are presented in Fig. 4 in which, for each stimulus, the average liveliness rating is plotted against the SD of F_0 expressed in Hz and in semitones. It is immediately clear from these diagrams that a linear scale of frequency (in Hz) is not appropriate to describe the responses of the subjects. Consider, *e.g.*, that the 5-year old's utterance with an F_0-variation of 118 Hz was given approximately the same (actually a slightly lower) liveliness rating as the man's utterance with an F_0 variation of only 50 Hz. The semitone scale, on the other hand, seems to fit the data rather well. On this scale, the two utterances have the same F_0-variation, 5.4 semitones. As distinct from Fig. 4a, in Fig. 4b there is no fanning of the lines which describe liveliness as a function of F_0-variation. Allowing for some noise in the data, the slopes of all the different lines in Fig. 4b can be said to be the same. This means that if expressed in semitones, a given increase in F_0-variation leads to a constant increase in perceived liveliness.

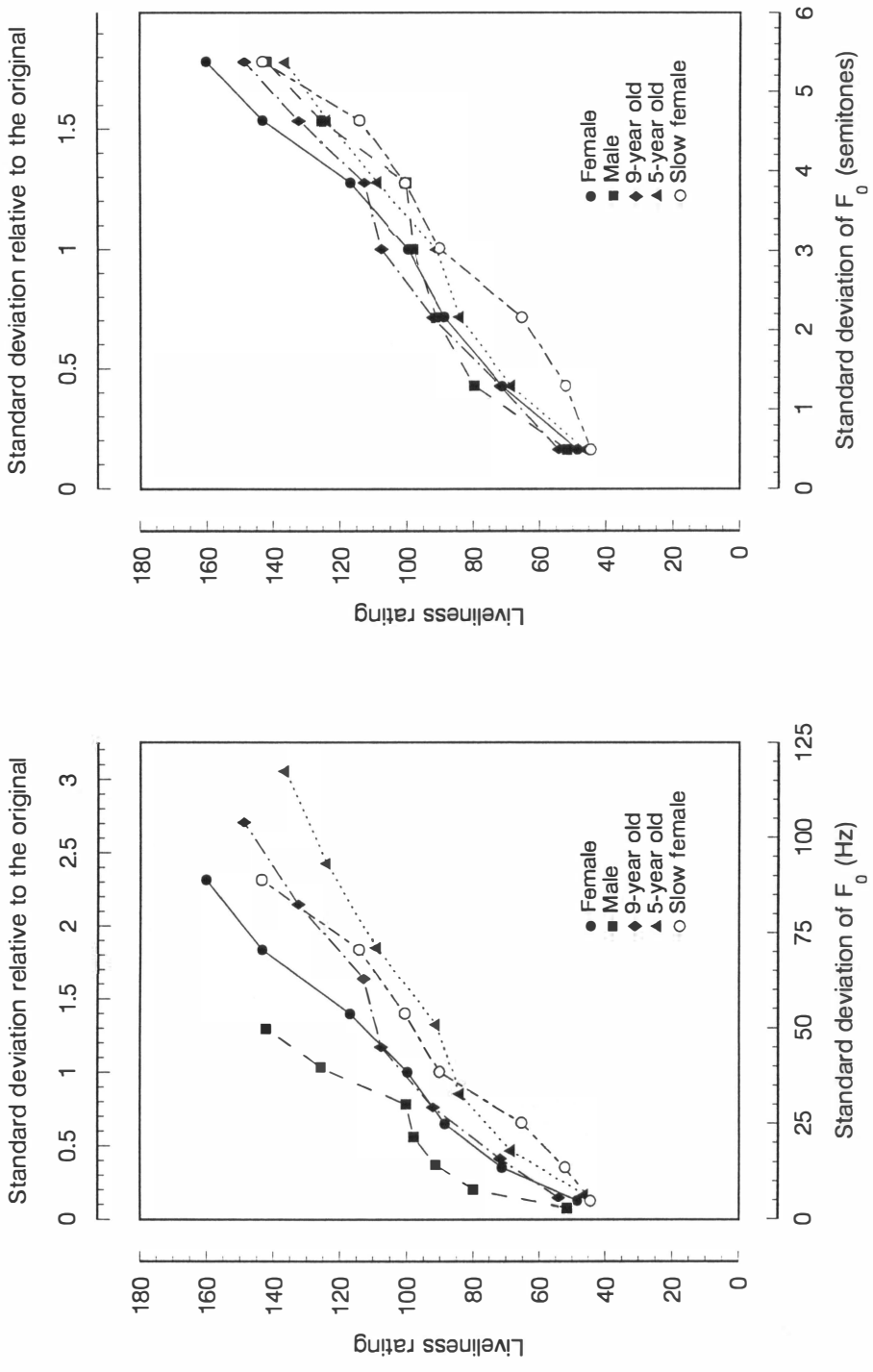


Figure 4. Liveliness ratings obtained in Exp. 1 for five types of speech shown as a function of the extent of the F_0-excursions expressed **(a)** in Hz and **(b)** in semitones.

However, the lines describing the subjects' ratings of the liveliness of stimuli with different speaker characteristics do not quite coincide in Fig. 4b either.

It can be seen quite clearly that the ratings given to the slow female speech are all lower than those of the other types. The ratings of the speech that was intended to represent a 5-year old are also in the lower range. This appears to reflect the extent to which the liveliness ratings were influenced by speech rate. Since the subjects were asked to rate 'liveliness', it was anticipated that slower speech might result in lower ratings. However, the only speech type with an intentionally deviant speech rate was the slow female speech. The other stimuli were given a speech rate that is typical for the type of speaker that was simulated. Over the whole range, the liveliness ratings given to the slow female speech were about 15 units lower than those given to female speech at the normal speech rate. If this is a representative result it would indicate that, in liveliness perception, the effect of speech rate adds linearly to the effect of the F_0-excursions.

A multiple regression analysis performed on the pooled data from the 17 subjects revealed that only k_e and speech rate (k_r) contributed significantly to the results. If the responses to the slow female speech and to that of the intended 5-year old are excluded, the analysis reveals no significant contribution of k_r. The contribution of k_r becomes significant if the slow female speech or that of the intended 5-year old is included ($p < .01$ and $p < 0.05$, respectively).

The design of Exp. 1 does not allow definite conclusions to be drawn about the influence of speech rate in quantitative terms. Nor is it possible to give a full explanation of the low liveliness ratings of the utterances that were intended to represent a 5-year old. However, if that speaker has been perceived as substantially older than 5-years, then the speech rate would seem too slow and this could result in lower liveliness ratings. An inspection of the age ratings does provide evidence to support this suggestion. The median age rating for the intended 5-year old speaker was 10 years, which is considerably higher than the intended age.

In order to investigate the influence of speech rate more systematically, a second experiment was designed in which both the F_0-excursions and speech rate were systematically varied. The results of the age-ratings of the stimuli presented in group 4 will be presented and discussed together with the results of a similar task in Exp. 2.

Even if the pooled results show a consistent and rather expected pattern, with the possible exception of the influence of speech rate for the 5-year stimuli, there was substantial between-subject variation in the weights attached to the various cues for liveliness. Fig. 5 shows the variation in the liveliness ratings expressed as the difference in the ratings between the 75- and the 25-percentile of the response distributions plotted against the F_0-variation, expressed in Hz and in semitones. In

this way the most aberrant responses are excluded. It can be seen that the remaining 50% of the ratings agree most closely when the F_0-excursions in the comparison stimulus are exactly the same as those in the standard if compared in relative terms, such as on a semitone scale. In this experiment, there were three such pairs: female–female, male–male, and female–male. In the first two cases, the compared utterances were completely identical, but the last pair is up to the point. In this case there was, in fact, complete agreement among the responses shown: 14 subjects produced the rating 100 (there were two ratings above and two below 100).

For the group of subjects as a whole there was no significant contribution of (k_b k_e), i.e. of the F_0-variation in Hz, but for three individual subjects there was, and one of them appears to have relied almost entirely on it. Another subject seems to have relied only on speech rate. This between-subject variation appears to indicate that 'liveliness' is not a fundamental percept like loudness or pitch. The test variable 'liveliness' was, however, considered to be the most appropriate means to focus the subjects' attention on the extent of the F_0-excursions, assuming the latter notion not to be included in a naive subject's competence.

4. Experiment 2: The effect of articulation rate on the perception of liveliness

The object of Exp. 2 was to obtain more precise information about the interplay of F_0-excursions and speech rate in liveliness perception. The previous experiment gave an indication that an utterance with a lower than normal speech rate is perceived as less lively than an utterance with a normal speech rate, given the same amount of variation in F_0. It seems reasonable, then, to hypothesize that a faster than normal speech rate will have a similar effect in the opposite direction. In this experiment the effect of fast and slow speech rates was examined for two of the speech types used in the previous experiment.

4.1 Subjects

Nineteen listeners, 7 male and 12 female, served as subjects in this experiment.

4.2 Stimuli and Procedure

In this experiment the female version of the utterance with $k_e = 1.00$ was used as the standard in all comparisons. The comparison stimuli were adult male and 5-year old child, each with three rates of speech including the same as that used in Exp. 1 and, in addition, a slower and a faster version. The adult male and the 5-year old had been chosen for this experiment since they represent the extreme values of F_b. The values of k_r were 0.820, 1.000, and 1.220 for the adult male and 0.672, 0.820, and 1.000 for the 5-year old.

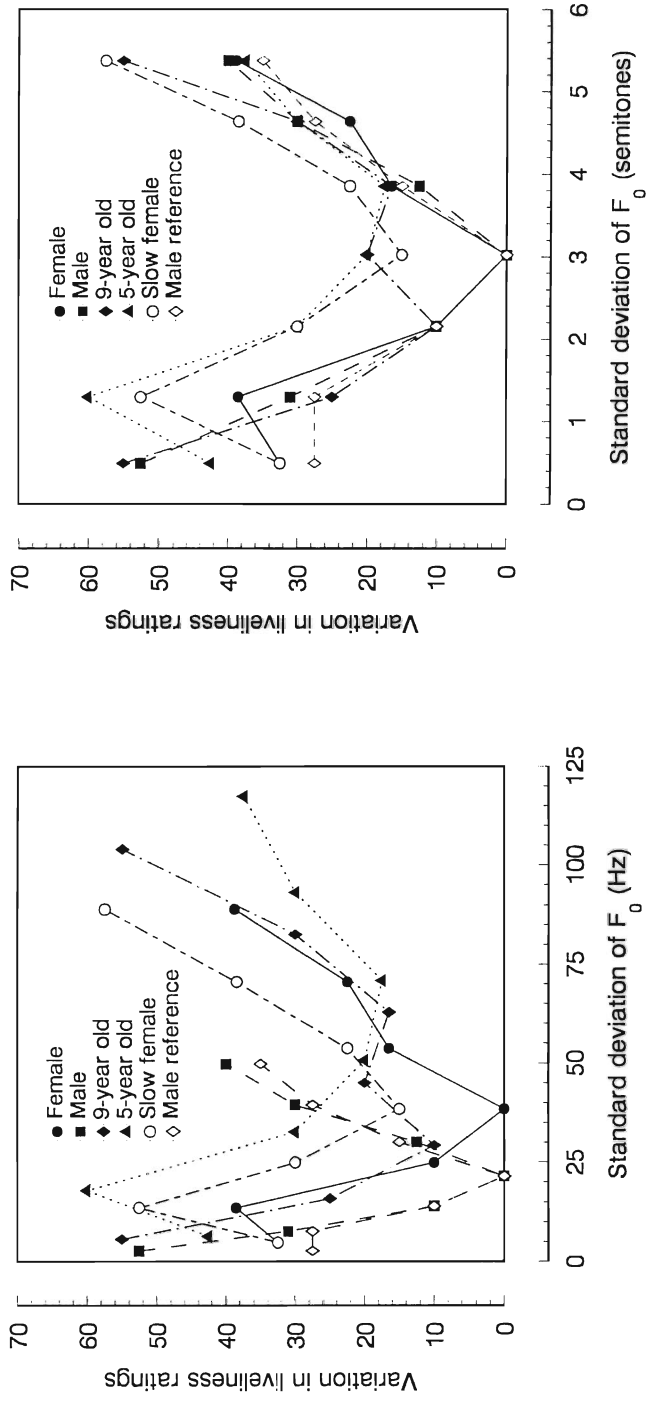


Figure 5. Standard deviation of liveliness ratings obtained in Exp. 1 plotted against the extent of the F_0-excursions expressed **(a)** in Hz and **(b)** in semitones.

The stimuli were presented in two groups. The first and major group contained 42 stimulus pairs with a female standard. These included, beside the mentioned adult male and the 5-year old's versions also seven female versions, namely the same as those in group 1 of Exp. 1. The pairs were presented in random order. Part 2 consisted of the 7 different versions with $k_e = 1.00$, presented alone for the purpose of judging age and sex of the speakers.

4.3 Results and discussion

The results from one of the subjects had to be discarded since this subject had left out several answers and, for some of the remaining answers, there was doubt whether they had been marked in the right place on the response sheet. The average of the pooled results of the remaining 18 subjects is shown in Fig. 6. As for the normal speech rate, the results were very similar to those obtained with the same stimuli in Exp. 1. Again, the utterances which received similar liveliness ratings were those in which the extent of the F_0-excursions was the same expressed in semitones and not in Hz.

As for the influence of speech rate on the perception of liveliness, it can now be seen that a change in speech rate results in an increase or decrease in liveliness ratings that is roughly constant throughout the entire range of F_0-excursions. For both types of speaker, the increase in speech rate of 22% per step from slow to normal to high, resulted in an average increase in liveliness ratings of 27 percentage units per step. In approximation, the contribution of speech rate to perceived liveliness appears to be added arithmetically to that of the F_0-excursions and there

Table V. Weights and significance levels of the factors k_e, k_e k_b, k_r, k_b and the constant obtained in a linear multiple regression analysis performed on the liveliness ratings by all subjects in Exp. 2 and by those who perceived the '5-year old' as younger (Group 1) or older (Group 2) than 13 years.

	All subjects		**Group 1**		**Group 2**	
	Weight	**Sign.**	**Weight**	**Sign.**	**Weight**	**Sign.**
k_e	30.4	.0000	35.8	.0000	22.0	.0003
$k_e \cdot k_b$	2.4	.5	–2.0	.6	9.3	.1
k_r	110.4	.0000	117.6	.0000	99.1	.0000
k_b	3.1	.5	14.5	.014	–14.7	.07
(const.)	–49.6	.0000	–66.0	.0000	–23.9	.1

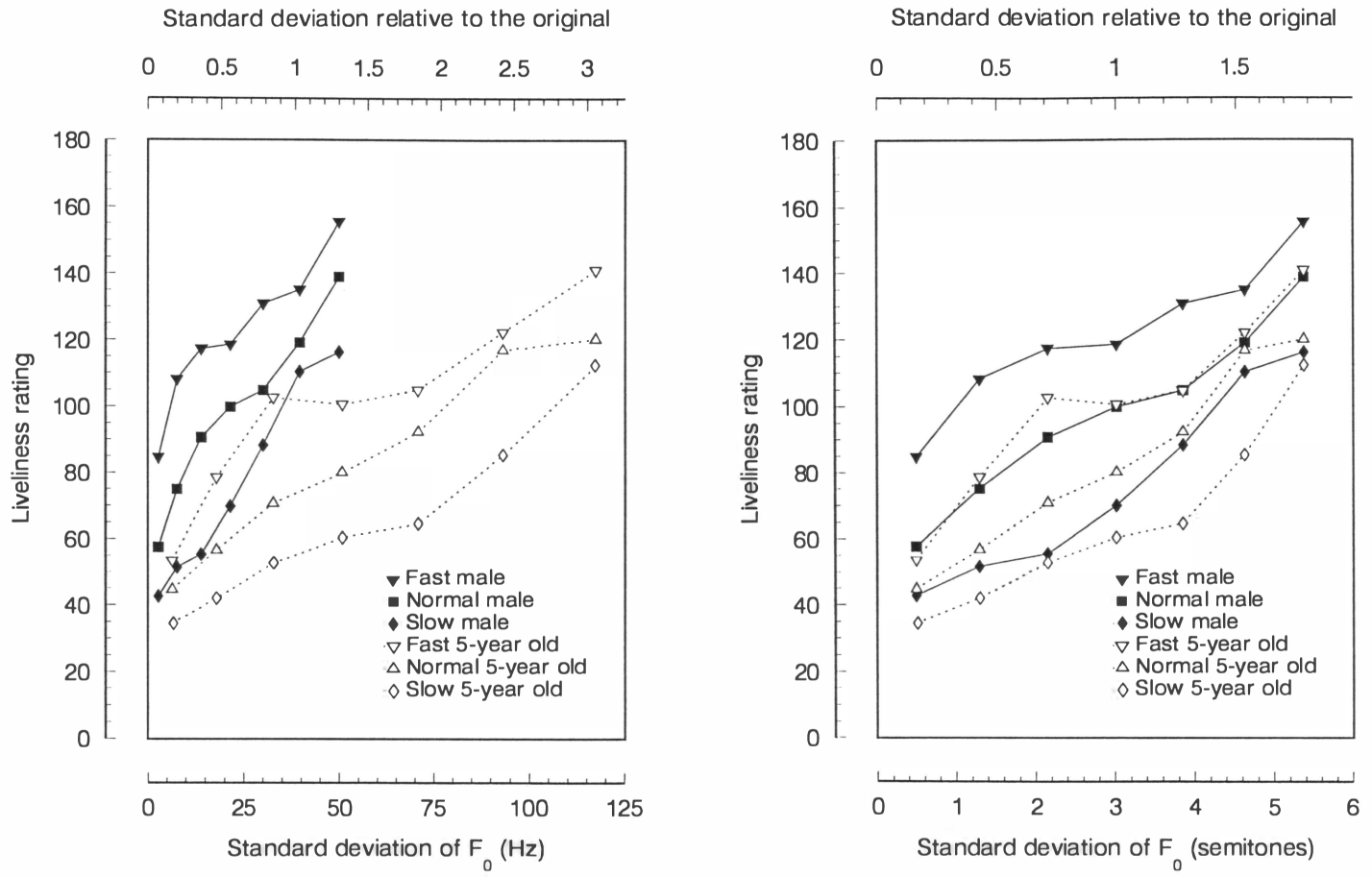


Figure 6. Liveliness ratings obtained in Exp. 2 for three rates of speech by two types of speaker shown as a function of the extent of the F_0-excursions expressed **(a)** in Hz and **(b)** in semitones.

is a roughly linear relationship between the extent of the F_0-excursions and liveliness as well as between speech rate and liveliness. It is, therefore, appropriate to subject the data to a linear multiple regression analysis, whose results are shown in Table V. The variables considered in this analysis include, beside the liveliness ratings, were k_e, $k_b k_e$, k_r, and k_b. The latter was 0.56 for the adult male and 1.32 for the '5-year old'. According to Table V, a 1% increase in speech rate resulted in a 1.1% increase in liveliness, while a 1% increase in the extent of the F_0-excursions resulted in a 0.3% increase in liveliness. The F_0-variation in Hz did not result in any significant additional effect, nor did k_b. As for the results of the individual subjects, there was a significant positive correlation with the absolute variation of F_0 in three cases but this was balanced by three other subjects who showed a significant negative correlation.

As in Exp. 1, the speech of the intended 5-year old, with a speech rate that was 18% lower than that of the adult male, which corresponds to the normal difference between these speaker groups, was perceived as less lively. Yet, if we compare the utterances whose speech rate was the same, we can see an adult–child difference in the opposite direction. It may be that the perception of the liveliness of an utterance is governed by the perceived age of the speaker. If, then, the intended 5-year old was perceived as an adult, we would expect the liveliness ratings for the utterances with the same speech rate to coincide. An analysis of the age ratings given by individual listeners may reveal whether there is such an interaction between perceived age and perceived liveliness.

The result of the age ratings of the utterances presented in Exps. 1 and 2 are shown in Fig. 7. Since the number of utterances presented for age-rating was small and we could not see any systematic effect of speech rate nor of spectral emphasis, the ratings of all stimuli with the same formants and F_0 were collapsed. Unexpectedly, however, we did observe a significant effect of the sex of the *listener* on the perceived age of the speaker. Therefore, the ratings by male and by female listeners are shown separately in Fig. 7. The age-ratings of the adult voices resulted in a median of 30 years for both male and female speech and for both male and female listeners. As for the ratings of the child voices, however, there was a clear difference between the ratings by men and by women. The median rating of the intended 5-year old was 8 years as judged by women and 15 years as judged by men. The difference between the response distributions is highly significant ($p < 0.0005$). This holds true whether age is scaled linearly or logarithmically. In the latter case, the distribution of responses assumes a more normal shape. As for the intended 9-year old (in Exp. 1), the discrepancy between the judgments by men and by women was not so large but still significant. There was also a marginally significant

sex difference in the ratings of the adult female stimuli, but not in the adult male stimuli.

We shall now test the hypothesis that the perception of liveliness is influenced by the perceived age of the speaker against the alternative that, for identical stimuli, the perceived age of the speaker has no influence on perceived liveliness. For this purpose, the set of subjects has been divided into two groups — those who perceived the speech as that of a child and those who did not. The dividing line has somewhat arbitrarily been chosen as 13 years. Group 1 includes 11 subjects, 1 man and 10 women and the median of their ratings was 8 years. Group 2 includes 7 subjects, 6 men and 1 woman with a median age rating of 35 years. Fig. 8 shows the average liveliness ratings given by these two groups to the adult male and '5-year old' stimuli as a function of speech rate. This reveals a clear difference. In Fig.8b, the liveliness ratings of the stimuli that were intended to represent 5-year olds but which were probably perceived as small adults appear to fall on the same trajectory as the responses to the stimuli with adult male characteristics. In Fig. 8a, showing the results for those subjects who perceived the intended 5-year old at least as a child, albeit somewhat older, the liveliness ratings of the child's utterances fall on a trajectory that is different from that describing the responses to the adult utterances. In order to test the significance of the contribution of the perceived age to the liveliness ratings, a multiple regression analysis was performed separately for the two subgroups whose results are shown in Figs. 8a and 8b. The result of these analyses is also entered in Table V. The difference between the two groups is primarily reflected in the weight of the variable k_b. The contribution of this variable is significantly positive ($p < 0.02$) in the results of those subjects who perceived the speaker as a child, while it its negative and approaching significance ($p < 0.08$), in the results of those who perceived the speaker as older than 13 years. We must, therefore, accept that the perceived liveliness is dependent on the perceived age of the speaker.

In choosing the stimuli, we had considered the possibility that the liveliness ratings of stimuli with the same "intended speech rate", i.e., slow, normal, high in relation to what is normal for the simulated type of speaker rather than in absolute terms, might coincide. Fig.8a shows at least a tendency in this direction. The agreement is very good for the slow versions, intermediate for the normal ones, and for the fast versions there is at least a slight tendency in this direction. Complete agreement is not to be expected since the age of the intended 5-year old was perceived as higher than that (8 years) even in this group of subjects and they would expect a higher speech rate.

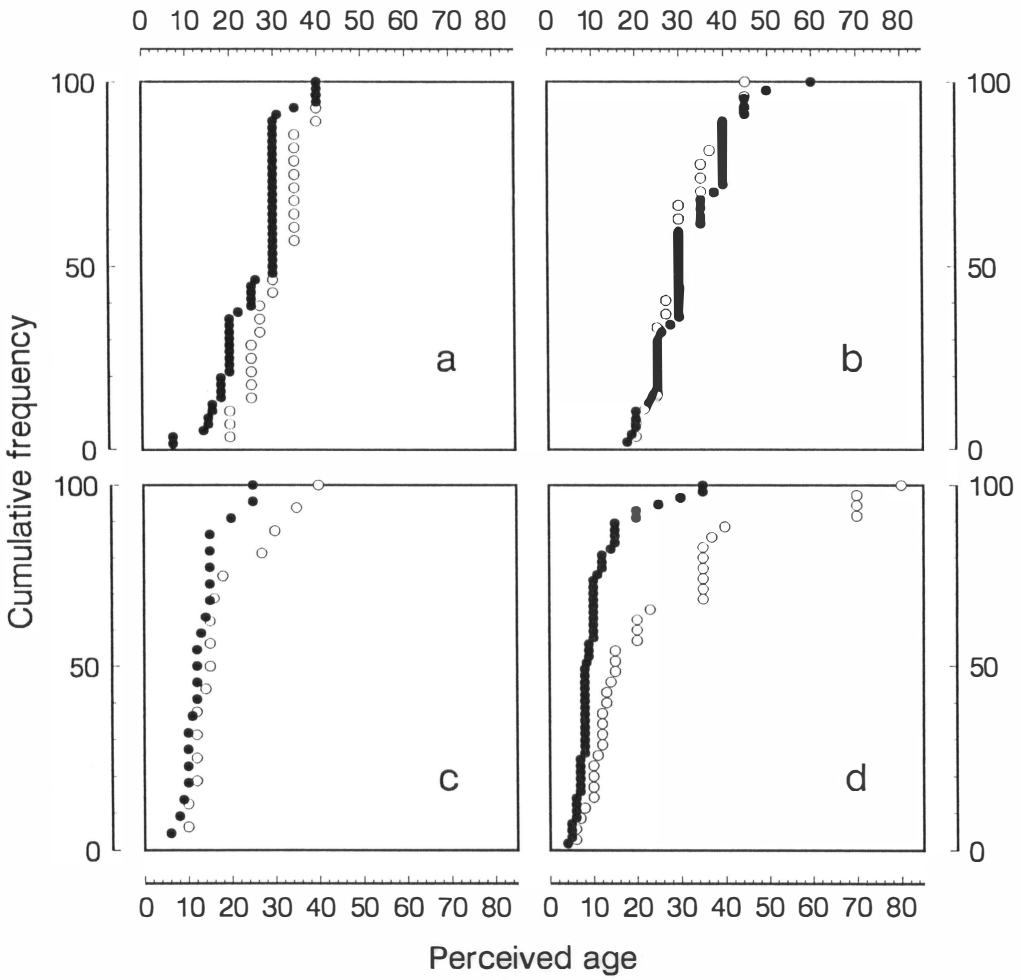


Figure 7a–d. Cumulative distribution of age ratings in response to resynthesized utterances. **(a)** Adult female speaker, 28 years of age. Two stimuli in Exp. 1, one with reduced articulation rate, and one stimulus in Exp. 2. Perceived sex: Female, with three exceptions (boy of age 7, 14, 16 years). **(b)** Adult male speaker, one stimulus in Exp. 1 and three stimuli in Exp. 2 with low, normal, and high articulation rates. Perceived sex: Male, without exception. **(c)** Speaker with intended age 9-years. Two stimuli in Exp. 1, with and without de-emphasis of the higher frequencies. Perceived sex: Female, with four exceptions. **(d)** Speaker with intended age 5-years. Two stimuli from Exp. 1 with and without high frequency de-emphasis and three stimuli from Exp. 2, all with de-emphasis but with low, normal, and high articulation rates. Perceived sex: Female, with two exceptions.

5. Experiment 3: The effect of virtual voice register on the perception of liveliness

The results of Exps. 1 and 2 clearly do not support the conclusions of Hermes and van Gestel (1991) who claimed that F_0-excursions are perceptually equivalent if they are the same expressed in ERB. The results on which this claim is based have, however, been obtained with stimuli that did not vary in speaker age nor in sex. In the stimuli used by Hermes and van Gestel, the formant frequencies remained the same for all stimuli, while only F_0 was varied. The paralinguistic variation simulated was similar to a switch between modal and falsetto register by an adult male speaker. Although we must reject the conclusion arrived at by Hermes and van Gestel, this is not to say that there must have been an error in the acquisition of their experimental results.

As for the phonetic quality of vowels, which is primarily varied by articulation, it is well known that in perception F_0 interacts with the formant frequencies, in particular with F_1 (Traunmüller, 1988). It would, therefore, hardly come as a surprise if formant frequencies, in particular F_1, were to interact with F_0 in perception of the prosody and the paralinguistic quality of speech, which are

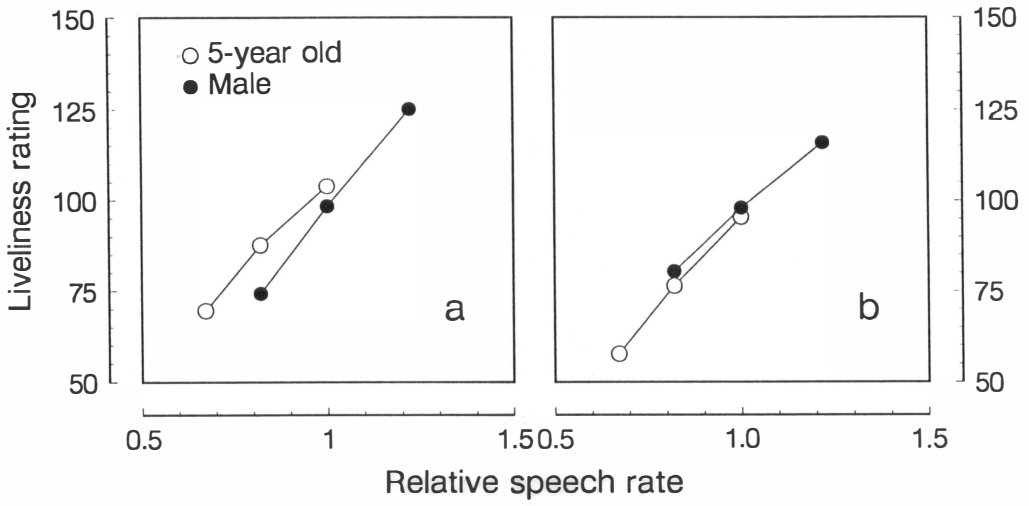


Figure 8. Average liveliness ratings obtained in Exp. 2 by subjects who perceived the intended 5-year old as **(a)** younger and **(b)** older than 13 years shown as a function of speech rate relative to the original. All responses to stimuli with the same speaker and speech rate have been pooled.

primarily varied by phonation. The apparent discrepancy between the results obtained by Hermes and van Gestel and those of the present Exps. 1 and 2 may be due to such an interaction. Exp. 3 is an attempt to clarify this question.

5.1 Subjects

Nineteen listeners, 8 male and 11 female, served as subjects in this experiment.

5.2 Stimuli and procedure

Five different voices were used for the stimuli in this experiment, the normal male and female voices as well as falsetto versions of both these speaker types and a female voice in a low register, reminiscent of the speech of some female smokers. For data see Table IV. The normal female voice served as standard also in this experiment.

5.3 Results and discussion

The average liveliness ratings of each stimulus are shown in Figs. 9 and 10. In Fig. 10, the ratings are plotted against the F_0-excursions on a semitone scale. It can be seen that the falsetto voices received higher ratings for both the female and the male voice types while the female voice in the low register received lower ratings than the normal female voice.

In Fig. 9, the liveliness ratings are plotted against the F_0-excursions in Hertz. Comparing these plots with those of the same data in Fig. 10, we can see a discrepancy in the opposite direction. The curves representing the ratings for the falsetto voices are now lower and those for the low register voice are higher. If we had used only a male speaker, i.e., only the data shown in Figs. 9a and 10a, we might conclude from these results that the scale on which F_0-excursions with the same extent are perceptually equivalent is neither a semitone scale nor a Hz scale but a scale somewhere in between those two, as suggested by Hermes and van Gestel (1991). If the data shown in Figs. 9 and 10 are plotted against the variation in ERB-rate of F_0 (lowest partial), the discrepancy between the ratings of the stimuli that were most similar to those used by Hermes and van Gestel, namely those simulating speech in the malefalsetto vs. modal register,is, indeed, diminished dramatically. This holds also in a comparison of female speech in the modal register with that in the low register. If we compare speech in the female falsetto register with that in the female modal register, we find that the results areeven compatible with scaling the F_0-variations in bark or in Hertz. However, the data obtained in Exp. 1 and 2 suggest this kind of approach to be fundamentally flawed. Scaling the extent of the F_0-excursions in ERB would, instead, lead to a substantial discrepancy between predicted and observed liveliness for speakers who differ in age or sex.

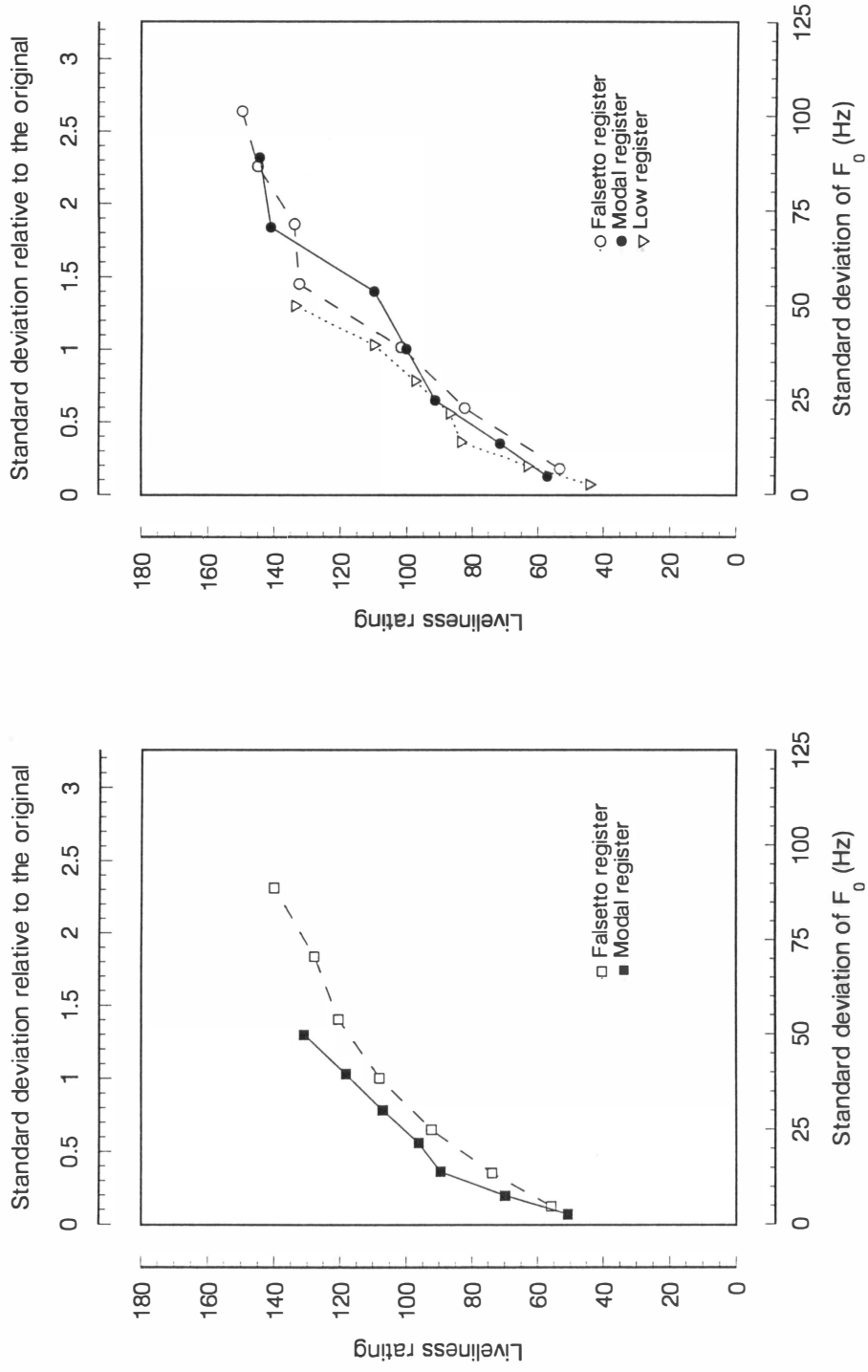


Figure 9. Liveliness ratings obtained in Exp. 3 for the male **(a)** and female **(b)** voices shown as a function of the extent of the F_0-excursions expressed in Hz.

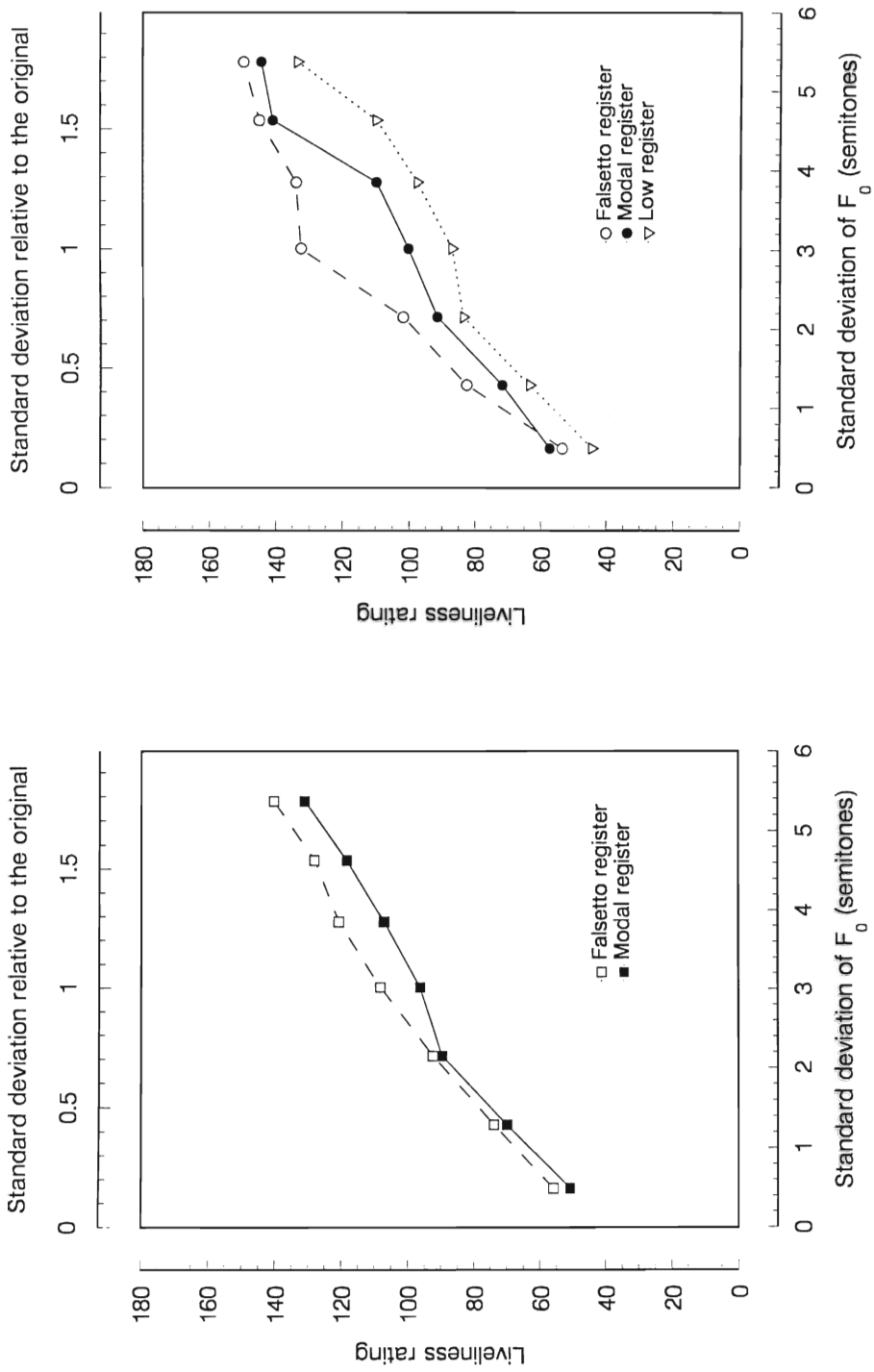


Figure 10. Liveliness ratings obtained in Exp. 3 for the male **(a)** and female **(b)** voices shown as a function of the extent of the F_0-excursions expressed in semitones.

Thus, we cannot escape the conclusion that liveliness ratings are influenced by some factor in addition to the extent of the F_0-excursions and speech rate.

It has been suggested above that the formant frequencies, in particular F_1, might be relevant for the liveliness ratings. This question can be answered by comparing the liveliness ratings given to stimuli in which the F_0-contour was the same while there was a difference in formant frequencies. The female low register versions had the same F_0-contour as the male modal register versions, while the male falsetto versions had the same F_0-contour as the female modal register versions. The female falsetto versions are excluded from this comparison. In a pairwise comparison test, the ratings given to the stimuli in the low female register were found to be significantly lower than those given to the stimuli in the male modal register, by 6.1 units on average. There was also a difference in the expected direction between the stimuli in the female modal register and in the male falsetto register, but for the pooled results, this difference of 2.6 units failed to attain the 5% significance level.

6. General discussion

The present investigation had two aims. First and foremost, to answer the question of how listeners evaluate F_0-excursions, i.e., what kind of scale should be used in order for perceptually equivalent F_0-excursions to be represented by equal intervals. The second aim concerned the perception of 'liveliness' as such. We wanted to know on which basis subjects rate this psychological variable.

The results of Exp. 1 showed that listeners judge F_0-intervals in the speech of men, women, and children to be equal if they are equal in semitones. The data shown in Fig. 4 may still be compatible with a scale that deviates to a slight extent from a semitone scale. However, the overwhelming agreement among all except the most 'noisy' listeners in their judgment that the liveliness of the male version of the utterance with a $k_e = 1.00$ agrees precisely with that of the female standard with the same value of k_e, shows us that the semitone hypothesis holds with as high a degree of precision as could be achieved given the resolution of the present experiments. Around this minimum the between-listener variation in the liveliness ratings increases with an increasing difference between standard and comparison in the relative extent of the F_0-excursions, as can be seen in Fig. 5b, but also with differences in speech rate. This is primarily a consequence of differences between subjects in the slopes of their liveliness functions with respect to these two cues.

In our analysis, we have chosen to oppose a logarithmic against a linear scale of frequency. By rejecting the relevance of the linear scale, we also reject the relevance of the bark-scale. Within the region of F_0-frequencies, the bark scale deviates only marginally from a linear scale of frequency while the ERB-scale can be said to be about half-way between a linear and a logarithmic scale. Replacing

the linear scale of frequency with an ERB scale would reduce the discrepancies in Figs. 4a and 6a, but the discrepancy between adult male and 5-year old would still be very large and we could not explain the between-subject consensus in their ratings of the male and female stimuli with $k_e = 1.0$.

While the results of Exps. 1 and 2 showed that Hermes and van Gestel's (1991) claim that F_0-excursions are perceptually equivalent if they are the same expressed in ERB is not a valid generalization, the results of Exp. 3 are compatible with the results on which that claim was based. Against the background given by the results of Exps. 1 and 2, this can only be understood if it is assumed that listeners take an additional factor into account when rating liveliness and this factor must be related to the frequency position of the formants. The liveliness evoked at a given speech rate by F_0-excursions with a given extent in semitones appears to depend on the amount of space available below F_1. If the spectral distance between F_1 and F_b is larger than what it normally is, the perceived degree of liveliness is decreased — if it is smaller, liveliness is increased. We shall not attempt to evaluate this hypothesis quantitatively. In order to do that we would need to decide upon a definition of F_1 in this context. What is relevant here is probably some kind of mean rather than the instantaneous position of F_1. We would also have to decide how to scale the distance between F_1 and F_0. A bark scale is probably more adequate than a logarithmic scale for describing the auditory mapping of formants, but even this approach has its intricacies. In normal phonation, the bark distance between the average F3 and the average or base-value of F_0 can be said to be independent of the age and sex of the speaker (Traunmüller, 1988). The same can be said about F_2, but it does not hold so nicely for F_1. Our present results do not tell us how to handle this problem. It is not even clear that what we are dealing with is an immediate interaction between F_0 and the formants. The observed interaction might, instead, be mediated by the prior perception of personal properties such as the speaker's 'size' or age and sex and the listener's expectations concerning the speech of such a speaker.

Although there was no other objective factor than the formant frequencies to which the deviation of the results of Exp. 3 from those of Exps. 1 and 2 could be attributed, the attempted demonstration of this effect failed to attain significance in comparing the male falsetto voice with the modal female voice. When looking for an explanation of that failure, we noticed that the range of ratings for stimuli in which the speaker was the same as in the standard was slightly larger than in the other cases. This is reflected in the lines describing the liveliness ratings of utterances with different speaker characteristics, e.g., in Fig. 4b in a reduced slope in the middle of the range. This explains why the expected effect was smaller in the comparison of the male falsetto voice with the modal female voice than in the

comparison of the modal male with the low-register female voice. Only the first mentioned case involved a judgment against a voice with the same speaker characteristics — the modal female voice. In all the other cases the ratings of the stimuli with a high k_e will be reduced since the speaker is not the same as in the standard. Thus, the increased liveliness due to the falsetto characteristics of the male voice will, to some extent, be neutralized by the decrease in liveliness ratings due to the difference in speaker characteristics.

As for the second aim of this investigation, we have seen that the perceived degree of liveliness depends mainly on the extent of the F_0-excursions and on speech rate. For an average listener, a 1% increase in liveliness can be achieved by increasing speech rate with 0.9%, while the extent of the F_0-excursions must be increased by about 3% in order to achieve the same increase in perceived liveliness. When analyzing our results, we considered the hypothesis that the perceived liveliness might be given by the derivative of the F_0-movement. This hypothesis had to be rejected since in that case, the effect of a 1% increase in speech rate should have been no larger than that of a 1% increase in the extent of the F_0-excursions. In this context it is relevant to mention that listeners are in general very sensitive to variations in speech rate. It has been found that ratings of speech rate are approximately proportional to the2nd power of speech rate expressed in syllables per second (Grosjean and Lane, 1981).

We had expected the liveliness ratings to vary in accordance with power functions of speech rate and of the extent of the F_0-excursions. Although we still believe a power-function approach to be theoretically preferable to a linear approach, at least for the description of the relation between speech rate and liveliness, we have chosen to follow the latter approach after finding that most, but not all, subjects would attach a non-zero rating to the liveliness of an utterance with a completely flat F_0-contour.

The age of the simulated speakers who were intended to represent children, especially that of the '5-year old' has not been perceived as intended in spite of the great care taken to ensure that F_0 as well as the formant frequencies and speech rate approximated the data published on speakers of that age. Results obtained in another study (Traunmüller and Bezooijen, 1993), completed after the present experiments, have shown that the verbal maturity of a speaker influences, not surprisingly, the perceived age of the speaker. In the present case, the voice characteristics were right for a 5-year old, but the speech style, in the widest sense, was indicative of an adult speaker.

The accidental finding that adult male and adult female listeners give different age ratings to these stimuli may be due to sex-specific ways of resolving the discrepancy between the verbal and the nonverbal cues contained in the stimuli. It

was pointed out to us by R. van Bezooijen, that it has been suggested previously that men might be more sensitive to verbal information and women more to nonverbal information. If this is true, our results would indicate that women integrate all kinds of cues, perhaps attaching slightly more weight to the nonverbal cues than to the verbal cues, thereby arriving at an age-rating that is higher than indicated by the nonverbal cues but lower than indicated by the verbal cues, while men attach more weight to the verbal cues and face difficulties in bridging the discrepancy between the verbal and the nonverbal cues without assuming an abnormal physiology of the speaker. They appear to have interpreted the intended 5-year old as an adult who is small in stature. One subject even described the voice as belonging to an old female dwarf. The question about the basis of this sex-specific behavior remains to be investigated further. Although we can claim to have shown that there *is* a sex-specific difference in auditory age perception, our accidental results are not sufficient to test the suggested hypothesis. The results could, e.g., equally well be due to a sex-specific reaction to the audible distortions that are introduced by the LPC-based speech synthesis.

Acknowledgments

This research has been supported, in part, by a grant from HSFR, the Council for Research in the Humanities and Social Sciences, within the frame of the Swedish Language Technology Program. We are grateful to Renée van Bezooijen for her suggestions concerning the sex-specific behavior of listeners.

References

Bezooyen, R. van. (1984): *Characteristics and Recognizability of Vocal Expressions of Emotion*, Foris Publications, Dordrecht.

Boë, L.-J., M. Contini and H. Rakotofiringa. (1973): "Etude statistique de la fréquence laryngienne," *Phonetica* **32**, 1–23.

Bruce, G. (1982): "Developing the Swedish intonation model," in *Working Papers* **22**, Lund university, Department of linguistics, 51–116.

Brown, B. L., W. J. Strong and A. C. Rencher. (1974): "Fifty-four voices from two: the effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech," *J. Acoust. Soc. Am.* **55**, 313–318.

Chen, G-T. (1974): "The pitch range of English and Chinese speakers," *Journal of Chinese Linguistics* **2**, 159–171.

Chevrie-Muller, C. and F. Gremy. (1970): "Contribution a l'établissement de quelques constantes physiologiques de la voix parlée de l'adulte," *Journal Français d'Oto-Rhino-Laryngologie* **XVI,** 149–154.

Elert, C-C. and B. Hammarberg. (1991): "Regional voice variation in Sweden," in in *Actes du XIIème Congres International des Sciences Phonétiques,* vol. **4**, Université de Provence, Service des Publications, Aix en Provence, 418–420.

Fairbanks, G., and W. Pronovost. (1939): "An experimental study of the pitch characteristics of the voice during the expression of emotion," *Speech Monographs* **6**, 85–105.

Fónagy, I., and K. Magdics. (1963): "Emotional patterns in intonation and music," *Phonetica* **16**, 293–326.

Fujisaki, H., N. Yoshimune and N. Nakamura. (1970): "Formant frequencies of sustained vowels in Japanese obtained by analysis-by-synthesis of spectral envelopes," Unpublished data, University of Tokyo.

Garnica, O. K. (1977): "Some prosodic and paralinguistic features of speech to young children," in *Talking to Children: Language Input and Acquisition,* edited by Catherine E. Snow and Charles A. Ferguson, Cambridge University Press, pp. 63–88.

Graddol, D. (1986): "Discourse specific pitch behavior," in *Intonation in discourse*, edited by Catherine Johns-Lewis, Croom Helm, London and Sidney, pp. 221–237.

Grosjean, F. and H. Lane. (1991): "Temporal variables in the perception and production of spoken sign language," in *Perspectives on the Study of Speech,* edited by P. P. Eimas and J. L. Miller, Lawrence Earlbaum Associates, pp. 207–237.

Haselager, G. J. T., I. H. Slis and A. C. M. Rietveld. (1991): "An alternative method of studying the development of speech rate," *Clinical Linguistics & Phonetics* **5**, 53–63.

Henton, C. G. (1989): "Fact and fiction in the description of female and male pitch," *Language & Communication* **9**, 299–311.

Hermes, D. J. and J. C. van Gestel. (1991): "The frequency scale of speech intonation," *J. Acoust. Soc. Am.* **90**, 97–102.

Johns-Lewis, C. (1986): "Prosodic differentiation of discourse modes" in *Intonation in discourse* edited by Catherine Johns-Lewis, Croom Helm, London and Sidney, pp. 199–219.

Kitzing, P. (1979): *Glottografisk frekvensindikering: En undersökningsmetod för mätning av röst-läge och röstomfång samt framställning av röstfrekvensdistributionen*, Lund University, Malmö.

Ladefoged, P. (1967): "Stress and respiratory activity," in *Three Areas of Experimental phonetics,* edited by Peter Ladefoged, Oxford University Press, London, pp. 1–49.

Maurer, D., N. Cook, T. Landis and C. d'Heureuse (1992) "Are measured differences between the formants of men, women and children due to F_0 differences?," *Journal of the International Phonetic Association* **21**, 66–79.

Mikeev, Y. V. (1971): "Statistical distribution of the periods of the fundamental tone of Russian speech," *Soviet Physics — Acoustics* **16**, 474–477.

Pegoraro Krook, M. I. (1988): "Speaking fundamental frequency characteristics of normal Swedish subjects obtained by glottal frequency analysis,"*Folia Phoniatrica* **40**, 82–90.

Rappaport, W. (1958): "Über Messungen der Tonhöhenverteilung in der deutschen Sprache," *Acustica* **8**, 220–225.

Rastatter, M. P. and R. D. Jaques. (1990): "Formant frequency structure of the aging male and female vocal tract," *Folia Phoniatrica* **42**, 312–319.

Rose, P. (1991): "How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency?," *Speech Communication* **10**, 229–247.

Scherer, K. R. (1974): "Voice quality analysis of American and German speakers," *Journal of Psycholinguistic Research* **3**, 281–290.

Takefuta, Y., E. G. Jancosek and M. Brunt. (1972): "A statistical analysis of melody curves in the intonation of American English," in *Proceedings of the 7th International Congres of Phonetic Sciences, Montreal 1971*, 1035–1039.

Traunmüller, H. (1988): "Paralinguistic variation and invariance in the characteristic frequencies of vowels," *Phonetica* **45**, 1–29.

Traunmüller, H., P. Branderud and A. Bigestans. (1989): "Paralinguistic speech signal transformations," in *Phonetic Experimental Research, Institute of Linguistics, University of Stockholm*, **10**, 47–64.

Traunmüller, H. and R. van Bezooijen. (1993): "The auditory perception of children's age and sex," in *Fonetik93, RUUL* **23**, Dep. of Linguistics, Uppsala University, 69–72.

Williams, C. E. and K. N. Stevens. (1972): "Emotion and speech: Some acoustical correlates," *J. Acoust. Soc. Am.* **52**, 1238–1250.

Phonetic Experimental Research,
Institute of Linguistics,
University of Stockholm (PERILUS),
No. XVII, 1993, pp. 35–44

Quality judgements by users of text-to-speech synthesis as a handicap aid

Olle Engstrand

1. Introduction

This paper reports a survey of the need for improved quality in text-to-speech systems for handicapped users. Quality criteria are concerned primarily with those phonetic aspects which affect the intelligibility and naturalness of the artificially simulated speech (rate of utterance, articulatory precision etc.) rather than phonetic expressions of age, sex, emotion etc. (paralinguistic variation). The aim here is to a) analyse comments collected from visually and speech-handicapped users, and to a certain extent also from dyslectic users, and b) base the specifications for text-to-speech systems on these analyses. These specifications can then be used when priorities are to be made between future improvements, with the requirements of the users in mind.

2. Method

Three groups of handicapped users with partly different requirements and conditions are involved here. The visually handicapped and dyslectic users want synthesis primarily as a means of reading texts, both those they have written themselves, and others. Users with a speech handicap use synthesis to communicate directly with others. For this group, it was therefore natural to consider the experiences of those close to the user, in particular, relatives and medical staff. In some cases, where the user was not available himself, it was necessary to let a relative supply the information.

Six speech-handicapped users and 11 relatives or medical staff were interviewed. The interviews with the speech-handicapped users were conducted by direct speech, text-to-speech synthesis or text telephone. Ten visually handicapped people were interviewed. One dyslectic person was interviewed. About half of the interviews were conducted in front of the synthesis terminal with demonstrations and examples, the others by telephone.

The interviews lasted from half an hour to one and a half hours. All interviews had a free and open format. Care was taken not to influence the informants with

leading questions, and I tried to obtain primarily spontaneous reactions. In addition, a series of follow-up questions were put to the informants, concerning the intelligibility and naturalness of the synthesis, its sensitivity to noise, the time required to become accustomed to it etc., all set in relation to the informants' judgement of phonetic dimensions such as stress, phrasing, intonation, rate of utterance, articulatory precision etc.

3. Results and comments

Most of what the interviews revealed is presented and discussed in this section. The users' reactions are fairly consistent, both between the different handicap groups and to different text-to-speech systems. The various informant groups will therefore generally be treated as a single group. I shall not distinguish between or compare different synthesis systems — there is not sufficient data for that. As far as the users are concerned, the systems seem to have the same fundamental faults and merits, although probably not to the same extent.

It must be stressed that the general attitude to text-to-speech synthesis as an aid is very positive in all the groups (although there are clear exceptions). Many describe synthesis as a wonderful aid which greatly facilitates work, study or communication in general. Criticism and requests for improvements should be considered with this in mind.

3.1 Overall limitations

The main points, expressed very consistently in various forms, can be summarized as follows:

Text-to-speech synthesis is quite intelligible with a certain effort from the listener in quiet surroundings, in a familiar context and after a period of familiarization. The speech style is, however, quite unlike natural speech — it is uniform and monotonous.

3.1.1 Psychological effort, sensitivity to noise

Listening to and understanding synthetic speech always requires greater concentration than natural speech. Synthesis is, therefore, much more sensitive to noise than natural speech. The following comment is typical: "You really have to concentrate, you get tired... You don't automatically understand the synthetic speech, you have to listen carefully."

Comment

The strongest and most consistent impression which the text-to-speech users have put forward concerns the sensitivity of synthetic speech to noise and the effort required to understand the contents of the text. Intelligibility is satisfactory as long

as the user can concentrate well enough. Most users benefit so much from synthesis that they are motivated to apply the necessary concentration, but this appears to cost them considerable psychological effort.

One consequence of this is that for those around the user, the ability to understand synthetic messages is largely a question of familiarity and attitude. This concerns primarily the speech-handicapped group, who mainly direct synthetic speech to other people. Several speech-handicapped users have noticed that many listeners lack the patience to listen and understand. Others have pointed out that the intelligibility of the synthesis is dramatically worse than natural speech for older people with moderate hearing impairments.

Another consequence comes of the following argument concerning aims and means for evaluation of the overall quality of text-to-speech synthesis:

The ability to perceive synthetic messages can be understood and measured in several ways. In what is known as the "sentence-by-sentence" method, which has recently been used in, among other places, Sweden (e.g. Carlson, Granström, Neovius and Nord, 1992; Neovius and Raghavendra, 1993), it is assumed that there is a positive correlation between the quality of the synthesis and the speed with which the test listeners get through the test. In this kind of investigation it has been shown that the listeners' performance in text-to-speech tasks is, in some situations, comparable to corresponding natural speech tasks. The sentence-by-sentence method does not, however, explicitly measure the degree of psychological effort associated with the perception of synthetic messages. For my informants, however, precisely this constitutes an essential difference between natural speech and text-to-speech synthesis. It is clearly desirable for future evaluatory work in the field of text-to-speech to develop adequate tests to explicitly capture this aspect.

3.1.2 Context

My informants have also observed that it is particularly difficult to perceive synthetic messages taken out of context. As one of the users expressed it: "You almost have to know what is being said before you can understand it."

Comment

It is, of course, always, even with natural speech, easier to perceive a message taken in context. However, the informants' claim that the problem is considerably aggravated in synthesis is so definite and consistent that it must be taken seriously.

The familiar context which is often necessary for the synthesis to be relatively easy to understand is usually less of a problem for the visually handicapped users than for the speech-handicapped users. The visually handicapped users I have talked to mostly use synthesis with familiar texts, often written by themselves. (I have not been able to reach e.g. visually handicapped readers of synthesized

newspaper text.) The problem seems to be considerable for the group of speech-handicapped users and those to whom their synthetic speech is addressed.

3.1.3 Length of training

Different informants estimated the time required to achieve relatively good intelligibility as 2–4 weeks of fairly regular use.

Comment

The permanent level of competence is achieved relatively quickly. This fairly short training period might, therefore, not seem to pose a problem. It is, however, worth noting that training is necessary before the synthesis can be dealt with relatively easily. It is also interesting to hear the informants accounts of their own and others' first encounter with the synthesis. The predominant experience is that the simulated speech initially sounds strange and is difficult to understand. This is not the case with perception of, for example, an adult with normal speech and no strong non-standard dialect.

All of these observations concerning sensitivity to noise, effort, context and training time show that the available text-to-speech systems have clear defects as far as phonetic quality is concerned. How can these defects be specified more clearly? Which improvements should be given priority to ease the problems?

3.2 Specific limitations

The informants specifically indicate the following reasons for the above problems.

3.2.1 Phrasing and pauses

The informants reported frequent difficulty in hearing how the text is composed. The following comment is representative: "You want to be able to hear the punctuation", i.e. how words and phrases are grouped together, where sentences begin and end, etc. At places where this does not work satisfactorily, it is easy to get stuck and lose one's place in the text. This makes it more difficult to make sense of the text as a whole.

Comment

Essentially, this means that the boundaries of the syntactic constituents of the text are not marked correctly or clearly enough by prosodic means (intonation, duration, sufficient pauses etc.). This may seem surprising in the light of the long and successful history of research in Swedish prosody. The relationship between syntax and prosody is, of course, a complex one, but I am sure that existing models of prosody, complemented by targeted research (see e.g. Bruce, Granström, Gustafson and House, 1992, 1993; Strangert, Ejerhed and Huber, 1993), would provide a good basis for quality improvements in this area.

3.2.2 Stress and focus

It is often difficult to perceive the relative prominence of words and phrases, i.e. it is not always clear what weight is to be attributed to words and phrases in the context of the text. This hinders the interpretation of the text as a whole.

Comment

This observation has two separate aspects. On the one hand, the problem of how the relative weight of words and phrases is to be marked prosodically can probably be solved on the basis of existing knowledge. We know quite a lot about how this happens using suitable adjustments in fundamental frequency, duration, intensity, articulatory precision etc. On the other hand, we are not yet equipped to solve the much more difficult problem of predicting when and by how much a word or a phrase is to be highlighted in its context.

Consider this example: It is well known that almost all Swedish compound words (words like *äppelträd*, "apple tree", *gräsmatta,* "lawn"...) are pronounced with what is known as the grave tonal word accent, i.e. with a falling tone on the first stressed vowel followed by a rise to the next stressed vowel. This tonal contour is necessary and sufficient to give the impression of a stressed grave accent word (Zetterlund, Nordstrand and Engstrand, 1978). The precise extent of the second rise is determined by how prominent the word is to be in the context of the sentence (cf. Bruce, 1977). If the rise is too small, the word will sound too unstressed, if it is too high, the word will sound too emphatic and be attributed too much prominence in its context. This kind of error of adjustment not only leads to unnatural sounding pronunciation, but also to the listener misinterpreting the intention behind the utterance. This co-ordination between the phonetic properties of speech and its semantics and pragmatics is extremely complex and little studied. There are no concrete models here which can be simply applied in a text-to-speech system, although there are a few rough rules of thumb, e.g. that new information is usually made phonetically prominent, while known information stays in the background. Tentative models are being formed in this area (cf. e.g. Horne, 1991, Horne et al., 1993).

An observation made by almost all the informants also belongs here: Stress is placed on the wrong syllable in many words and phrases (examples are provided, e.g. *kompe'tens* "competence" becomes *kom'petens*. This is disturbing and reduces the usefulness of the synthesis is various situations. For example, when proof-reading one's own texts (visually handicapped and dyslectic users) it is not possible to ascertain whether the spelling is incorrect or the rule system has generated an incorrect pronunciation.

3.2.3 Rate of utterance

It is desirable for the user to be able to vary the speed at which the text is read aloud (rate of utterance). This is, in principle, possible. However, the informants consistently observed that both a relatively low and a relatively high rate of utterance make the speech less natural and less intelligible. In particular, an increase in rate of utterance does not, in itself, lead to more natural sounding speech, and a low tempo does not result in increased clarity, on the contrary, the speech becomes "unclear", "drawled" or "blurred" (representative judgements). (This has given my dyslectic informant particular difficulty, since she is dependent on a slow and clear reading to give her time to check her spelling.)

Comment

It might seem natural that too fast reading would impair the intelligibility of the simulated speech. It is less intuitively obvious that too slow reading also impairs the quality of the synthesis: it should be easier to follow the text when it is read slowly. This is probably not the central issue. I believe that the informants' negative reactions to both high and low rates of speech should be interpreted as follows.

It is probably not primarily the informants' capacity for processing data which sets the limits regarding the speech rate of the synthesis. It is normally easy to understand natural speech, even if it is very fast or very slow. It can, thus, be supposed that the problem with text-to-speech synthesis is rather that its changes in rate of utterance do not accurately imitate those in natural speech. This is a rather complicated issue in itself. The complexity lies in the fact that changes in the tempo of natural speech are not linear. This non-linearity means, for example, that different parts of words and utterances are shortened or lengthened to different extents. Several studies have, for example, shown that unstressed syllables are shortened relatively more than stressed syllables and that vowels are shortened more than consonants when rate of utterance increases (this is the case for Swedish and English, but not necessarily for all languages). Another important characteristic of fast speech, especially in fast reading of text, is that pauses tend to disappear and that boundaries of syntactic constituents and the like tend to be marked less clearly than in slower speech.

A further important example of non-linearity in tempo variation is the following: Certain kinds of sounds are inherently dynamic in the sense that their very identity lies in extremely fast spectral changes. For example, the category of stops, /p t k b d g/ etc. is characterized by rapid spectral change in the VC and CV boundaries, while the corresponding fricatives start and end more gradually. This means that there is a theoretical limit on how much e.g. a vowel-stop-vowel sequence can be linearly extended in time. At a certain point, the inherent abrupt-

ness of the stop will be threatened, and it will then be perceived either as a drawled stop or as the corresponding fricative or affricate. This does not normally happen in natural slow speech. There the vowels and closed portions of the stops are lengthened, while the transitions between vowels and stops are still characterized by rapid spectral transitions.

The conclusion must be that it would be worthwhile testing whether variations in speech tempo in text-to-speech synthesis could be simulated in a more natural way, taking the non-linearity discussed above into account.

Finally, it should be observed that variations in the tempo of natural speech interact with yet another important dimension, i.e. the variation displayed in natural speech between casual and formal phonetic performance. This aspect will be discussed in the next section.

3.2.4 Speech style

Most informants point out the stylistic monotony and lack of variability of the synthesis as a clear failing. I shall attempt to capture this in terms of the variation found in natural speech between *casual* and *formal* and the phonetic correlates of this stylistic dimension.

Comment

The factors we have discussed so far are in many ways related to the intelligibility of the synthetic speech, and thus fairly directly to its usefulness in practical situations in communicating, at work or study. What I have referred to as speech style is really more concerned with the naturalness and flexibility of the synthesis and therefore also with its possible applications. A representative comment from one of the visually handicapped informants was: "Text-to-speech synthesis works well as an aid to write memos and the like, but you would never dream of using it to read fiction.".

It would, however, be an oversimplification to divide the limitations of text-to-speech synthesis into those characteristics purely associated with the conveying of information on the one hand and purely aesthetic or stylistic characteristics on the other. Firstly, there are stylistic aspects to all the above phenomena. Secondly, it is very probable that greater potential for stylistic variation, as was called for by most of the informants, would considerably improve the intelligibility of the text-to-speech synthesis. Let me substantiate this conclusion by drawing attention to a universal aspect of natural speech situations, which we can call *situational adaptation to the listener*.

According to Lindblom (e.g. 1987) natural speaker-listener strategies are characterized by an aspiration to achieve what he calls "sufficient phonetic contrast". In principle, this means that the speaker in a dialogue adjusts his speech in a flexible

way to the requirements of clarity, speech tempo etc. which the listener can be assumed to have in a given situation. Imagine, for example, a group of elderly Social Democrats involved in a discussion about Swedish domestic politics during the 1950s. We would probably often hear utterances like [sos'jamkratna] or [tag'lande]. The low degree of precision in the pronunciation of *socialdemokraterna* "Social Democrats" and *Tage Erlander* (the Prime Minister of Sweden from 1946–1969), an example of what is known as phonetic reduction, is typically found in situations where the vocabulary is limited by a predetermined choice of subject matter. In such a situation, very small phonetic means are sufficient to steer the listeners' perception in the right direction. In other situations, for example when listing a number of unpredictable names or numbers, much greater clarity is required. The pronunciation of a given word varies, therefore, from phonetically reduced "weak" forms to phonetically explicit "strong" forms.

How predictable is the choice between more or less reduced forms? This is really an open question. A rough rule of thumb is that new information is generally assigned a phonetically strong form while previously known information is assigned a weaker form. Another tentative generalization is that "grammatical" words (conjunctions, prepositions etc.) are reduced more than "lexical" words (verbs, nouns etc.). A further relationship which can be observed in natural speech is that the variation between phonetically strong and weak forms is related to the speed at which they are produced, i.e. rate of utterance (cf. above). It is therefore probable that a reduced form such as [tag'lande] is pronounced more rapidly than the corresponding strong form [tage erlander]. This adjustment gives natural speech much of the rhythm which is lacking in the synthesis. (Some informants characterized the rhythm of the synthesis as "monotonous" or "soporific").

The two phenomena (strong vs. weak, slow vs., fast) can, however, also occur independently, that is, a) rapid speech can contain strong forms and b) slow speech can contain weak forms. Similar effects can be found in different situations with a single speaker. A radio commentary of a fast ice-hockey match can contain examples of the first kind. Lennart Hyland's sometimes machine-gun-like commentaries are often given as a classic example. Different combinations of speech tempo and phonetic reduction can also contribute to the formation of personal styles of speech.

The possibility of synthetically simulating personal styles of speech was not given high priority by my informants — there are more important requirements. One such basic requirement is without doubt the ability of text-to-speech synthesis to *accurately and appropriately imitate the normal stylistic variation of natural speech*.

Our theoretical knowledge in this area is, however, still insufficient to be easily applied in a text-to-speech system. This is, thus, yet another area where significant improvements in quality will be attainable only as research makes new models available. This is particularly so regarding the *range* of phonetic-stylistic variation (how great is the difference between strong and weak forms?); its *systematics* (what are the intermediate forms like?); its *perceptual relevance* (what is its contribution to the naturalness and intelligibility of the speech?) and its *predictability* (what factors trigger the choice of strong or weak forms?). Research is, however, underway at a number of international centres, and many useful theories are being formed (see e.g. the special issue on speech styles of *Speech Communication*, vol. 11, nos. 4–5).

4. Summary and conclusion

On the basis of judgements made by a number of users of text-to-speech synthesis as a handicap aid, I have concluded that the following areas should be given priority when seeking to improve the intelligibility and naturalness of the simulated speech: *a) phrasing and pauses, b) stress and focus, c) temporal and stylistic variation.*

This is a tall order, involving several of the central areas of phonetics, prosody in particular. It is not possible on the basis of the available interviews to make priorities between these areas. They are not, in any case, independent of each other; they influence each other in many ways. In some cases, progress might be made using existing theoretical knowledge.On the whole, however, it seems likely that significant improvements in text-to-speech synthesis will have to wait until experimental speech research makes adequate models available, probably a long and trying wait.

Acknowledgments

My warmest thanks to all who most generously contributed with their experiences or otherwise helped me in this investigation. This work was supported by The Swedish Institute for Disabled Persons (Handikappinstitutet).

References

Bruce, G. (1977): *Swedish word accents in sentence perspective.* Travaux de l'Institut de linguistique de Lund **XII**. Lund: Gleerups.

Bruce, G., Granström, B., Gustafson, K. and House, D. (1992): "Prosodic phrasing: a perceptual experiment." In D. Huber (Ed.): *Papers from the Sixth Swedish Phonetics Conference, Gothenburg*, pp. 1–4.

Bruce, G., Granström, B., Gustafson, K. and House, D. (1993): "Prosodic modelling of phrasing in Swedish." In D. House and P. Touati (Eds.): *Proceedings of an ESCA Workshop on Prosody, September 27–29, 1993, Lund, Sweden* (Working Papers 41, Lund University, Department of Linguistics), pp. 180–183.

Carlson, R., Granström, B., Neovius, L. and Nord, L. (1992): "The 'listening speed' paradigm for synthesis evaluation." In D. Huber (Ed.): *Papers from the Sixth Swedish Phonetics Conference, Gothenburg*, pp. 63–66.

Horne, M. (1991): "Phonetic correlates of the 'new/given' parameter." *Proceedings from the XIIth International Congress of Phonetic Sciences, vol. 5*, pp. 230–233.

Horne, M., Filipsson, M. Johansson, C., Ljungqvist, M. and Lindström, A. (1993): "Improving the prosody in TTS systems: morphological and lexical-semantic methods for tracking "new" vs. 'given' information." In D. House and P. Touati (Eds.): *Proceedings of an ESCA Workshop on Prosody, September 27–29, 1993, Lund, Sweden* (Working Papers 41, Lund University, Department of Linguistics), pp. 208–211.

Lindblom, B. (1987): "Adaptive variablity and absolute constancy in speech signals: two themes in the quest for phonetic invariance." *Proceedings from the XIth International Congress of Phonetic Sciences, vol. 3*, pp. 9–18.

Neovius, L., Raghavendra, P. (1993): "Evaluation of comprehension of text-to-speech: the sentence-by-sentence listening paradigm." In J.S. Pettersson (Ed.): *Fonetik-93: Papers from the Seventh Swedish Phonetics Conference held in Uppsala, May 12–14, 1993*, pp. 45–48.

Strangert, E., Ejerhed, E. and Huber, D. (1993): "Clause structure and prosodic segmentation." In J.S. Pettersson (Ed.): *Fonetik-93: Papers from the Seventh Swedish Phonetics Conference held in Uppsala, May 12–14, 1993*, pp. 81–84.

Zetterlund, S., Nordstrand, L. and Engstrand, O. (1978): "An experiment on the perceptual evaluation of prosodic parameters for phrase structure decision in Swedish." In E. Gårding, G. Bruce and R. Bannert (Eds.): *Nordic Prosody: Papers from a symposium*, pp. 15–21.

Phonetic Experimental Research,
Institute of Linguistics,
University of Stockholm (PERILUS),
No. XVII, 1993, pp. 45–54

Word-prosodic features in Estonian conversational speech: some preliminary results[1]

Diana Krull

Abstract

Estonian has three distinctive degrees of quantity: short, long and overlong. This paper reports an investigation on the temporal and tonal correlates to quantity in the natural conversation of one Estonian speaker. The results show statistically significant differences between quantities only for the temporal correlates. The tonal correlates display a considerable overlap between quantities.

1. Introduction

Over the years, several investigations have been addressed to the Estonian quantity system. A number of theories about the nature of the three distinctive quantities and their acoustic correlates have been put forward. (For an overview, see Lehiste 1970, 1988; a new theory is presented by Eek and Help 1987). The phonetic material on which those theories have been based, consists of so called "laboratory speech", that is, words or sentences prepared by the investigator and read by the speaker. However, more recent work has shown that more spontaneously produced speech can differ considerably from such laboratory productions (see e.g. Lindgren, Krull and Engstrand (1987). Nevertheless, the acoustic correlates to phonological distinctions can, at least to some extent, be present also in conversational speech. The stability of such acoustic correlates can differ with language. For example, Engstrand (1992) has shown that duration relations are much more robust in Finnish when compared to Swedish. This may be due to the fact that Finnish has principally one acustic correlate to quantity, whereas Swedish has two: duration and vowel quality.

It can be hypothesized that if an acoustic parameter is used as a primary correlate to phonological distinctions in a language, the freedom of the speakers of the

1) This is a slightly expanded version of the paper with the same title in: D. House and P. Touati (Eds.): *Working Papers*, **41** (Proceedings of an ESCA workshop on prosody, 1993*)*, Department of Linguistics and Phonetics, Lund University.

language to use this parameter for other purposes will be restricted. That is, differences in the acoustic parameter should remain relatively robust across speaking styles. Therefore, an investigation of natural Estonian conversation is important for two reasons: It can help to test the hypothesis, and, at the same time, shed some more light on the question relative importance of different cues to quantity in Estonian.

Estonian has three phonologically distinct degrees of quantity: short (Q1), long (Q2) and overlong (Q3). They are signalled by the duration ratio between the first (main stressed) and the second syllable of a word. The typical ratio for Q1 is 2:3, for Q2 3:2 and for Q3 2:1 (Lehiste 1960). To distinguish Q3 from Q2, listeners use an additional tonal cue (Lehiste 1970): falling F0 for Q3 and flat or slightly rising for Q2.

Is the relatively small temporal difference between Q2 and Q3 maintained in conversational speech? Earlier results with words read in isolation and in a carrier phrase (Krull 1992) showed that the duration relation between the two initial syllables remained stable even when the syllables involved were shortened as a result of increased word length. In most of the cases, the differences in the F0 contour also remained stable.

Other tonal cues described in the literature but not studied in Krull op. cit., are an F0 stepdown from the end of V1 to the beginning of V2 (Lehiste 1970b), and an earlier location in time of an F0 maximum within V1 for Q3 (Eek 1990).

The aim of the present study is to assess the stability of these cues in natural conversational Estonian speech, to begin with, of one speaker.

Table I. Schematic representation of the two initial syllables of Q1, Q2 and Q3 words in Estonian. Only the underlined forms were used in this investigation.

Quantity degree	**Schema**
Q1	(C)VCV
Q2	(C)VVCV
	(C)VCCV
	(C)VVCCV
Q3	(C)VVVC(V)
	(C)VCCC(V)
	(C)VVVCCC(V)

2. Method

The subject was a male phonetician, native speaker of standard Estonian, resident in Estonia. Seated in an anechoic chamber together with the author and prompted by a few short questions, he related episodes from his childhood, schooldays and travels. The talk — over an hour and a half of a lively near-monologue — was recorded digitally. Lexical non-compound words of the form underlined in Table I were located and sampled into a computer at a rate of 10 kHz/s. V1 was a short, long or overlong vowel, or — only when long or overlong — a diphthong; C was a short consonant. The form with a short intervocalic consonant was chosen because the exact duration of a single short consonant in syllable initial position is of no consequence for the quantity degree and therefore V1 and V2 can be used to represent syllables (Lehiste 1960). The three-way quantity contrast is connected to the two initial syllables of a word, of which the first carries the main stress.

For the analysis, the Kay CSL 3400 system was used. Vowel durations were measured and the duration ratio V1/V2 was calculated. The beginning and end of the vowel were defined as the onset and and offset of a clear formant pattern. When the vowel was preceded by a stop, the burst release was defined as the beginning of the vowel. The aspirative phase sometimes appearing at the end of vowels, especially in word final position, was not included.

Next, F0 was measured at the beginning and end of V1 and V2. After stop consonants, the voice onset instead of the burst was now defined as the beginning of the vowel. If there was an F0 maximum within V1, its frequency and temporal location were marked. Utterance final words were not included. The remaining material consisted of 157 words, 48 of Q1, 45 of Q2 and 64 of Q3. Words where V1 was a long or an overlong vowel were measured separately from words where it was a diphthong. Moreover, disyllabic words — which were in a clear majority — were measured separately from words of three or four syllables.

Mann-Whitney U-tests showed no significant change in syllable duration due to the lengthening of the word as was reported by Krull (1992) for laboratory speech, neither was any significant difference found between long vowels and diphthongs. Therefore, these groups were analyzed together.

3. Results

The durations of V1 and V2 are shown in the so called "box and whiskers" plots in Fig. 1, and the V1/V2 duration ratios can be seen in Fig. 2. The data were, in general, not normally distributed, therefore median values are given instead of means, and the Mann-Whitney U-test was used. The test showed statistically significant differences between Q1–Q2, and Q2–Q3, both for the durations of V1 and V2 and for the V1/V2 ratios. For all cases $p<.001$.

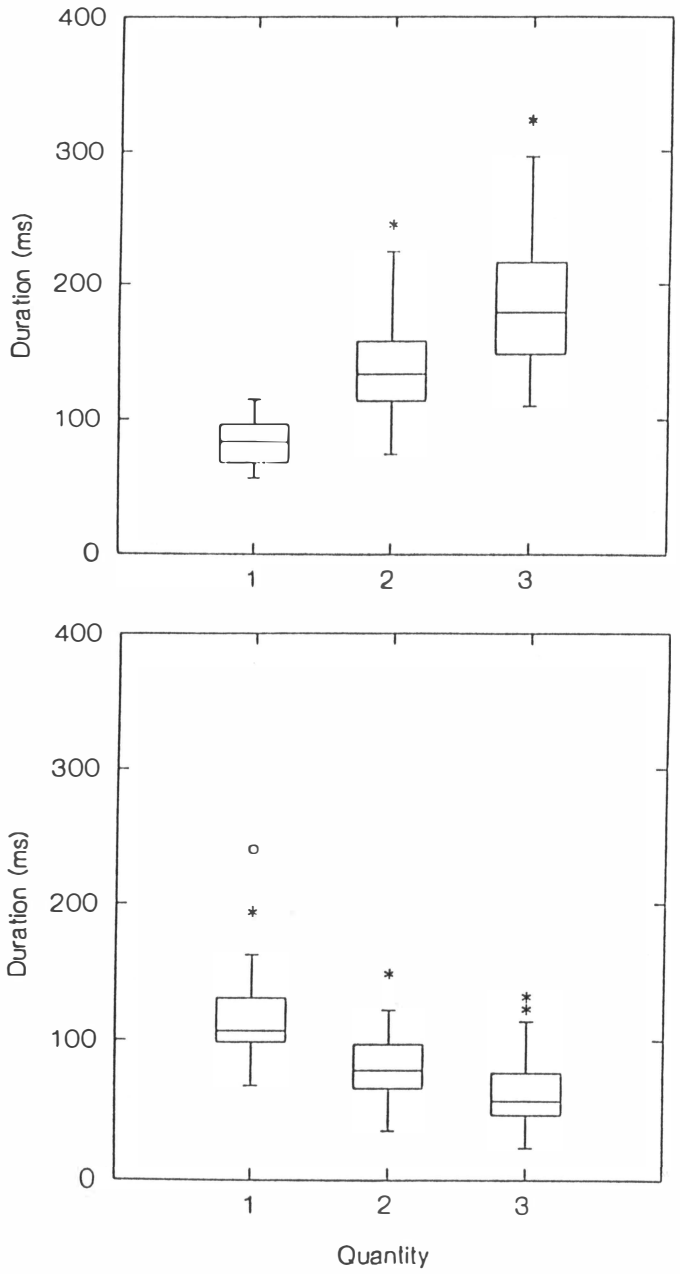
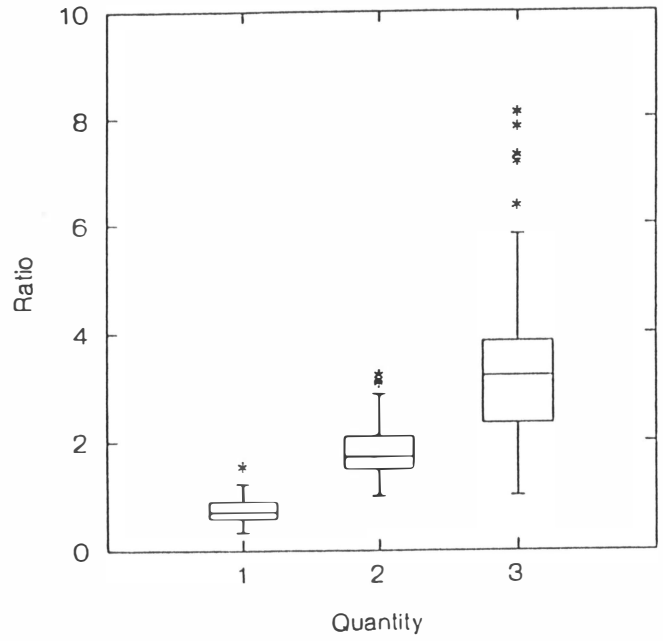


Figure 1. Box-and-whiskers plots of the duration of V1 (top) and V2 (bottom) in three degrees of quantity. The boxes represent the spread of 50% of the sample, 25% (quartile) above and 25% below the median. The whiskers (above and below the boxes) are drawn to the nearest value not beyond a standard span of the quartiles. The points beyond that are represented by astrisks and circles. (For a definition of box-and-whiskers plots, see Velleman and Hoaglin, 1981.)



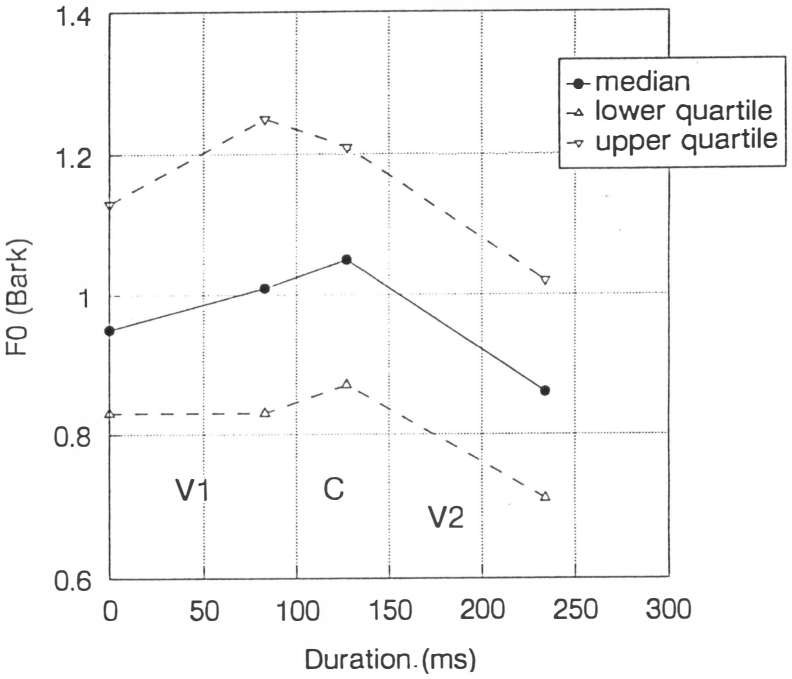


Figure 2 (top). V1/V2 ratio in three degrees of quantity.

Figure 3 (bottom). F0 movement from the beginning of V1 to the end of V2 in Q1 words words . The segment durations are median values.

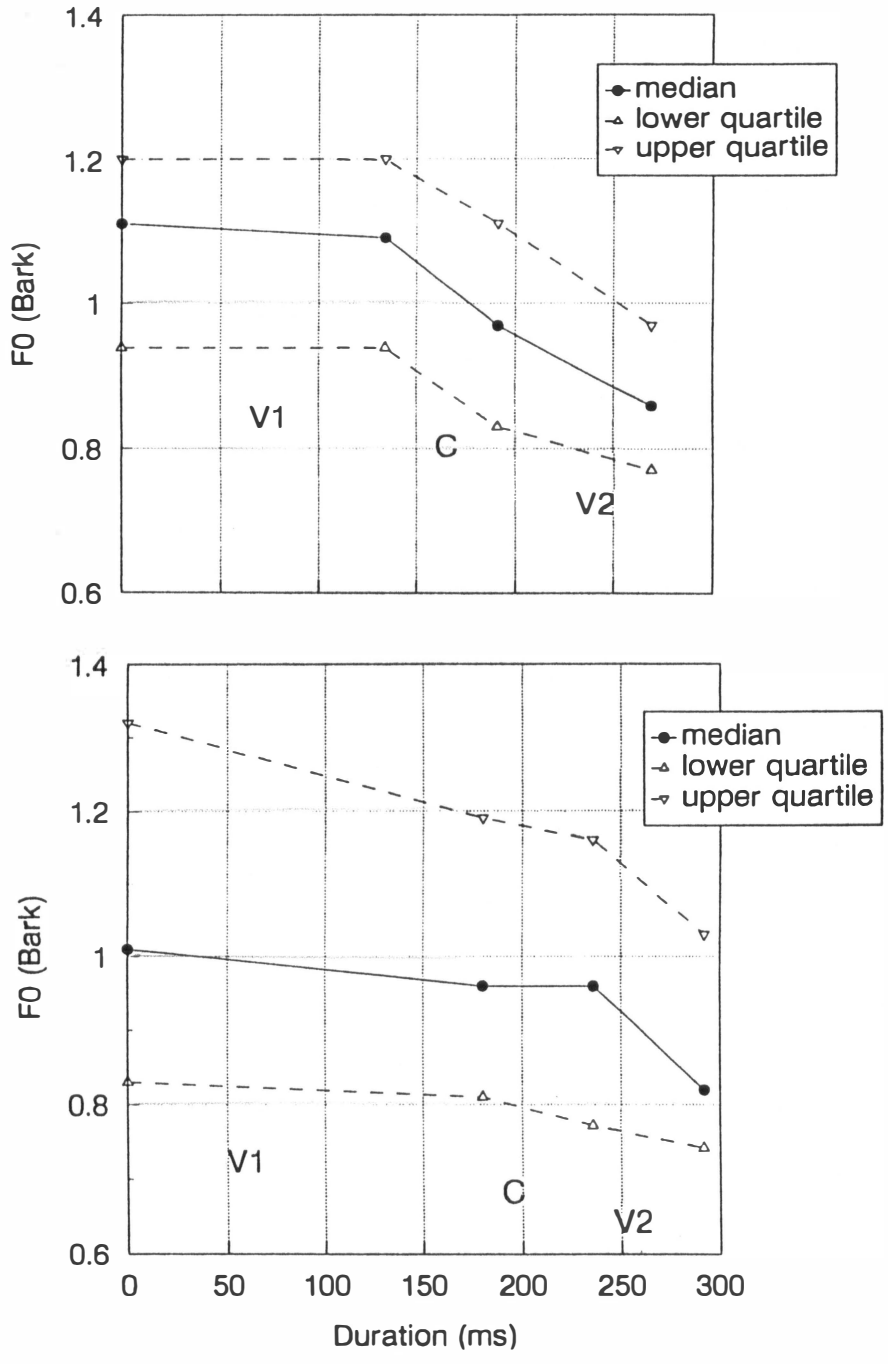


Figure 4. F0 movement from the beginning of V1 to the end of V2 in Q2 (N=45) **(top)** and Q3 words (N=64) **(bottom)**. The segment durations are median values.

The corresponding differences in F0-change, on the other hand, were not statistically significant, although there was a tendency for F0 to fall more during V1 in Q3 than in Q2 words[2] (Figs. 3 and 4). There was also a tendency for F0 to stay unchanged between the end of V1 and the beginning of V2 in Q3 words, and to fall in connection with Q2, but even here the difference was small and the variation considerable.

An F0 maximum within V1 was found in 27% of the Q2 words and in 59% of the Q3 words; Q1 words had no such maximum. The location of the maximum on the time axis was earlier in Q3 than in Q2 words: for Q2 median location of the maximum was at 55% of the entire duration of V1, for Q3 the corresponding value

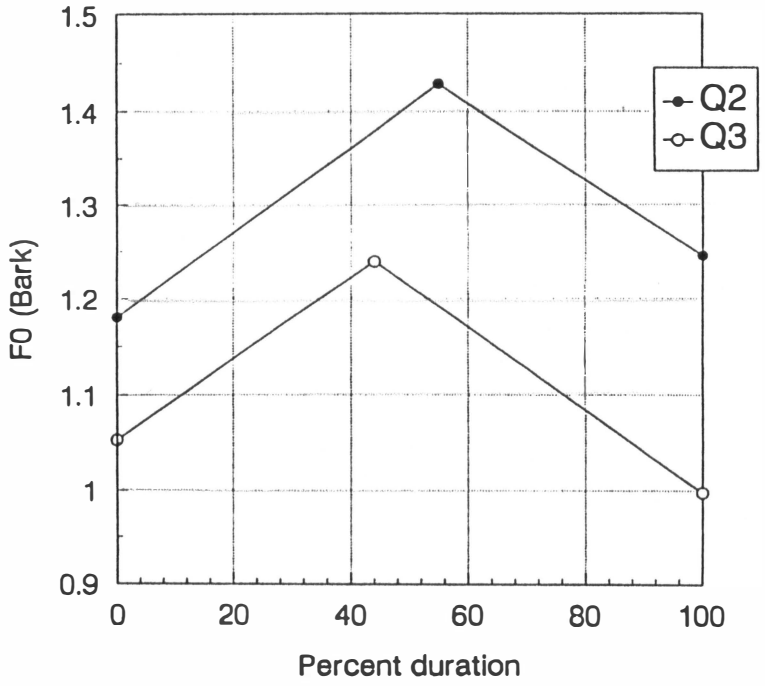


Figure 5. Schematic representation of the F0 contour in words that had an F0 peak within V1. For Q2 N=12, for Q3 N=37.

2) For a better correspondence to what is received by the human ear, the F0 frequencies were converted into Bark, using the formula presented by Traunmüller (1990): z = (26.81 f/(1960+f) –.53.

was at 44% (Fig. 5). However, the difference in location was not statistically significant.

Finally, the material was checked for possible correlations between the different acoustic correlates to quantity. No such correlation — positive or negative — was found.

4. Discussion

It was hypothesized in the Introduction that if an acoustic parameter is used as a primary correlate to phonological distinctions in a language, the freedom of the speakers of the language to use this parameter for other purposes will be restricted. This hypothesis, if true, would predict that the temporal correlates to Q1 and Q2, and at least one — temporal or tonal — correlate to the distinction between Q2 and Q3 would remain stable. That is, there should be some acoustic differences between quantities found in "laboratory speech" that remain stable even in a more spontaneous speaking style.

The results of this investigation showed that — for this one speaker — the most stable acoustic difference between quantities was temporal. Not only the duration ratios, but even the absolute durations of V1 and V2 were significantly different between quantity degrees. Although there was a certain overlap, it was small and involved few items. A comparison of these results with temporal data from words of the same form read in a carrier phrase by the same speaker (Eek 1975) showed, surprisingly, that the temporal differences between quantities were enhanced rather than weakened in a more natural speech style. A possible reason for this may be that, in conversational speech, the speaker may increase the duration of overlong syllables, for example, to signal emphasis.

The tonal correlates, on the other hand, were present only as a statistically not significant tendency. The stability of the temporal cues in Estonian therefore suggests that these cues may be the most crucial cues for signalling quantity. However, further investigation with more speakers is necessary.

Acknowledgement

Many thanks to Arvo Eek for acting as a speaker.

References

Eek, A. (1975): "Observation on the duration of some word structures II", *Estonian Papers in Phonetics* (EPP) 1975, Academy of Sciences of the Estonisn S.S.R., Tallinn, 7–51.

Eek, A. (1990): "Units of temporal organization and word accents in Estonian". *Linguistica Uralica XXVI*, 251–264.

Eek, A. and Help, T. (1987): "The interrelationship between phonological and phonetic sound changes: A great rhythm shift of old Estonian." *Proceedings of the 11th ICPhS,* Tallinn 1987, Vol. **6** pp. 218–233.

Engstrand (1992): "Durational correlates of quantity in Swedish and Finnish: Data from natural speech," *Fonetik '92* (Papers from the Sixth Swedish Phonetic Conference), Technical report 10, Gothenburg: Dept of Information Theory, Chalmers University of Technology, 47–50.

Krull, D. (1992): "Temporal and tonal correlates to quantity in Estonian", *PERILUS,* **XV**, Institute of Linguistics, Stockholm University.

Lehiste, I. (1960): "Segmental and syllabic quanity in Estonian", *American Studies in Uralic Linguistics.* Indiana University, Bloomington, Vol. **1**, 21–82.

Lehiste, I. (1970a): *Suprasegmentals.* MIT Press, Cambridge, Mass.

Lehiste, I. (1970b): "Experiments with synthetic speech concerning quantity in Estonian", *Congressus Tertius Internationalis Fenno-Ugristarum*, Pars I, Acta Linguistica Tallinn.

Lehiste, I, (1988): "Current debates concerning Estonian quantity." In *FUSAC '88: Proceedings of the Sixth Annual Meeting of the of the Fenno-Ugric Studies Association of Canada*, University press of America, Lanham, MD, 77–86.

Lindgren, R., Krull, D. and Engstrand, O. (1987): "Akustiska studier av fonetiskvariation i svenskan." In: P. Linell, V. Adelsvärd, T. Nilsson och P.A. Pettersson (eds.), *Svenskans beskrivniong **16**.* Förhandlingar för att dryfta frågor rörande svenskans beskrivning. Vol **2**, 326–338.

Traunmüller, H. (1983): "Analytical expressions for the tonotopical sensory scale", Journal of the Acoustical Society of America **88**, 97–100.

Velleman, P. and Hoaglin, D. (1981): *Applications, Basics, and Computing of Exploratory Data Analysis,* Duxbury Press, Boston, Mass.

Appendix

Median and upper/lower quartile values of duration and F0 data. For Q1, N=48; for Q2, N=45, and for Q3, N=64. An F0 maximum within V1 was found in 12 Q2 words and 38 Q3 words.

Table AI. Duration (ms) of V1, C, and V2, location of an F0 peak (if any) within V1 (ms from the beginning of the vowel), and the V1/V2 ratio.

	Quantity	V1 dur	Max loc	C dur	V2 dur	V1/V2
Upper quartile	I	96	–	51	131	.9
Median		83	–	44	107	.7
Lower quartile		68	–	34	99	.6
Upper quartile	II	158	118	71	97	2.1
Median		134	76	57	78	1.7
Lower quartile		114	49	39	65	1.5
Upper quartile	III	216	103	66	76	3.9
Median		180	87	56	56	3.2
Lower quartile		149	65	45	46	2.3

Table AII. F0 (Hz) at the beginning, peak (if any) and end of V1, and at the beginning and end of V2.

	Quantity	V1 init	Peak	V1 final	V2 init	V2 final
Upper qtl	I	130	–	139	136	121
Median		115	–	120	123	107
Lower qtl		105	–	105	108	95
Upper qtl	II	135	168	135	128	116
Median		128	155	126	117	107
Lower qtl		114	142	114	105	100
Upper qtl	III	145	162	134	132	121
Median		120	139	115	116	104
Lower qtl		105	125	103	100	98

Phonetic Experimental Research,
Institute of Linguistics,
University of Stockholm (PERILUS),
No. XVII, 1993, pp. 55–63

Sonority contrasts dominate young infants' vowel perception[1]

Francisco Lacerda

1. Introduction

Most earlier studies assessing young infants' perception of vocalic contrasts have concentrated on reporting successful discrimination of the vowel contrasts on which the infants were tested. Recently, however, a shift in focus appeared in infant speech perception studies. Instead of exclusively addressing the issue of how young infants sort speech sounds into meaningful phonemic categories, research has moved on to study the nature of the infants' internal representations of speech sounds. Some years ago, studies carried out by Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy and Mehler (1988), suggested that 2-month old infants were able to detect common denominator phonetic properties in series of speech sounds. In addition, work by Kuhl (1985; 1987) suggests that 6-month-old infants are able to perceive phonetic invariance in the presence of acoustic differences due to the speaker's sex and age. Thus, infants seem to be able to discriminate between speech sounds used distinctively in their ambient languages and also to recognise phonetic equivalence even in the presence of paralinguistic variability. Also cross-linguistic developmental studies indicate that, initially, language general speech perception processes are shaped into efficient language specific perceptual strategies, in many cases already by the end of the first year of life (Werker and Tees, 1984; Best, McRoberts and Sithole 1988; Werker and Polka, 1993). Recently, Kuhl, Williams, Lacerda, Stevens and Lindblom (1992) reported that infants develop language specific vowel prototypes as early as by 6-months of age. According to the prototype theory proposed by Kuhl (1991), prototypes can be pictured as "perceptual magnets". When variants of a prototypical vowel are heard they are assimilated to the prototype reference. Thus, the theory predicts that there will be more discrimination errors when variants are located in the neighbourhood of a prototype (more generalisations of the prototype) than for variants of a non-prototypical sound. This type of pattern was, in fact, observed in the cross-language study carried out by Kuhl et al. (1992). American infants listening to variants of the prototypical

1) Paper presented at the 125th meeting of the Acoustic Society of America,17–22 May 1993, Ottawa, Canada.

American /i/-vowel made more discrimination errors than when listening to the variants of the foreign Swedish /y/-vowel, whereas the discrimination pattern obtained from the Swedish infants was the other way around. Swedish infants produced significantly more generalisations when listening to variants of the Swedish /y/ prototype than when listening to variants of the American /i/ prototype. Hence, in terms of the prototype theory, Kuhl et. al's (1992) results suggest that infants develop prototypes for the vowels in their native language already by 6 months of age and that these prototypes influence perception of neighbouring vowel sounds in the way described by the analogy with the perceptual magnets.

The notion that perception of vowel sounds may be structured around the development of language specific prototypes is of great importance for the insight it provides into the organisation and development of speech perception strategies. In this paper we describe the results of a series of experiments in which perceptual discrimination of vowels in the low region of the vowel space was studied with Swedish infants. The contrast examined was that between [ɑ] and [a].

The Swedish vowel system makes extensive use of quantity contrasts between vowels. One of these contrasts involves the long and short vowel phonemes in the low region of the vowel space, /a:/ and /a/. In the Swedish dialect of the Stockholm area (and in most Swedish dialects) this quantity contrast is enhanced by a correlated quality contrast, resulting in a phonetic [ɑ] vs. [a] distinction. Thus, by

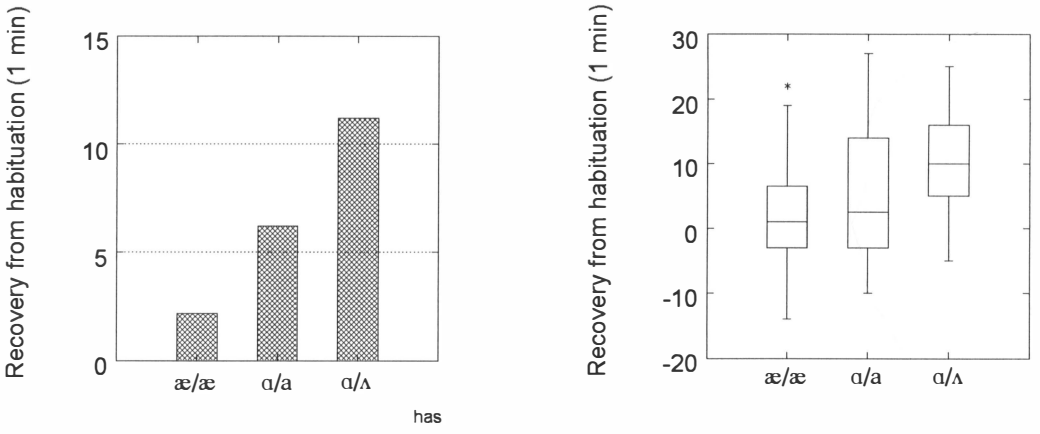


Figure 1. Recovery from habituation. **(a)** Average increment in the sucking frequency for the no-change contrasts and for two discriminable vowel contrasts. **(b)** Box plots for the same data.

investigating the Swedish infants' ability to discriminate between the [ɑ] and the [a] sounds, we are trying to assess the infants' capacity to focus on a quality difference that conveys relevant phonemic information in the infants' ambient language.

A recent study by Lacerda (1992a) indicated that 2–3 month-old infants were apparently unable to discriminate between [ɑ] and [a] synthetic vowels that differed only in their F_2 values. This null result is informative in the context of other speech perception experiments that have examined young infants' ability to discriminate vowel contrasts of similar magnitude. Along with the successful discrimination of an [a] vs. [ʌ] contrast reported by Jusczyk, Bertoncini, Bijeljac-Babic, Kennedy and Mehler (1990), which is a contrast involving approximately the formant distance (in Bark) that used by Lacerda (1992a), it suggests that infants may preferentially focus on vowel contrasts involving F_1 differences.

This paper summarises the results of an investigation designed to assess the possibility of differential sensitivity to changes in F_1 and F_2. The basic experimental methods used in this investigation have been described in Lacerda (1991; 1992a).

2. Differential formant sensitivity in 2–3 month-old infants

To examine the possibility that the differences in discrimination performance observed between Jusczyk et al's. (1990), and Lacerda's (1992a) results might be due to procedural differences, like the use of different recovery measures in the high amplitude sucking procedure (Lacerda, 1993) or discrepancies in the phonetic detail of the stimuli used in those experiments, we extended our initial set of stimuli to include a [ɑ]/[ʌ] contrast differing only in F_1 and by the same amount (in Bark) that F_2 differed in the original [ɑ]/[a] contrast.

The average +/–1 minute recovery from habituation for the F_1 and F_2 contrasts are shown in fig. 1a, along with the data from a no-change control condition. Fig. 1b displays the same data as box diagrams which provides a better indication of the different ranges of the recovery from habituation results. The results are based on 33 measures for the [ɑ]/[a] contrast, 26 measures for the [ɑ]/[ʌ] contrast and 29 measures from the control group. A Kolmogorov–Smirnov two sample test indicated a non-significant difference between the [ɑ]/[a] and the control group ($p<0.551$) but a highly significant difference between the [ɑ]/[ʌ] and the control group ($p<0.005$).[2]

2) A parametric analysis of variance indicates an overall significance difference between the three conditions (two experimental and one control), $p<0.030$, $F(2, 85)=3.639$. The analysis of the contrasts between each experimental condition and the control condition agrees with the results of the non-parametric analysis, $p<0.204$, $F(1, 85)=1.642$ and $p<0.008$, $F(1, 85)=7.270$, respectively.

Thus, the results from the discrimination tests carried out with equivalent F_1 and F_2 contrasts indicate that 2–3 month old infants have a better discrimination performance when the vowel sounds contrast along the opening dimension than when they contrast along the front–back dimension (Lacerda, 1992b).

3. Differential formant sensitivity in 6 to 12 month-old infants

Acoustic correlates of variation in a vowel's opening degree are both variations in F_1 *and* in the overall vowel intensity. Therefore, because infants tested with the classical high-amplitude sucking paradigm are exposed to a single stimulus during the habituation phase which is subsequently replaced by a single new stimulus, it is important to investigate whether infants might be simply attending to intensity cues to discriminate between the vowels contrasting along the high/low dimension. Strictly, it cannot be excluded that infants might respond to intensity changes in the stimuli, even though intensity variations associated with different opening degrees are normally ignored by adult listeners (Ladefoged, 1967).

Thus, to rule out the possibility of a systematic influence of the intensity levels upon the discrimination results, we carried out new experiments using a different speech synthesis method, parallel speech synthesis. In addition, the experiments were run with older infants to investigate if the contrast had already been established by 6 months of age. For these older infants a different infant speech perception technique was used.

The infants were tested with the head-turn paradigm. They were first trained to respond to a single stimulus difference involving 2 Bark shifts in both F_1 and F_2. After successful training the infants were submitted to a criterion phase in which both test and catch trials were presented. To meet the criterion, the infants had to provide 7 correct responses out of 8 in a row. Under the criterion phase the

Table I. Specification of the stimuli and their use in each of the test phases.

Stimulus	Use	F1(Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)	DF
v00	reference	733	1067	2467	3200	—
v11	test	503	1067	2467	3200	F1: 1 Bark
v21	test	733	1467	2467	3200	F2: 1 Bark
v31	test	503	1467	2467	3200	F1, F2: 1 Bark
v12	criterion	400	1067	2467	3200	F1: 2 Bark
v22	criterion	733	1695	2467	3200	F2: 2 Bark
w32	crit.+train.	400	1695	2467	3200	F1, F2: 2 Bark

differences between the discriminable stimuli were still 2 Bark but they were conveyed either by a frequency shift in F_1, in F_2 or in both formants (the same contrast that had been used in the training phase). The infants who met the criterion proceeded to the test phase. In this phase contrasts involving 1 Bark shifts in the formant frequency were used. The contrasts were produced by either F_1, F_2 or both formants.

The vowels used in these experiments were generated by a parallel synthesiser to minimise the dependence of the overall intensity on the formant frequencies. The formant frequencies and the stimuli used in each of the test phases are specified in table I.

Figure 2 shows the average discrimination scores obtained when the test stimuli, v11, v21 and v31, were compared to the reference stimulus, v00. The results displayed correspond to the data obtained from 18 infants whose false-alarm rates were less than 0.35.

An analysis of variance carried out on the discrimination scores obtained when these 18 infants listened to v11, v21 and v31 against the v00 reference, revealed an overall within subjects significance level of $p<0.000$, $F(3,51)=16.376$. An analysis of the difference between the scores obtained for the discriminations along F_1 and those along F_2 also revealed a significant difference ($p<0.001$, $F(1,17)=14.695$, for the comparison between v11 and v21). To evaluate the effect

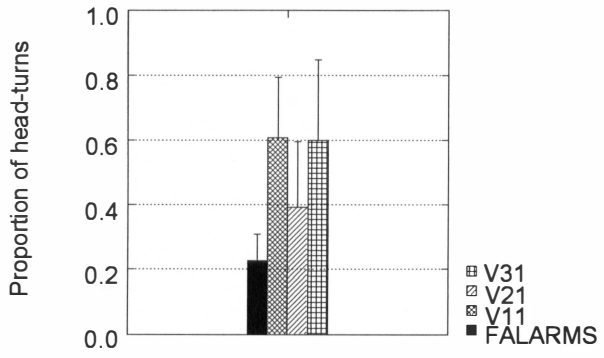
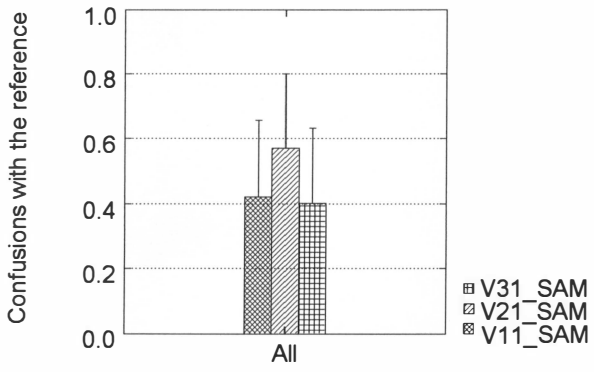


Figure 2. Average percent of head-turns observed for the control situation (false alarms) and for the three target vowels.

of the false alarms on the significance scores, all the available data was re-analysed using the individual false alarm rates as a covariant. This variance analysis was based on 29 infants and gave the following within subjects results:
1. Significant overall differences between the discrimination scores for v11, v21 and v31, $p<0.014$, $F(2,54)=4.636$.



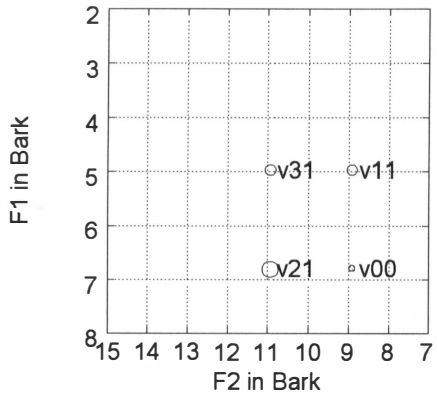


Figure 3 (top). Average percent of confusions of the each target vowel with the reference vowel.

Figure 4 (bottom). Illustration of the proportion of variants (filled circles) that were assimilated to the reference vowel (open circle).

2. Non-significant interaction between the dependent variables and the covariant, $p<0.252$, $F(2,54)=1.414$.
3. Significant difference between the discriminations along F_1 and those along F_2, $p<0.007$, $F(2,27)=5.926$.

An additional evaluation of the robustness of the data was made using a non-parametric test. The data from the 18 infants whose false alarm rates were below 0.35 was submitted to a Wilcoxon test that compared the discrimination scores of v11 with those of v21. The difference between discriminations along F_1 and F_2 was highly significant ($p<0.003$, two-sided).

The generalisation errors committed are illustrated in figure 3. It shows that more generalisations from the reference stimulus were made along F_2 (front/back dimension) than along F_1 (openness dimension). Finally, figure 4 displays the test stimuli in the F_1xF_2 plane and indicates by the size of the filled symbols the percentage of assimilations to the reference (open circle) by the size of the solid circles.

5. Discussion

The results from the high-amplitude sucking experiments and from the head-turn experiments indicate an advantage in the discriminations of the vowel sounds involving sonority contrasts in spite of the procedural differences.

Apparently, the present results indicate that infants do not focus equally on all acoustic dimensions used to distinguish between vowel sounds. Rather, everything else being equal, they seem to process more easily contrasts involving F_1 differences than F_2 differences.

This asymmetry also seems to extend to other regions of the vowel space such as the region of the front high vowels /i/ and /y/ (Kuhl, pers. comm.) although that asymmetry is not as clear as the one obtained for the low–back vowels. Kuhl's analysis of the discrimination data from American and Swedish infants listening to variants of /i/ and /y/ indicates that Swedish infants have a significant advantage for discriminations involving variations in F_1 as compared to those involving variation in F_2, [$F(1,15)=7.3$, $p < 0.016$ for Swedish infants listening to Swedish /y/ and $F(1,15)=14.78$, $p < 0.002$ for Swedish infants listening to American /i/]. The American infants reveal only a tendency to asymmetry. At this stage, additional data and replications of the present observations are necessary to establish the extent of sonority advantage over chromaticity.

From a developmental perspective the present results seem to indicate that the auditory systems may be preferentially tuned to pick up vowel contrasts involving F_1.

The particular case of the [ɑ]/[a] contrast underlines this perceptual advantage. In fact, since the [ɑ]/[a] contrast is used in the Swedish infants' ambient language, it could be expected, on the basis of Kuhl et al.'s (1992) results, that Swedish 6-month-olds would also be able to discriminate between those two vowel qualities. On the other hand, the phonological status of the stimuli used by Kuhl et al. (1992) is different from the ones used here. Whereas /i/ and /y/ represent different phonemes in Swedish, [ɑ] and [a] do not. In the present case, the vowel quality difference is only part of the quantity contrast between /a/ and /a/. At a phonetic level, however, [ɑ] and [a] occur in the Swedish infants' ambient language and it might be expected that the infants learn to attend to the quality differences as a cue to the quantity contrast occurring in their native language. That does not seem to be the case. Swedish infants under 12 months of age seem to be simply unable to detect the [ɑ]/[a] contrast, whereas they discriminate between [ɑ] and [ʌ].

The fact that the infants showed a dominance in the discrimination of the sonority contrast, even though their ambient language would lead them to focus on the chromaticity contrast, suggests that sonority contrasts may be easier to perceive than equivalent chromaticity contrasts. This is in accordance with the observation that natural languages tend to prefer high-low contrasts over front-back contrasts.

A search through the UPSID database (Maddieson, 1984) supports this notion. The front/back contrasts of the type used in the perception tests described here are observed in only about 1% of the UPSID languages. High/low contrasts are much more frequent. In addition, when vowel systems use front/back contrasts for a constant degree of vowel opening, the contrast is usually accompained by a default rounding of the back vowels, as if an additional acoustic property would be necessary to mark the front/back difference in an otherwise narrow perceptual space.

Dominance of the sonority contrasts in vowels can also be related to babbling. When the infant opens and closes the jaw, on which the tongue rests, the resulting vowel sounds fall primarily along the sonority dimension (MacNeilage and Davis, 1990). Thus, babbling may further enhance the focus on the high/low vowel contrasts.

Finally, the interpretation of the present results in terms of the prototype theory points to a hierarchical development of vowel prototypes. From the perspective of the prototype theory, the present data indicate that exposure to ambient language does not result in the simultaneous establishment of all vowel prototypes.

Acknowledgments

The research on which this report is based is supported by The Bank of Sweden Tercentenary Foundation (grant 90–0150).

References

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P., Kennedy, L. and Mehler, J. (1988): "An investigation of young infants' perceptual representations of speech sounds", *J. Exp. Psychology: General,* **117**, 21–33.

Best, C., McRoberts, G. and Sithole, N. (1988): "Examination of perceptual reorganization for non-native speech contrasts: Zulu click discrimination by English-speaking adults and infant", *J. Exp. Psychology: Human Perception and Performance,* **14**, 345–360.

Jusczyk, P., Bertoncini, J., Bijeljac-Babic, R., Kennedy, L. and Mehler, J. (1990): "The role of attention in speech perception by young infants", *Cognitive Development,* **5**, 265–286.

Kuhl, P. (1985): "Categorization of speech by infants", in J. Mehler and R. Fox (Eds.), *Neonate Cognition: Beyond the blooming buzzing confusion*, Erlbaum, Hillsdale, N.J., 231–262.

Kuhl, P. (1987): "The special-mechanisms debate in speech research: categorization tests on animals and infants", in S. Harnad (Ed.), *Categorical Perception: The groundwork of cognition*, Cambridge Univ. Press, New York, 355–386.

Kuhl, P. (1991): "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not", *Perception and Psychophysics,* **50**, 93–107.

Kuhl, P., Williams, K., Lacerda, F., Stevens, K. and Lindblom, B. (1992): "Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age", *Science,* **255**, 606–608.

Lacerda, F. (1991): "Perception of CV-utterances by young infants: Pilot study using the high-amplitude sucking technique", *PERILUS,* **XII**, 161–177.

Lacerda, F. (1992a): "Young infant's discrimination of confusable speech signals" in M. Schouten (Ed.) *The Auditory Processing of Speech: From Sounds to Words*, V. van Heuven and L. Pols (Eds.), Speech Research 10, Mouton de Gruyter, Berlin, 229–238.

Lacerda, F. (1992b): "Young infants prefer high/low vowel contrasts", in *Tech. Report,* **10**, Dep. of Information Theory, Chalmers University of Technology, Göteborg, 75–78.

Lacerda, F. (1993): "Dependence of High-Amplitude Sucking discrimination results on the pre- and post-shift window duration" (in preparation).

Ladefoged, P. (1967): *Three Areas of Experimental Phonetics*, Oxford University Press.

MacNeilage, P. and Davis, B. (1990): "Acquisition of speech production: The achievement of segmental independence", in Hardcastle, W. and Marchal, A. (Eds.), *Speech production and speech modelling*, Kluwer: Dordrecht, 55–68.

Maddieson (1984): *Patterns of sound,* Cambridge: Cambridge University Press.

Werker, J. and Tees, R. (1984): "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life", *Infant Behavior and Development,* **7**, 49–63.

Werker, J. and Polka, L. (1993): "Developmental changes in speech perception: New challenges and new directions", *Journal of Phonetics,* **21**, 83–101.

Phonetic Experimental Research,
Institute of Linguistics,
University of Stockholm (PERILUS),
No. XVII, 1993, pp. 65–74

Word accent 2 in child directed speech: A pilot study[1]

Ulla Sundberg

Abstract

The tonal characteristics of disyllabic accent 2 words in child directed speech were studied and compared to adult directed speech. The fall parameter in the first, primary stressed syllable and the rise parameter in the secondary stressed syllable were measured. Words marked with sentence accent directed to a three month old infant had wider pitch excursions than those directed to an adult listener. Sentence accent was found to have a great impact on the F0 contour.

1. Introduction

Child Directed Speech (CDS) is an important linguistic input to infants as they are exposed to it during one of the most dynamic phases of their language acquisition period (Vihman, et al., in press).

The prosodic characteristics of CDS have been examined in a number of studies. Results have revealed that mothers of newborn babies use higher pitch, wider pitch excursions and more prosodic repetitions when they speak to their infants than when they speak to adults (Fernald & Simon, 1984). Cross-linguistic investigations of CDS have shown a consistent use of prosodic and other types of linguistic modification in different kinds of languages (Fernald et al.,1989, Grieser & Kuhl, 1988).

Previous investigations have mainly focused on the sentence and phrase levels. It is likely that the prosodic modifications found on those levels also affect the intonation contour in the word domain (Grieser & Kuhl, 1988). The goal of the present investigation was to analyze modifications at the word level by collecting quantitative data on the tonal characteristics of the Swedish word accent 2 in disyllabic words in CDS.

---

1) This is a revised version of an article published in *Nordic Prosody VI, Papers from a symposium*, Stockholm: Almqvist & Wiksell International, 1993, 199–206.

2. Background

In Central Standard Swedish the characteristic F0 contour of disyllabic accent 2, or grave accent, words is two-peaked. The primary stressed syllable is marked by an F0 fall. Provided the words have sentence accent (i.e., are pronounced with emphasis or are in focus), the secondary stressed syllable is marked by an F0 rise (Bruce, 1977; Engstrand, 1989). In contrast, accent 1 (acute) words lack an F0 fall on the primary stressed syllable. When in focus, such words display an F0 rise on that syllable.

Engstrand (1989) investigated all disyllabic accent 2 words in adult directed spontaneous speech. The F0 fall turned out to be a robust characteristic of these types of words across speaking rates and styles. Later, Engstrand et al (1991)

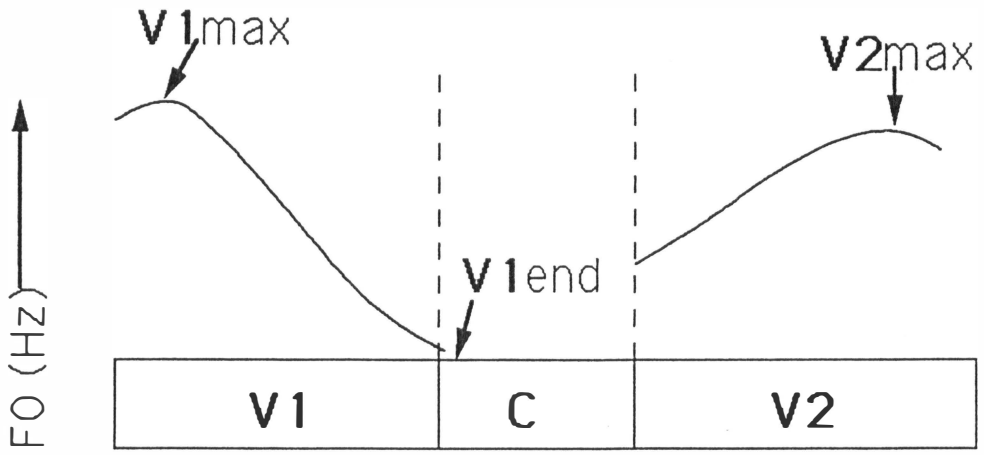


Figure 1. Stylized F0 contour of a disyllabic accent 2 word with sentence accent. The bar represents vowel and consonant segments.

analyzed accent 2 word candidates in vocalizations and babbling from 17 month old children. Their data showed that, on the average, the Swedish children displayed higher F0 values in the second syllable than did the American children. There was no significant difference in the behavior of F0 in the first syllable.These results raises the question regarding the nature of the childrens linguistic input. What are the phonetic characteristics ot these types of words in speech directed to children? The purpose of this study is to investigate 1) whether the secondary F0 rise is more prominent in CDS than in ADS, and 2) whether the secondary F0 rise is influenced more than the primary F0 fall in CDS. If confirmed, and in view of Engstrands et ats results, this might suggest a link between CDS as a linguistic input to the child and the childs acquisition process.

3. Method

In order to obtain comparable speech samples from CDS and *Adult Directed Speech* (ADS) a mother, speaking Central Standard Swedish interacted with her three months old son in a sound isolated booth. The infant sat in a baby seat opposite to its mother. She was asked to play with her son the way she would normally do at home. To elicit disyllabic words with accent 2 and accent 1 the investigator had selected some toys with names consisting of such words. This part of the session lasted about ten minutes. The investigator then entered the booth and talked informally with the mother about the infants' reactions to the toys, the babies favourite toys at home, baby clothing and so on.

The mother's speech was recorded using a DAT tape recorder. The microphone was a Sennheiser Red Dot Mic mounted on a head set so as to ensure a constant distance to the mothers mouth. Also, the session was videotaped.

Disyllabic accent 2 words were identified in the mother's CDS and ADS. The words that were judged by the investigator to be the most prominent in each utterance, were selected for analysis. In what follows, these words are referred to as having sentence accent.

The same method for determining the Fall and Rise parameters (see below) was used as in Engstrand et al (1991). F0 was measured at a) the acoustic onset of the first vowel, b) the F0 turning point, c) the acoustic offset of the first vowel, d) the acoustic onset the second vowel e) the F0 maximum, f) the acoustic offset of the second vowel. The Fall parameter was defined as the F0 difference between turning point and offset in the first vowel. The Rise parameter was defined as the F0 difference between the turning point of the second vowel and the offset of the preceding first vowel (see figure 1). Both these F0 differences were expressed in semitones. The acoustic analysis was done with the Soundswell computer program (S. Ternström, Royal Institute of Technology, Stockholm). This program provides

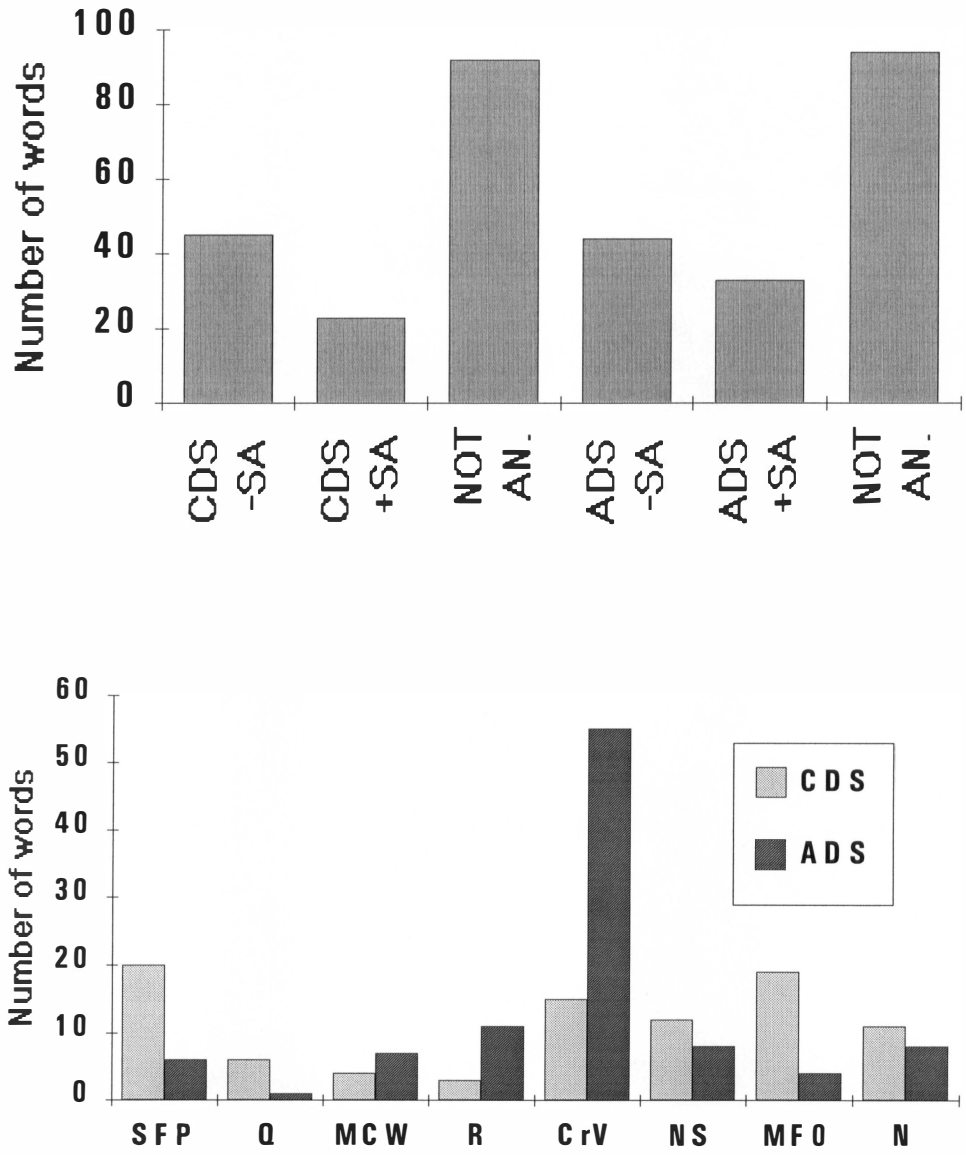


Figure 2. (a) Number of analyzed and not analyzed words with and without sentence accent (+ and – SA, respectively) in CDS and ADS. **(b)** Causes for discarding words. SFP = sentence final position; Q = question; MCW = morphologically complex word; R = reduced articulation; CrV = creaky voice; NS = impossible to locate segment boundaries; MF0 = F0 impossible to measure; N = Noisy recording.

a simultaneous display of spectrogram, F0 and speech wave signal. The analysis was carried out on all disyllabic words with accent 2 that appeared in CDS and ADS. A total of 68 CDS words were analyzed, 45 without sentence accent and 23 with sentence accent; in the ADS 77 words were analyzed, 44 without sentence accent and 33 with sentence accent. (See Figure 2a).

As illustrated in Figure 2b, several disyllabic accent 2 words were not included in the analysis. For example, the F0 contour of words in sentence final positions and in questions are heavily influenced by the F0 changes associated with these particular conditions. Compounds were also excluded since these words often contain medial consonant clusters that may influence the F0 contour.

In addition, F0 extraction turned out to be impossible in a number of words. There were several reasons for this difficulty. First, in some cases voice source mechanisms, e.g. creaky voice, or noise eliminated the possibility of obtaining reliable F0 data. Notably, creaky voice occurred much more often in ADS than in CDS. Also, some of the words were heavily reduced; this was more common in ADS than in CDS. Moreover, it was impossible to make a reliable segmentation in some cases. In CDS many words had a weak fundamental in the second and/or the first syllable, since the mother often spoke with a very soft or even whispering voice. For these reasons, 92 words from CDS and 94 words from ADS had to be discarded.

4. Results

As illustrated in Figures 3 a and b the dispersion was generally rather small, particularly for the Fall parameter, amounting to no more than one semitone, approximately. However, the Rise parameter showed a greater dispersion, especially in CDS and with regard to words marked by sentence accent.

The measurements were treated statistically using ANOVA. The results are listed in Tables I and II.

Comparison between CDS and ADS disregarding sentence accent, showed highly significant differences with respect to both the Fall and the Rise parameters, see Figure 4; the F0 changes were clearly greater in CDS than in ADS.

With regard to the words *with* sentence accent the Fall parameter did not differ significantly between CDS and ADS. However, the Rise parameter was significantly greater in CDS. In words *without* sentence accent no significant differences could be observed between CDS and ADS.

In CDS, sentence accent resulted in great F0 changes, both in the Fall and, in particular, the Rise parameter, see Figure 5. In ADS, on the other hand, sentence accent did not result in any statistically significant increase of the amplitudes of the F0 gestures. This suggests that the variation of F0 plays an important role in

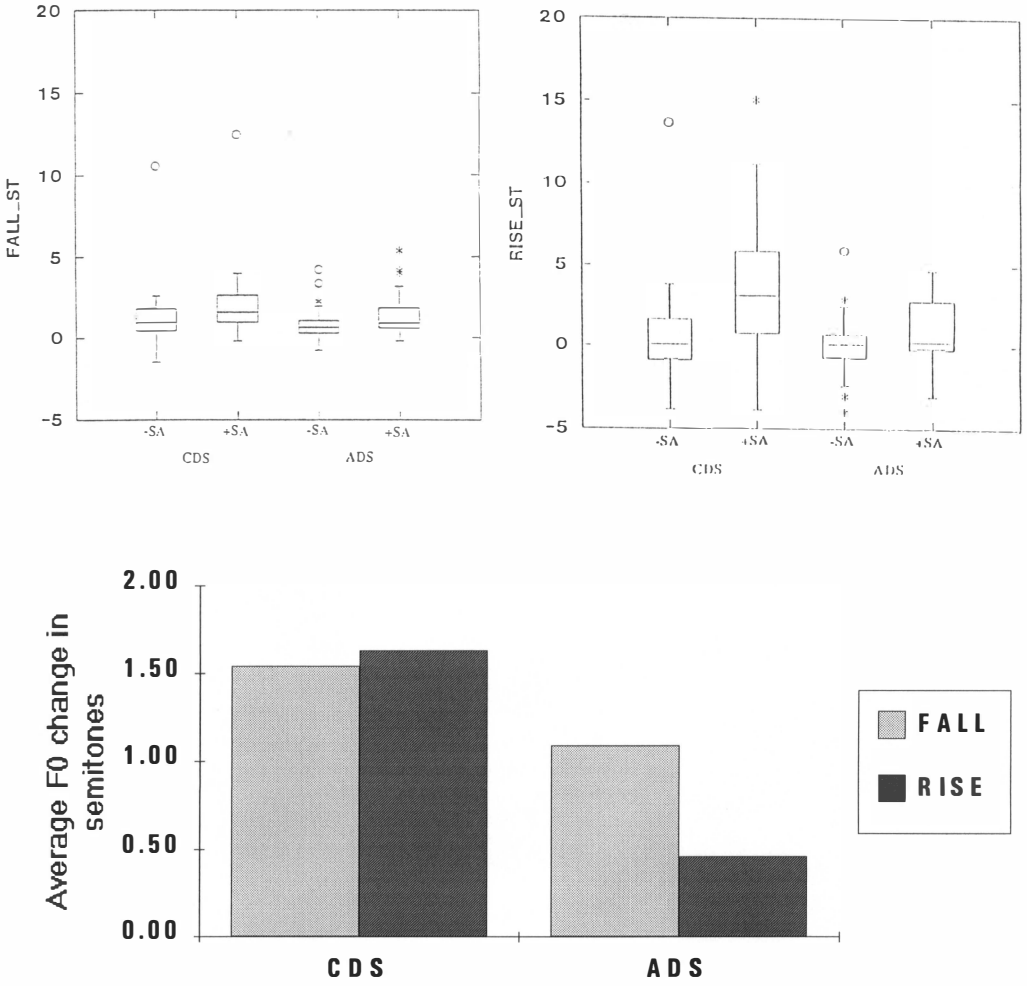


Figure 3, top. (a) Density plot for F0 values for the Fall parameter, expressed in semitones. The boxes contain 50 % of the values. The whiskers show the range of values that falls within 1.5 Hspreads of the hinges. (Hspread is comparable to the midrange.) The stars and circles represent outside values. The horizontal lines in the boxes mark the median value. **(b)** Density plot for F0 values for the Rise parameter, expressed in semitones.

Figure 4, bottom. Average F0 change in semitones (st) for words, disregarding sentence accent CDS and ADS.

signaling sentence accent in CDS. It is interesting that in both speech styles a F0 rise occurred instead of the expected F0 fall on the first syllable in some of the words, see Figure 6. Similarly, and much more frequently, a Fall replaced the expected F0 rise in the second syllable in a number of cases. In CDS *without* sentence accent such cases were not unusual: 42% (19 cases out of 45) of the words were produced with such inverted F0 contours while for words *with* sentence accent this happened in 22% (5 cases out of 23) of the words. The corresponding occurrences for ADS were 45% and 39% (20 cases out of 44 and 13 cases out of 33), respectively.

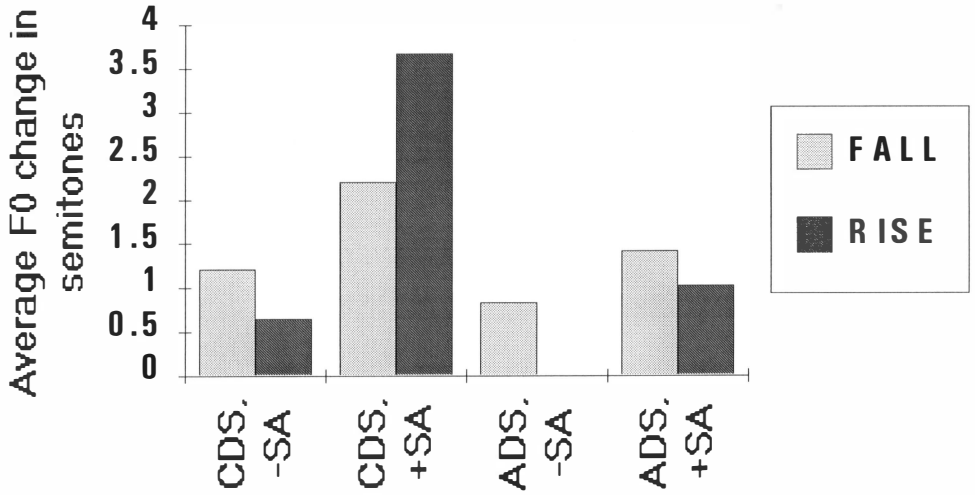
These data demonstrate a considerable variability in the tonal realization of spontaneous speech and are very different from those obtained under well-controlled laboratory conditions.

Table I. Mean F0 change in semitones in CDS and ADS disregarding sign and sentence accent.

	F0-change	FALL	RISE
CDS	1.58	1.54	1.63
ADS	0.77	1.09	0.46
Difference	0.808*	0.45	1.16

Table II. Means (M) and Standard Deviations (SD) in semitones for the various experimental parameters in child and adult directed speech (CDS and ADS) with and without sentence accents (SA).

	+SA					–SA					+/–SA	
	FALL		RISE		Diff	FALL		RISE		Diff	Diff	
	M	SD	M	SD		M	SD	M	SD		FALL	RISE
CDS	2.2	2.4	3.7	4.2	1.5	1	1.2	0.7	1.8	0.3	0.7*	1.2*
ADS	1.4	1.3	1	1	0.4	0.8	0.9	0	1.7	0.8	0.6	1
Diff	0.8		2.6*			0.2		0.6				



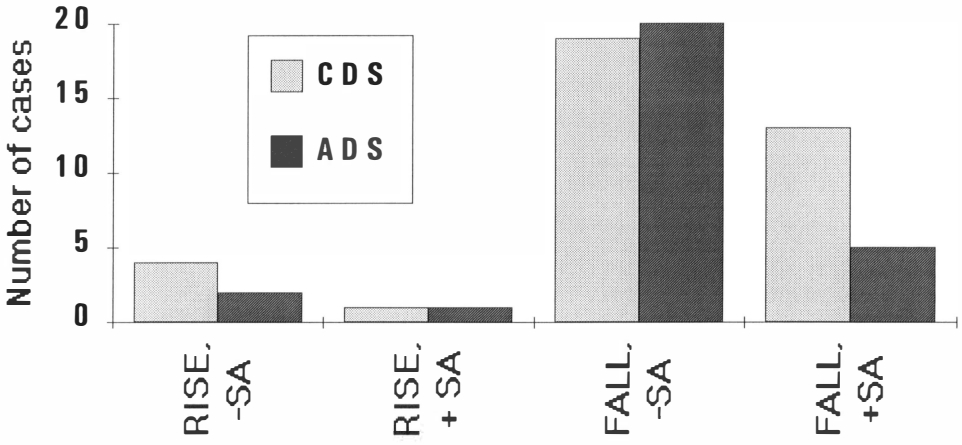


Figure 5, top. Average F0 change in words with and without sentence accent (+ and – SA) found in CSD and ADS.

Figure 6, bottom. Number of words with inverted F0 contours with and without sentence accent (+ and – SA) found in CDS and ADS.

5. Discussion

The results show how the mother drastically changed her speech style when addressing her infant. The main difference was the exaggerated tonal characteristics in disyllabic accent 2 words. The highly significant differences between words with and without sentence accent were not unexpected, especially for the Rise parameter. Several earlier investigations have shown that the Rise parameter is very sensitive to sentence accent (Bruce 1977). Words *without* sentence accent showed a much greater average Rise parameter in CDS than in ADS. These results confirm the first hypothesis posed above, i.e., the secondary F0 rise is more prominent in CDS than in ADS.

In ADS the Fall parameter average was greater than the Rise parameter average in words *with* sentence accent. This finding is coherent with earlier investigations, for example Engstrand (1989). He found that the Fall parameter in the primary stressed syllable served as a very robust marking of accent 2 in adult directed spontaneous speech. This is also in agreement with the low standard deviations found for the Fall parameter in the present study. The second hypothesis, that the secondary F0 rise is influenced more than the primary F0 fall in CDS, was also supported. In particular, the F0 fall in the first syllable was seldom replaced by an F0 rise, whereas the F0 rise in the second syllable was realized as a F0 fall in many cases.

The results from this investigation may, in view of those reported by Engstrand et al (1991), suggest that CDS is an important factor as linguistic input during the childs language acquisition process. However, the results from the present study must be regarded as preliminary since the speech of only one mother has been investigated. Analysis of data from more informants is in progress.

Acknowledgements

I wish to thank Francisco Lacerda for his kind assistance with the statistical computations, and Olle Engstrand for many helpful suggestions on the manuscript.

References

Bruce, G. (1977): "Swedish word accents in sentence perspective", *Travaux de L'Institut de Linguistique de Lund* ,**12**, Lund: Gleerup.

Engstrand, O. (1989): "Phonetic features of the acute and grave word accents: data from spontaneous speech", *PERILUS* **X**, Stockholm University, 13–37.

Engstrand, O., Williams, K. and Strömqvist, S. (1991): "Acquisition of the Swedish tonal word accent contrast", *PERILUS* **XII**, Stockholm University, 189–193.

Fernald, A. (1985): "Four-month-old infants prefer to listen to motherese", *Infant Behaviour and Development* **8**, 181–195.

Fernald, A., Kuhl, P. (1987): "Acoustic determinants of infant preference for motherese speech", *Infant Behaviour and Development* **10**, 279–293.

Fernald, A., Simon, T. (1984): "Expanded intonation contours in mothers speech to newborns", *Developmental Psychology* **20**, No. 1, 104–113.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardie, B. & Fukui, I., (1989): "A cross-language study of prosodic modifications in mothers and fathers speech to preverbal infants", *Journal of Child Language* **16**, 477–501.

Grieser, D.L. Kuhl, P.K. (1988): "Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese,*Developmental Psychology* **24**, No. 1, 14–20.

Ternström, S. Soundswell signal workstation software, Soundswell Music Acoustics, Solängsvägen 46, S-191 54 Sollentuna.

Vihman, M.M., de Boysson-Bardie, B., Durand, C., Kay, E., & Sundberg, U. (1993): "External sources of individual differenses? A cross-linguistic analysis of the phonetics of mothersspeech to one-year-old children", *Developmental Psychology*, in press.

Phonetic Experimental Research,
Institute of Linguistics,
University of Stockholm (PERILUS),
No. XVII, 1993, pp. 75–79

Swedish tonal word accent 2 in child directed speech — a pilot study of tonal and temporal characteristics

Ulla Sundberg and Francicso Lacerda

Abstract

Tonal and temporal aspects of disyllabic accent 2 words in child directed speech and adult directed speech were investigated. The speed of the F0 change in the first syllable F0 fall and the second syllable F0 rise in words with and without sentence accent were compared in both speech situations. In child directed speech words with sentence accent had a greater and faster F0 change than words without sentence accent. In adult directed speech the contrast was marked by duration. The results suggest a differentiated use of F0 and duration in child directed and adult directed speech.

2. Background

In a previous study (Sundberg,1993, this issue) we analyzed the amplitudes of the F0 excursions associated with accent 2 in child directed spech (CDS) and adult directed speech (ADS).The results showed that the subject drastically changed her speech style when addressing her three months old infant. The main effect was the exaggeration of the tonal characteristics in disyllabic accent 2 words in CDS as compared to ADS. In words *with* sentence accent the F0 rise was significantly greater in CDS than in ADS. However, temporal effects may also play a role. In particular, the previous investigation showed no difference in the F0 excursions between CDS and ADS for words *without* sentence accent. This poses the question whether there is a difference with respect to durations; one and the same F0 difference may be perceptually different depending on the rate of F0 change. Also, the tonal difference found in the words with sentence accent may be accompanied by differences in temporal characteristics. The purpose of the present investigation was to analyze the rate of pitch change associated with the differences in the F0 excursions previously observed betwen ADS and CDS.

3. Method

Durations were measured on a subset of the disyllabic accent 2 words used in the previous study. Segmentation and measurement points are described in that paper.To obtain comparable sample sizes of the four variables involved (CDS and ADS words, and words with and without sentence accent) as well as relevant measuring points, words were selected which met the following two criteria: (1) the highest pitch of the first vowel occurred at its onset; (2) the highest pitch of the second vowel did not occur at its onset but later. In this way, duration of the F0 fall in the primary stressed syllable comprises the first vowel segment, and duration of the F0 rise in the secondary stressed syllable comprises the medial consonant up to the point of the F0 maximum of the second vowel.

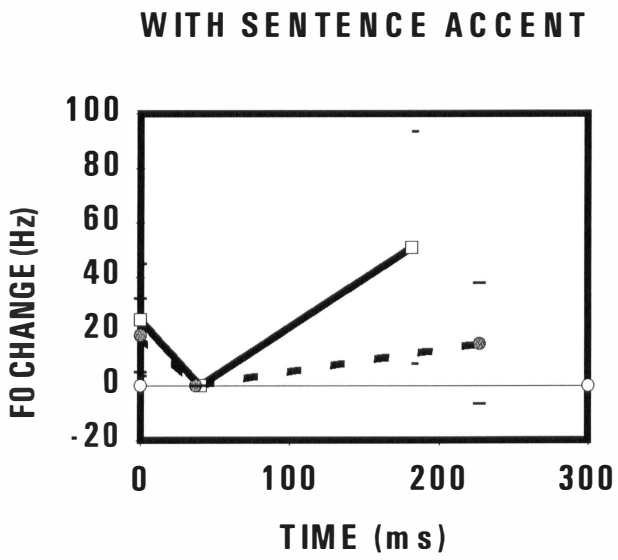
The analysis included 48 words; 14 CDS words and 13 ADS words without sentence accent, and 10 CDS words and 11 ADS words with sentence accent. The speed of the F0 change, in Hz/ms, was computed for both the Fall and the Rise in each word, and the means and standard deviations were calculated.

4. Results and discussion

Figure 1 shows the results for the words uttered *with* sentence accent in terms of the mean F0 changes and the mean rates of F0 change. In words marked with sentence accent there were almost no difference between CDS and ADS in the F0 Fall, and the rates of F0 change was also similar. In the subsequent Rise, though, an exaggerated F0 change and a steeper slope can be seen in the CDS data. In ADS the Rise took somewhat longer. The dispersion was moderate in the Fall, but considerable in the Rise, especially in CDS as can be seen from the standard deviations.

In words *without* sentence accent the Fall had a less steep slope in CDS than in ADS as can be seen in Figure 2. The F0 Fall was the same in both cases, but the duration was somewhat longer in CDS indicating a slower speed. The Rise took much longer in CDS so that the duration was as long as for words *with* sentence accent. Both the F0 Rise and the dispersion were greater in CDS than in ADS. The small mean F0 changes in both ADS and in words without sentence accent was mainly due to the fact that many of the F0 contours were inverted, i e the expected F0 fall in the primary stressed syllable was replaced by an F0 rise, and the expected F0 rise in the secondary stressed syllable was realized as an F0 fall.

A comparison between the overall F0 contours in CDS and ADS for words *with* and *without* sentence accent reveals a striking difference between the speech styles. In CDS the F0 contours were very similar in their timing characteristics. The difference is in the range of the F0 changes during the Fall and the Rise. In ADS on the other hand, the timing was clearly different. In words *with* sentence accent



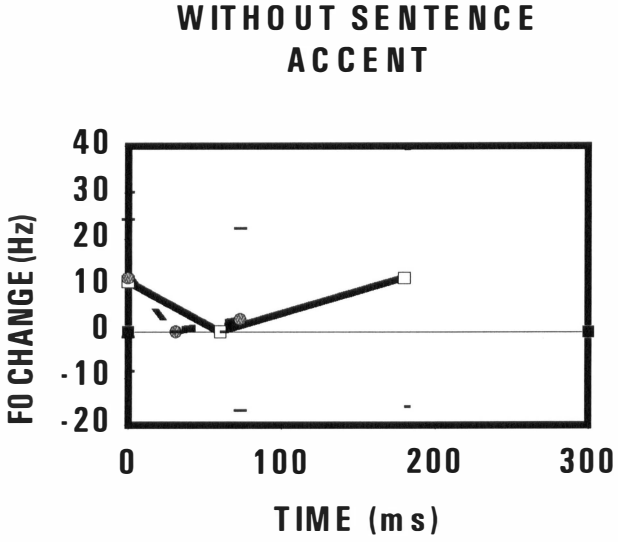


Figure 1. Mean F0 change as a function of time in words with sentence accent. CDS is marked with solid lines and ADS is marked with dotted lines. The bars represent +/– one standard deviation. The first section of the curves pertain to the F0 fall in the first syllable, and the second one relate to the rise in the second syllable.

Figure 2. Mean F0 change as a funcion of time in words without sentence accent. CDS is marked with solid lines and ADS with dotted lines. The bars represent +/– one standard deviation.

the duration of the Rise parameter is approximately 4 times longer than in words without sentence accent.

The temporal pattern in disyllabic accent 2 words seemed very different in CDS and in ADS. In CDS the timing characteristics were almost identical in words *with* and *without* sentence accent while a great difference was found in the F0 changes. In ADS, on the other hand, the F0 changes were almost identical while a great difference was found in the timing characteristics. These results are in agreement with observations reported by Fernald & Kuhl (1987) who found that 4 months old infants showed a significant preference for F0 patterns but not for duration patterns in CDS.

More data are currently being gathered to check these preliminary conclusions. The perceptual significance of the exaggeration of the F0 contours and time characteristics in CDS should be tested in formal experiments with infants. In such a study the results from the present investigation would provide valuable quantative data from a realistic mother — child interaction, in creating the stimuli.

Reference

Fernald, A., Kuhl, P. (1987): "Acoustic determinants of infant preference for motherese speech," *Infant Behavior and Development* **10**, 279–293.

Sundberg, U (1993): "Word accent 2 in child directed speech: A pilot study", *PERILUS,* **XVII** (this number).

Phonetic Experimental Research,
Institute of Linguistics,
University of Stockholm (PERILUS),
No. XVII, 1993, pp. 81–106

Stigmatized pronunciations in non-native Swedish

Una Cunningham-Andersson

Abstract

This paper reports the results of two experiments exploring the relationship between native Swedish speakers' perception of, on the one hand, different immigrant groups and, on the other hand, different phonetic features of immigrant Swedish. The first experiment showed that a single speaker can be judged differently as regards how friendly and educated he sounds when he uses some non-native pronunciations than others. This was interpreted as evidence that attitudes to non-native pronunciations are separate from attitudes to speaker groups. Evidence was also found that a single NNP — non-native pronunciation — in the same phonetic material produced by different speakers will be judged differently. This shows that the particular non-native pronunciation being used is not the only factor which influenced the listeners in their judgement of the stimuli. Other candidates are speaker characteristics, such as the perceived personality, accentedness, voice quality and sex of the speaker, as well as the assumed ethnic origin of the speaker. One of the naive listener groups (with lower socio-economic and educational status) was found to be significantly influenced in their judgements of the speakers by their beliefs about the speakers' backgrounds. The second experiment resulted in a list of the 94 NNPs occurring in an extended material showing the median judgements they elicited from 91 new listeners on the three judgement dimensions of NNP importance, speaker friendliness and speaker education. This list should be of interest to teachers of Swedish as a second language.

1. Introduction

Much existing work on attitudes to language variants is based on the assumption that these attitudes are intimately associated with attitudes to the groups who use the variants in question. For example, the classic matched-guise work of Lambert, Hodgson, Gardener and Fillenbaum (1960), studying attitudes of French and English speaking Canadians to the same bilingual speakers' French and English speech assumes that listeners who judged a speaker's French and English guises

differently were reacting to the very "Frenchness" or "Englishness" of the speaker, rather than any linguistic differences. Similarly, in previous experimental work on foreign accent by Cunningham-Andersson and Engstrand (1988), it was found that informants perceived a single non-native speaker of Swedish as significantly more friendly and significantly less successful (5% level) when he was believed to be a Kurd than when he was believed to be a German.

It is, however, important to distinguish between, on the one hand, attitudes to ethnic groups, which can evidently be elicited using non-native speech samples as stimuli, and on the other hand, attitudes to the non-native accents associated with these groups. Giles (1970) showed that this is possible. He found that British 12 and 17 year-olds perceived various British and foreign accents differently for status content (reflecting attitudes to the group speaking with the accent in question) and for aesthetic content (reflecting attitudes to the accent itself).

The effect of accent strength on this kind of speaker evaluation is well-documented. Stronger accents elicit less favourable reactions. Ryan, Carranza and Moffie (1977) and Sebastian, Ryan and Corso (1978), for example, found that Spanish-accented speakers of English with stronger accents were perceived more negatively on a number of non-linguistic dimensions (Ryan et al: likelihood of being a friend, eventual occupation, pleasant-unpleasant; Sebastian et al: social class, similarity in beliefs, desirability in a range of social relationships) than those with weaker accents. Similarly, Cunningham-Andersson and Engstrand (1988) found that two non-native speakers of Swedish, with different degrees of accentedness, both believed to be Kurds, were perceived by native Swedish speakers as being significantly different from each other on a number of dimensions representing perceived degree of education, friendliness, wealth, kindness, trustworthiness, intelligence, accentedness, comprehensibility, clarity of speech, pleasantness of voice and irritation, the more accented speaker being in all cases more negatively regarded.

The first of the two experiments reported here is designed to explore the relationship between the way different categories of native speakers of Swedish see different immigrant groups and the way they react to different phonetic features of non-native Swedish.

2 Experiment 1

2.1 Hypotheses

To establish that there are, in fact, phonetically conditioned attitudes to foreign accent, as distinct from attitudes to immigrant groups or immigrants in general, an attempt can be made to show that different non-native pronunciation features

produced by the same non-native speaker elicit different native speaker responses. Rather than eliciting a judgement of accent and personality based on an entire reading or speech sample, which would be difficult to distinguish from an attitude to the speaker's known, assumed or believed ethnic background, listeners can be asked to base their judgements on one non-native pronunciation (NNP) at a time. If the listeners are asked to judge not only how important a particular NNP is, but also to make judgements concerning the personality and status of the speaker, and if they are found to judge a single speaker's various NNPs differently on these dimensions, this means that genuine phonetically conditioned attitudes to foreign accent have been isolated.

The first hypothesis is, therefore, that different phonetic features of a certain non-native speaker's pronunciation cause native listeners to react in different ways (*hypothesis 1*), i.e. that it is possible to react positively to one feature of a particular foreign accent and negatively to another feature of the same accent. This kind of discrimination is clearly independent of the speaker and the listener's attitude to the ethnic group he knows, believes or assumes the speaker to represent.

Although it has been hypothesized that there are differences between native speaker judgements of different NNPs in a single speaker's speech, this does not imply that a given NNP will be judged similarly when produced by different speakers. The difference between speakers is too great for that. Also a single NNP can be more or less noticeable. It seems likely that a single NNP in similar phonetic contexts in different speakers' speech will be judged differently by native listeners, which would indicate that phonetically conditioned attitudes to the particular NNPs used by the speaker are tempered by differences in attitudes to the speakers' voice characteristics, accent strength, or known, believed or assumed ethnic origin of speakers. The second hypothesis is, therefore, that the same NNP in different non-native accents will elicit different native responses (*hypothesis 2*).

Phonetically conditioned attitudes to speech and speakers and inter-speaker variation have been dealt with in the previous two hypotheses, and an attempt has been made to distinguish them from attitudes to ethnic groups elicited using linguistic stimuli. Let us now consider this kind of attitude, which can be called ethnically conditioned attitudes to non-native speech and speakers, and which have been the subject of many investigations in the past, including the pioneering matched guise work of Lambert et al (1960).

A good deal about the way the Swedish population views various immigrant groups is known from investigations such as that reported in Westin (1984), which includes a table (page 205) where eighteen ethnic groups were ranked by Swedes for cultural similarity, similarity of values, familiarity and an objectively calculated measure corresponding to cultural distance. Other Scandinavians (including Finns)

were found to rank highly, closely followed by three Anglo-Saxon categories. Southern Europeans and Middle-Eastern peoples ranked lower, while Africans and Gypsies were given the lowest rankings. Given this, and given the results of previous attitude measurements, e.g. Lambert et al (1960), Cunningham-Andersson and Engstrand (1988), it is hypothesized that native speakers will perceive non-native speech and speakers differently depending on what they know, or believe they know about the speakers' linguistic, and, therefore, ethnic, origin (*hypothesis 3*).

Let us now turn our attention to the native listeners and judges in whose ethnically conditioned attitudes to non-native speech and speakers we are interested. Thus far, native Swedish listeners have been referred to as though they were a single group. They are, of course, not a homogenous group in this respect. Westin (1984, p. 262) reports that blue-collar workers are, on average, less tolerant of immigrants than are other groups. It seems likely that this difference might have an effect on the ethnically conditioned attitudes to speech and speakers mentioned in hypothesis 3.

It is hypothesized that the socio-economic and educational level of the listener group will influence linguistically elicited ethnically conditioned attitudes, or, in other words, native listeners with a lower socio-economic and educational level will be more influenced by what they believe about a speaker's linguistic origin than will similar listeners with a higher socio-economic and educational standard (*hypothesis 4*).

2.2 Materials and method

The four hypotheses listed above make certain requirements on the material used as stimuli: firstly, it must be possible to compare several NNPs from a single speaker; secondly, it must be possible to compare similar NNPs from several speakers in identical phonetic contexts; thirdly, NNPs must be presented to the listeners as nearly in isolation as possible, so that any accompanying material does not influence their judgements. The NNP cannot be electronically isolated, since this would sound too unnatural to naive listeners. Fourthly, listeners must not be distracted by any difficulty in understanding what the speaker is saying.

The first two requirements can be met using readings of a single text by a large number of non-native speakers of Swedish. The non-native speech material used in this experiment is from the IRIS data base (described in Engstrand and Cunningham-Andersson (1988)) which contains a large number of recorded speech samples from immigrants to Sweden speaking both their native languages and Swedish. Part of the Swedish material is readings of the Swedish version of "The North Wind and the Sun". From these readings, five NNPs which occurred in a number of different accents were chosen. These deviant pronunciations were: A) a velar nasal

is followed by a velar stop, e.g. *tvungen* (''obliged'') [tvuŋgɛn]; B) voicing assimilation is anticipatory, such that sequences such as /sv/ in *svepte* (''wrapped'') and /std/ in *just då* (''just then'') become voiced in their entirety. The native pronunciation would be unvoiced; C) a prothetic vowel is inserted before /s/ + stop in, for example *starkaste* (''strongest'') [ɛstarkastɛ] or an epenthetic one in consonant clusters, such as between /t/ and /v/ in *tvistade* (''were arguing'') [təvɪstadɛ]; D) the lexical stress in trisyllabic words is shifted from the first syllable (most usual in native Swedish words) to the second (penultimate) syllable; E) the quality distinction between long and short /a/: [ɑ:], [a], is not maintained, the long vowel being pronounced farther forward than in a native pronunciation e.g. *gav* (''gave'') [gav].

Five versions of each of these non-native pronunciations, each taken from one of two places in the text, produced by non-native speakers, were chosen with the smallest possible total number of speakers, such that the maximum number of comparisons between native reactions to a single speaker's deviant pronunciations could be made. This gave a total of 25 tokens from 11 speakers, with up to three tokens from each speaker.

Labov (1966, p. 482) claimed that his listener group's attention focused ''only on those items which have risen to the surface of social consciousness, and have entered the general folklore of language (...) most perception of language is not perception of sense experience, but of socially accepted statements about language''. Nonetheless, it was necessary to cause our listeners to focus their attention on the NNPs under investigation to satisfy the third requirement made by the hypotheses. The NNPs were extracted from the database along with a little accompanying material from the phrase or sentence in which they occurred in the text. To this end, the stimulus phrases were made as short as possible, including just the minimum necessary to facilitate comprehension (requirement 4 above). As a further aid to comprehension and focusing of the listeners attention, they were provided with a response form where the whole stimulus phrases were written, with the letters corresponding to the relevant NNP marked. The listeners were instructed to base their judgement solely on the marked portion of each stimulus phrase. This is clearly not an entirely satisfactory method for isolating the NNPs; there is a risk that the listeners are influenced by other parts of the utterance, but it is important that stimuli presented to naive listeners sound natural. Each phrase was repeated three times, and two sets of these 25 triplets were then arranged on a tape, each set in random order as follows, where S1, S2 etc. are stimuli embedded in their phrases: S1x3, S2x3 S25x3. One stimulus was omitted from the tape and therefore excluded from analysis in both stimuli sets.

There are also certain requirements on the listener groups used to test the hypotheses. Listeners on different socio-economic and educational levels are to be compared in hypotheses 3 and 4. To meet these requirements, two groups of 17–19 year-old native speakers were used. The first group was intended to represent listeners with a relatively high socio-economic and educational level. They were 72 students of technical subjects in the final year of three-year theoretical courses at upper secondary school in Uppsala (henceforth known as T3). The course of studies they have embarked on is aimed at university admittance to subjects such as Engineering. These T3 listeners are deemed to have a normal or average level of familiarity with immigrants.

The second group of listeners (here referred to as P2) was to represent listeners with a lower socio-economic and educational level. They were 33 native Swedish students of technical subjects in the final year of two-year practical (vocational) courses at the same upper secondary school as the T3 listeners. These listeners are also deemed to have had average exposure to immigrants.

The only difference between groups P2 and T3 is assumed to be their educational, and therefore social status, in relation to the employment they will be expected to have at the end of their studies. Faringer (1982) performed a set of dialect and sociolect identification experiments on groups of this kind, and found that the distinction between students studying practical and theoretical courses is a useful one which reflects socio-economic and sociolinguistic differences.

Cunningham-Andersson and Engstrand (1988, 1989) report extensive preliminary experiments on listener groups similar to the T3 and P2 categories of listeners, investigating the capabilities of these informants regarding identification of foreign accents and estimation of how strong accents are. Both groups had great difficulty in recognizing any accents but those associated with other Scandinavian languages and the school languages, English, French, and to some extent, German. It is very important to have knowledge of the accent identification capabilities of these two groups, since the intention is to present them with two or more conflicting sets of language background information for each speaker, without them being aware that they are hearing the same speakers several times. It must be clear not only which accents they recognize easily, but also which accents they can exclude from the set of possible accents when they hear a particular speaker. Preliminary attitude experiments have been carried out on groups of listeners similar to the T3 students.

Previous work on linguistic attitudes, for example, Bourhis, Giles and Tajfel (1973), Lambert, Giles and Picard (1975), Williams (1974), Carranza and Ryan (1975), Edwards (1977), Cunningham-Andersson and Engstrand (1988), has shown the need for two basic dimensions to describe judgements of speaker characteristics elicited from linguistic stimuli, corresponding to the status of the

speaker and solidarity felt by the listener with the speaker. These two dimensions are both related to how listeners perceive the group they believe the speaker to come from. An additional dimension is clearly necessary to elicit the listeners' direct and explicit attitude to the speech they are hearing as distinct from the speaker producing the speech. In the preliminary experimental work reported in Cunningham-Andersson and Engstrand (1988), 12 judgemental dimensions were used for registering listener perceptions of non-native speech. Many of these dimensions were correlated with each other, and it seems likely that three dimensions, corresponding to what e.g. Williams (1974) and Carranza and Ryan (1975) have called status and solidarity, with the addition of what e.g. Romaine (1980) calls speech or code characteristics are sufficient. Three dimensions representing (a) the perceived importance of eliminating each NNP as uttered by a particular speaker (which is assumed to be a measure of attitude to the phonetic material itself), (b) the perceived friendliness (corresponding to speaker-listener solidarity) and (c) the perceived educational level (corresponding to status) of each speaker when he or she uses particular NNPs have therefore been used here.

A good deal of response data from the listener groups were necessary to test the hypotheses. The two stimuli sets were combined in different ways with information about the speakers' backgrounds to allow all the hypotheses to be tested. Two answer forms were designed, one for each of the sets of stimuli. On each page there were columns showing the words in which the deviant pronunciations occurred, with the relevant letters marked where appropriate, e.g. "ju*st d*å". The informants were instructed to base their judgements as far as possible on the marked portion of the token phrase. There were also three columns where informants could indicate on a scale from 1–5 a) how important they felt it was that the speakers should try to eliminate each of the deviant pronunciations from their speech, b) how friendly the speakers sounded, c) how highly educated the speakers sounded. Both forms had mixed correct and incorrect information, arranged so that each speaker was described as having two different mother tongues (one assumed to have higher status than the other), one in each of the stimuli sets (i.e. forms). This was done to enable us to directly elicit attitudes to ethnic minority groups. Great care was taken at this point to ensure that none of the speaker background information presented to the informants would sound unlikely. Previous work (Cunningham-Andersson and Engstrand 1988) has shown that these types of listeners are likely to accept information they are given about speaker background.

All the stimuli were not presented to all the listeners. Twenty-four T3 listeners heard both stimulus sets, the remaining 48 heard only the second set. The 33 P2 listeners heard both sets.

The stimulus tape was presented to the informants in groups of 12–30, and they were asked to indicate their reactions on the answer forms. The listening tests were carried out in the informants' school, during their regular 40 minute Swedish lesson. The rooms used were intended for language lessons, and the students were used to listening to taped material in these rooms. The walls were clad with material designed to improve the acoustic characteristics of the rooms to some extent. The listeners heard the stimuli through a loudspeaker. The listeners were instructed to spread themselves around the room as much as possible, so that they would not be tempted to look at their classmates' responses. They were instructed not to speak during the test, and to raise a hand if they wanted to say anything. Spontaneous comments were made in some cases but the listeners were asked to save their comments until the test was over. The experimenter could see all the listeners. After each listening session, the listeners were asked to give their impressions of the task.

2.3 Results

The results obtained in relation to each hypothesis are presented in this section. Hypothesis 1 was that different phonetic features of a single non-native speaker's pronunciation cause naive native listeners to react in different ways, suggesting that the listener reacts to the features in question separately from the alleged ethnic background or any other characteristic of the speaker. Let us consider how each speaker's various NNPs were judged by the 72 native T3 listeners.

The judgements of each speaker's NNPs as heard in the first stimulus set (with mixed correct and incorrect information about the speaker) were subjected to Friedman's two-way ANOVAs by rank. A significant result in this test means that the individual speaker's NNPs were judged differently from each other. These results are shown in Table I.

There are quite a few significant results in table I. The NNPs spoken by each speaker are often judged differently from each other as regards how important it is to eliminate each NNP. This is, of course, not surprising. All it shows is that the listeners were able to perform the task at hand. What is much more interesting is that a single speaker can be judged differently as regards his or her perceived educational level and even, in a few cases, how friendly he or she is, depending on which NNP is used at any given time. This constitutes clear evidence for phonetically conditioned attitudes to foreign accent. Individual non-native pronunciations from a single speaker can be judged very differently. It is as though the listener were to say to the speaker not only: "I don't like that particular non-native pronunciation", but also: "When you use that particular non-native pronunciation you sound uneducated or unfriendly". This gives us corroboration for hypothesis one.

Our second hypothesis was that the same NNP in different non-native accents will elicit different native responses, which would indicate that phonetically conditioned attitudes are tempered by differences in attitudes to the speakers' voice characteristics, accent strength, and known, believed or assumed ethnic origin of speakers. To test this hypothesis, Friedman's two-way ANOVAs by rank were performed, testing whether T3 judgements of each NNP produced by the various speakers as heard in the first stimulus set (with mixed correct and incorrect speaker information) were drawn from the same population. A significant result in this test means that the individual speakers were judged differently. These results are shown in Table II.

Table I. Friedman two-way analysis of variance by ranks testing whether T3 judgements of each speaker's various NNPs as heard in the first stimulus set have been drawn from the same population. A significant result means that the NNPs were judged differently.

Speaker	NNP	NNP-Importance	Friendliness	Education
1	A,C,E	p<0.0001	p<0.01	p<0.0001
2	B,D,E	p<0.0001	p<0.001	p<0.0001
3	A,C,D	p<0.0001	NS	NS
4	C,D,E	NS	NS	p<0.01
5	A,B	p<0.01	NS	NS
6	C,D	NS	NS	p<0.01
7	A,B	NS	NS	p<0.01
8	D,E	p<0.0001	NS	NS
9	C,E	NS	NS	NS

Table II. Friedman two-way analysis of variance by ranks testing whether T3 judgements of each NNP produced by the various speakers as heard in the first stimulus set have been drawn from the same population. A significant result means that the speakers were judged differently.

NNP	Speakers	NNP-Importance	Friendliness	Education
A	1,3,5,7	p<0.0001	NS	p<0.0001
B	2,5,7,10,11	p<0.0001	NS	p<0.001
C	1,3,4,6,9	p<0.0001	p<0.001	p<0.0001
D	1,3,4,6,9	p<0.0001	NS	p<0.001
E	1,2,4,8,9	p<0.0001	p<0.001	p<0.0001

Table II shows that different non-native speakers' productions of single NNP categories were usually judged differently from each other. Only three tests failed to give a significant result at the 1% level, indicating that there was no significant difference between the speakers using NNP A (velar plosive following velar nasal), B (voicing assimilation) or D (misplacement of lexical stress) as regards their perceived friendliness while using these NNPs. In all other cases the speakers were not judged similarly although they were using the same NNP in the same phonetic context. This means that our hypothesis is supported. There is no significant tendency for similar non-native pronunciations to be judged in the same way when they are produced by different speakers. For example, NNP D (misplacement of lexical stress) was much less well tolerated when produced by speaker 8 than by speaker 4.

Other factors than the particular non-native pronunciation used are clearly affecting the judgements. At this point it is worth noting that the stimuli were in fact very short: usually just one or two words. Nonetheless, one such factor may be the overall accentedness of the speakers. The listeners were given information about the linguistic background of the speakers, so they are not likely to have made any kind of identification attempt on their own. We know from our previous work, Cunningham-Andersson and Engstrand (1988), that T3-type listeners are not very good at correctly identifying foreign accents.

The third hypothesis predicts that a speaker's pronunciation will be judged differently depending on what the native informants know, or believe they know about the speaker's linguistic origin. This was tested by studying the 24 T3 listeners' judgements of the first and second stimulus sets (with conflicting information about the speakers' linguistic backgrounds). A speaker might, for example, have three different NNPs in the first stimulus set and the same three occurring again in the second stimulus set. Of these six tokens from this single speaker, three will be identified as coming from a native speaker of, say, Spanish, while the other three may be identified as having been spoken by a native speaker of Kurdish. The Mann-Whitney U-tests compare the judgements of the speaker in this example elicited while the listeners believe him to be a speaker of Spanish with those elicited using the same phonetic material while the listeners believe him to be a speaker of Kurdish. The fact that the listeners accepted this conflicting information is a reflection of their limited capabilities of identification of foreign accents, as mentioned above, and also shows that it is hard to recognize an unfamiliar speaker after hearing several other speakers in between. The results of these tests are shown in table III.

Hypothesis 3 is not corroborated for the T3 listener group on any dimension. There are very few significant differences between the judgements of the speaker

guises, and those that there are, are only at the 5% significance level. This means that, while the listeners were not aware that they were hearing the same speaker more than once, the information they were given about the speakers' backgrounds did not affect their judgements.

This finding is not consistent with earlier work, e.g. Lambert et al (1960), Cunningham-Andersson and Engstrand (1988), Giles (1970). This could have several explanations. Perhaps this kind of young Swedish listener is so tolerant of immigrants that he does not distinguish between different ethnic groups. After all, the hypothesis concerns a difference between judgements of stimuli which were phonetically identical. Lambert et al's investigation was set in Canada and focused on French and English speakers. The stimulus material being compared in Lambert et al's study was not identical, being the same speaker speaking in both French and English. Perhaps it is more difficult for even the most enlightened listener to disregard such differences. Also Giles' work (1970) was based on a single speaker producing many stimuli, and so it is not directly comparable to the current investigation. As for the difference between these findings and those published by Cunningham-Andersson and Engstrand (1988) (where a single speaker was judged as being significantly more friendly and less successful when he was believed to be a Kurd than a German) the T3 listeners who took part in this current experiment were resident in Uppsala: a fairly large university town with a fairly large proportion of immigrants from many different countries. The listeners in the experiment

Table III. Mann-Whitney U-Test comparing judgements by the T3 listeners of each speaker's NNPs when the listeners were given two sets of information about the speaker's native language.

Speaker	Guises	NNP-Importance	Friendliness	Education
1	Spanish/Kurdish	NS	NS	S>K, p<0.05
2	French/Persian	NS	NS	NS
3	Japanese/Persian	P>J, p<0.05	NS	NS
4	Spanish/German	NS	NS	NS
5	Turkish/Russian	NS	NS	NS
6	Swahili/English	NS	S>E, p<0.05	NS
7	German/Vietnamese	NS	NS	NS
8	French/Vietnamese	NS	NS	NS
9	Turkish/English	NS	NS	NS
10	Tigrinya/Japanese	NS	NS	NS
11	Dutch/Arabic	NS	NS	NS

reported in Cunningham-Andersson and Engstrand (1988) were resident in Tierp, a much smaller town with a lesser proportion of immigrants from fewer language groups. Westin (1984, pp. 232–235) found that those native Swedes who had least contact with immigrants were least tolerant of immigrants. Perhaps these Tierp listeners were more likely to be influenced by information about the speakers they were hearing. Clearly, it cannot be assumed that all native speakers are equally influenced by information about speaker background. The fourth hypothesis concerns precisely these differences.

Hypothesis 4 was that native listeners with lower educational status will be more influenced by information about speaker background than will more educated listeners. This was tested by performing the same operations on the P2 listener data as were performed on the T3 data to test hypothesis 3. That is to say that 33 P2 listeners' judgements of the first and second stimulus sets (with conflicting information about the speakers' linguistic backgrounds) were compared using Mann-Whitney U tests. Table IV shows the results of the Mann-Whitney U-tests comparing judgements by the P2 listeners of each speaker's NNPs when the listeners were given two sets of information about the speaker's native language.

The picture here is completely different. As can be seen in table IV, there are many significant differences between the P2 listener judgements of identical phonetic material depending on the information the listener is given about the linguistic background of the speaker. In about half the cases of double speaker

Table IV. Mann-Whitney U-Test comparing judgements by the P2 listeners of each speaker's NNPs when the listeners were given two sets of information about the speaker's native language.

Speaker	Guises	NNP Importance	Friendliness	Education
1	Spanish/Kurdish	NS	NS	NS
2	French/Persian	F>P, $p<0.001$	F>P, $p<0.05$	P>F, $p<0.05$
3	Japanese/Persian	P>J, $p<0.001$	J>P, $p<0.01$	J>P, $p<0.001$
4	Spanish/German	NS	NS	NS
5	Turkish/Russian	NS	NS	T>R, $p<0.05$
6	Swahili/English	S>E, $p<0.01$	S>E, $p<0.001$	NS
7	German/Vietnamese	G>V, $p<0.01$	NS	NS
8	French/Vietnamese	V>F, $p<0.01$	V>F, $p<0.01$	NS
9	Turkish/English	NS	NS	NS
10	Tigrinya/Japanese	NS	NS	J>T, $p<0.0001$
11	Dutch/Arabic	A>D, $p<0.05$	NS	NS

identification the importance of the speaker's errors is judged differently depending on what the listener believes about the speaker. This means, taking speaker 3 as an example, that the same non-native pronunciations produced by the same speaker are deemed more important if believed to have been spoken by a Persian speaker (relatively low status) than by a Japanese speaker (relatively high status). By the same token, the "Japanese" speaker is judged as being significantly more friendly and more highly educated than the identical "Persian" speaker. Similarly, on the 1% level of significance, speaker 6 was judged as having more important NNPs and as being more friendly as a Swahili speaker than as an English speaker; speaker 10 was thought more highly educated as a Japanese speaker than as a Tigrinya speaker; speaker 2 had more important NNPs as a French speaker than as a Persian speaker; speaker 7's NNPs were more important when he was thought to be a German than a Vietnamese speaker; speaker 8's NNPs were thought more important when she was believed to be a speaker of Vietnamese than of French. Notice, however, that what we have assumed to be the higher status speaker guise (on the basis of Westin's study) is not always judged more positively.

Hypothesis 4 is obviously corroborated. The listeners with lower socio-economic and educational status as represented by the P2 group (17–19 year-old students on vocational courses) were considerably more influenced by information about the speakers' backgrounds than were the higher status group, T3 (17–19 year-old students on theoretical courses).

2.4 Discussion

It has been found that an individual speaker can be judged differently as regards how friendly (in some cases) and how highly educated he or she is perceived to be, when using different NNPs (non-native pronunciations) (hypothesis 1, table I). This shows that there are genuine phonetically conditioned attitudes to foreign accent. Different speakers were found to be judged differently although they were using the same NNP at the same place in the same text (hypothesis 2, table II). This shows that speaker characteristics such as the perceived personality, accentedness, voice quality and sex of the speaker, as well as the believed ethnic origin of the speaker are more important than the particular NNP being used, and that this kind of judgement of speaker characteristics can be based on very short stimuli. No evidence at all was found to suggest that the informants in group T3 were swayed in their judgements of a speaker or an NNP by information about what mother tongue the speakers had. Listeners in group P2, however, were influenced by this kind of information (hypotheses 3 and 4, tables 3 and 4). This constitutes a difference between listener groups with different social status.

Let us at this point step back and consider the implications of these findings. The first finding was that the impression T3 listeners got of a single speaker changed as the speaker used different non-native pronunciations. The speaker characteristics were unchanged; the only difference between the tokens which were judged differently was that they were taken from different parts of the same text. This means that the listeners were being influenced either by the text itself or by the speakers' pronunciation at the different points in the text where the tokens were extracted. Since table II showed significant differences between the speakers using a particular NNP, it is very unlikely to be some feature of the text itself which is causing the varied judgements for single speakers. It is, rather, the non-native pronunciation itself which is causing the variation.

This constitutes a clear case of distinct phonetically conditioned attitudes to different non-native pronunciation features. A single speaker's NNPs can be judged differently by naive native listeners as regards the importance of the NNPs used, and, more surprisingly, how friendly (in a few cases) and how highly educated the speaker is perceived to be. If a non-native speaker is perceived as having a lower level of education when he or she lets a final velar nasal be followed by a voiced velar stop, such that a word like *gång* (''time'') is pronounced [gɔŋg] (NNP A), than when he or she inserts a vowel in a consonant cluster in a word like *tvistade* (''were arguing''), such that the sequence /tv/ becomes [tɛv] or [təv] (NNP C) this has implications for the teaching of Swedish as a second language since it shows that listener reactions may possibly be influenced by the systematic elimination of stigmatized non-native features from the individual's speech.

It was found that when the same part of the text was read by different speakers using the same non-native pronunciation, the listeners judged the speakers significantly differently on all dimensions. It may be the case, then, that the NNP used is accorded more or less importance for judgements on a given dimension depending on other speaker characteristics, such as the overall accentedness, the believed ethnic group or the sex of the speaker, making it difficult to compare listener judgements of one speaker with those of another.

The findings here, particularly those presented in table I, show clearly that it is possible to distinguish between attitudes to minority groups and attitudes to non-native speech. This is not to say that attitudes to minority groups do not exist in Sweden. Westin (1984) gives ample evidence of this. In this experiment, we have found evidence of a difference between listener groups. Westin found that blue collar workers were less tolerant of immigrants than were groups with higher socio-economic status. It has been shown here that this kind of difference between social groups is also reflected in linguistic attitudes held by 17–19 year olds. The judgements elicited from the lower status P2 group were influenced by what the

informants believed to be the mother tongues or ethnic groups of the speakers. The higher status T3 group, however, was not influenced by this information. The relative youth of our listener groups is also interesting, suggesting that these social differences in attitude are likely to be present in the future too.

It would clearly be useful for immigrants to learn to avoid the most stigmatized NNPs. The five NNP categories studied here can only give an indication that differences exist. Obviously, more NNP categories are necessary if this area is to be investigated in enough detail to give results of any practical use to learners. The following experiment is an attempt to do just that. This experiment is designed to establish which kinds of NNPs elicit the least favourable reactions from native listeners. It is an elaboration of experiment 1, which showed that a single speaker's various NNPs can be judged differently as regards NNP importance (which is unsurprising) and perceived friendliness and educational level of the speaker (which is much more serious). The purpose of this experiment is to try to identify the most stigmatized NNPs.

3 Experiment 2

3.1 Material and method

The material used in this experiment is also taken from the IRIS data base (Engstrand and Cunningham-Andersson 1988). NNPs occurring in both the readings of texts (once again, "The North Wind and the Sun") and in spontaneous speech (the speaker was encouraged to tell the "story of his life") were extracted from the data base, along with as little accompanying material as was deemed appropriate (in the same way as was discussed in experiment 1, above). The non-native pronunciations (NNPs) were divided into NNP categories, according to table V.

All in all, 94 stimulus tokens were selected. 21 speakers of 13 languages were involved. These tokens were then arranged on a tape in random order. A set of five tokens were also placed at the beginning of the tape, to be used for training purposes.

Only one listener group was used for this experiment, although they were tested in smaller groups of 20–30 listeners. This group was similar to the T3 group in the last experiment, composed of 91 upper secondary school students of technical subjects in the final year of a three-year theoretical course in Uppsala, although none of the students took part in both experiments. The same three judgement dimensions as were used in Experiment 1: NNP importance, friendliness and educational level were used again here.

As in the previous experiment, the informants had an answer sheet with the stimulus material in written form, to aid comprehension. Once again, the letters

corresponding to the NNP under investigation in each case were marked, in an attempt to direct the listeners' attention to just that feature of the token. The informants were instructed to base their judgements as far as possible on the marked portion of the token phrase. Each stimulus was repeated three times. Here too, there were three columns where informants were instructed to indicate on a scale from 1–5 a) how important they felt it was that the speakers should try to eliminate each of the deviant pronunciations from their speech, b) how friendly the speakers sounded, c) how highly educated the speakers sounded.

The first five stimuli on the tape and the form were dummies (they all occurred again at a later point in the tape). The informants were told that they could use these five to train on, and that their answers would not be analyzed. After this training

Table V. NNP categories.

Vowels
V1 [ɑ:] — [a]
V2 front — back
V3 reduction
V4 vowel + nasal — nasalized vowel
V5 orthographically influenced pronunciations (e.g. *dom kom* "they came" [dom kom] for [dɔm kɔm])

Phonotax
PH1a prothetic vowel e.g. *starkast* ("strongest") [estarkast]
PH1b epenthetic vowel e.g. *tvistade* ("were arguing") [təvɪstade]
PH2 cluster reduction
PH3 cluster reordering
PH4 consonant insertion

Prosody
P1a long vowel shortening e.g. *tätare* [tɛtare] for [tɛ:tare]
P1b short vowel lengthening e.g. *till sist* ("at last") [ti:l si:st] for [tɪl sɪst]
P2 stress placement (prepenultimate to penultimate) e.g. *vandrare* ("traveller")
P3 grave accent — acute accent

Consonants
C1 problems with "sje" fricative [ɧ]
C2 problems with /r/
C3 consonant devoicing
C4 voiceless — voiced
C5 stop — fricative
C6 retroflex — [r] + fricative
C7 retroflex — [r] + dental
C8 retroflex — dental
C9 retroflex — [r] + retroflex
C10 orthographically influenced pronunciations (e.g. *skina* "shine" [ski:na] for [ɧi:na] etc.)
C11 problems with [l]
C12 [v] — [w]
C13 [ŋ] — [ŋg]

session, the tape was stopped, and the informants were invited to ask questions. The 94 stimuli were then presented one by one, with a pause after each one to allow the students to fill in their responses on the form.

3.2 Results and discussion

The appendix lists all the 94 NNPs according to the median judgement of NNP importance from the 91 listeners on the five point judgement scale, such that a median score of 1 represents less important NNPs, while a median score of 5 represents more important NNPs. Note that the list shows those NNPs which were judged as being more important to eliminate with a higher score.

From the appendix, we can see, as we did in experiment 1, that similar NNPs from different speakers are judged differently. Let us consider the PH1a type NNPs such as RO's prothetic vowel before the /st/ cluster in *starkast* ("strongest"), transcribed as [estarkast] which elicited a median judgement of 1 (i.e. not a very important NNP) and speaker SA's similar NNPs in *starkare* ("stronger"), [estarkare], and *Sverige* ("Sweden"), [esverje] which elicited median judgements of 4 (a fairly important NNP). It is, however, possible to see that some kinds of NNPs are considered more important than others. Vowel insertion into a cluster, NNP type PH1b, seems to be considered a more serious NNP on the whole: JH and HS elicited median judgements of 5 for their [jenasit] *genast* ("at once") and [setrɔlar] *strålar* ("rays") respectively, and FG scored median 4 for his [tɔvistade] *tvistade* ("were arguing"), while NA's production of the same NNP elicited a median 3.

It is interesting to note that of all the various NNPs of the Swedish retroflex consonants found in our material, C6, C7, C8 and C9, the "worst" is to change the retroflex to [r] + a dental consonant (which is odd, given that this is the solution adopted by Southern Swedish dialects with uvular /r/), while the best tolerated is to use an [r]+ retroflex consonant (this latter solution was used by two Bengali speakers).

NNP type C13, where a velar plosive is permitted to follow a velar nasal was deemed to be quite an important NNP, scoring medians of 4 and 5, whether word final, such as in PLs [gɔŋg], *gång*, ("time"), or in medial position, such as in ED's [tvuŋgen], *tvungen*, ("obliged").

All the NNPs in category PH3, cluster reordering were judged as very important, with medians of 5, e.g. [strakare], *starkare*, ("stronger"); [strakats], *starkast*, ("strongest"), and [varandra], *vandrare* ("traveller"), which is really syllable reordering, and also forms another Swedish word: *varandra*, ("each other").

NNP categories C4, where voiceless segments become voiced, e.g. [dvuŋen], *tvungen*, ("obliged"); [juzdɔ], *just då*, ("just then"), and P2, which involves misplacement of lexical stress on the penultimate syllable, e.g. *slutade*, ("fin-

ished") are examples of mid-range NNPs, with median importance scores of 3 and 4.

Of the NNPs of type P1a and P1b, concerning vowel length, there is a clear difference between P1b, involving a too long pronunciation of short vowels, e.g. ED, OT, RO and PU's [ti:l si:st], *till sist* ("at last"), which were not considered very important (medians 2 and 3) and P1a involving shortening of a long vowel: HI, JE, FG and SA's [tɛtare] or [tatare], tätare, ("more tightly") (medians 4 and 5). This might also have something to do with this NNPs similarity with the Swedish word *tattare* ("tinker"). As an illustration of the frequency distribution of the listener judgements, figure 1 shows the number of judgements in each category for the P1a and P1b NNPs.

The appendix shows the 94 NNPs according to the median judgement of how friendly the speaker was judged to be while uttering just that NNP. The results here are less varied. Most NNPs elicited a median judgement of 3, indicating neutral friendliness/unfriendliness, although there was considerable variation between the 91 listeners' judgements. Notice, however, that speaker OT was assigned a median

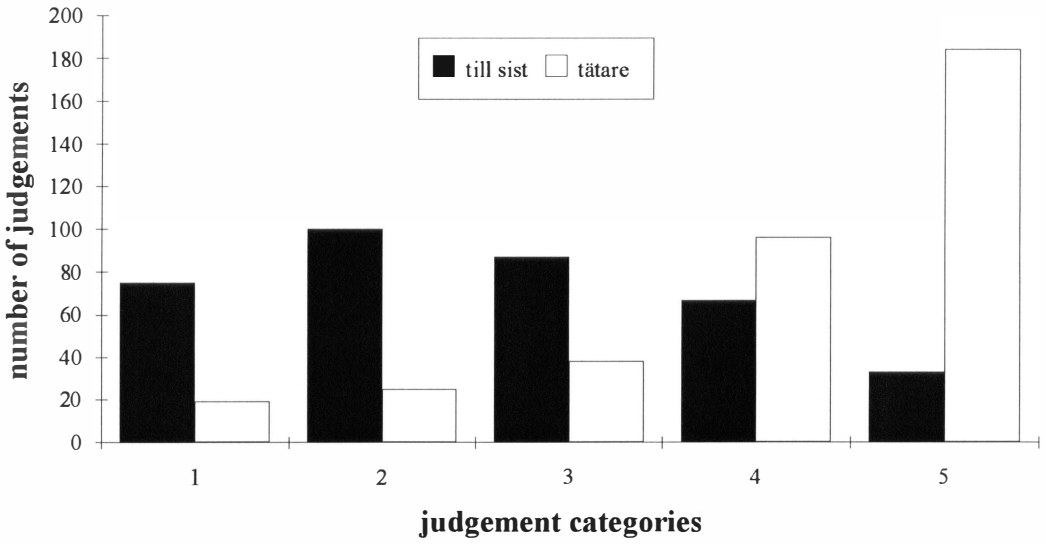


Figure 1. Histogram showing the frequency with which each judgement was assigned to the two NNPs p1a vowel shortening in tätare ("more tightly") [tɛtare] and p1b vowel lengthening in till sist ("at last") [ti:l si:st]. The dark bars represent [ti:l si:st] while the light bars represent [tɛtare].

friendliness score of 2 (fairly "low" friendliness) in four of her five NNPs, while speaker MI was assigned median 4 (fairly "high" friendliness) in 2 of 5 NNPs. This seems to indicate that there is not much connection between the NNPs used and the perceived friendliness of the speaker, but rather differences between speakers. This result is consistent with the findings of Experiment 1 (tables I and II).

Judgements of how educated speakers sounded were more varied than friendliness judgements. It is interesting to note that there are no median judgements of 5 (high educational level). Here the variation seems to be connected to both speakers and to NNP categories. Speakers JH and SH with their phonotactic NNPs, e.g. [jɛnasit], *genast*, ("at once"), and [strakats], *starkast* ("strongest") were judged as having very low educational levels (median scores of 1 and 2). Conversely, the highest estimated educational levels were assigned to the easily identified American speaker CE, and to ED and HI, all of whom also got some median 3 judgements. The difference between the median 2 and median 3 judgements appears to be associated with the judgements of the importance of each NNP, for example, P1a type NNPs involving vowel shortening, e.g. [tɛtarɛ] or [tatarɛ], *tätare*, ("more tightly") were here too judged more severely than P1b involving vowel lengthening, e.g. [ti:l si:st], *till sist* ("at last") etc.

The results of this experiment have obvious pedagogical implications. The NNPs associated with the least favourable overall impressions would, naturally, be worth avoiding for non-native speakers. It is, however, difficult to capture any generalities concerning overall impressions on the basis of the appendix. Figure 2 shows the median judgements of the importance of eliminating all the NNPs pertaining to each NNP category. From this it can be seen that three NNP categories, c10, (orthographically influenced pronunciations, such as [ski:na] instead of native [ɧi:na] for *skina* ("shine")), p1a, vowel shortening in *tätare* ("more tightly"), [tɛtarɛ/tatarɛ] instead of native [tɛ:tarɛ] and ph3, (cluster reordering e.g., [strakats] instead of [starkast] for *starkast* ("strongest")) are particularly undesirable, with median scores of 5. Eleven other NNP categories are worth avoiding for the learner, having a median score of 4. They are: c1, problems with /ɧ/, e.g., *sjuttio* ("seventy") [xɔti] instead of native [ɧɵti]; c7, retroflex consonants being pronounced as [r] + a dental, e.g., nordan ("north") [nɔrdan] instead of native [nu:ɖan]; c8, retroflex consonants being replaced by the corresponding dental consonants; c13, [ŋ] — [ŋg] or [ŋk]; c11, problems with [l]; p2, misplacement of lexical stress, ph1b, epenthetic vowel e.g. *tvistade* ("were arguing") [tɘvɪstadɛ]; ph4, consonant insertion in clusters; v2, front vowels being pronounced further back, e.g. *började* "began" [bɔrjadɛ]; v3, vowel reduction; v5, orthographically influenced vowel pronunciation, e.g. *fram* "to the end" [frɔm].

The very limited amount of existing research carried out with the aim of establishing priorities among native Swedish pronunciation goals, e.g. Bannert (1990) has been steered by consideration of speaker comprehensibility, not, as is the case here, of native listeners' reactions to the non-native speech they hear. Obviously, the primary aim of the non-native speaker must be to make himself understood. The work reported here deals with the usually infinite period from mastery of this aim to native-like pronunciation. This is of the utmost importance to all who are obliged to carry on their lives through the medium of a language other than their native language. The kinds of phonetically conditioned attitudes studied in this paper must not be ignored in the setting of targets for second language learners.

4 Summary

This paper deals with two experiments exploring the relationship between native Swedish speakers' perception of, on the one hand, different immigrant groups and, on the other hand, different phonetic features of immigrant Swedish.

In the first experiment, five categories of non-native pronunciations (NNPs) were presented in short phrases to two distinct listener groups. The listeners were required to indicate their impressions of each phrase on a form (where the NNP in question was marked) using three response dimensions. It was found that there were

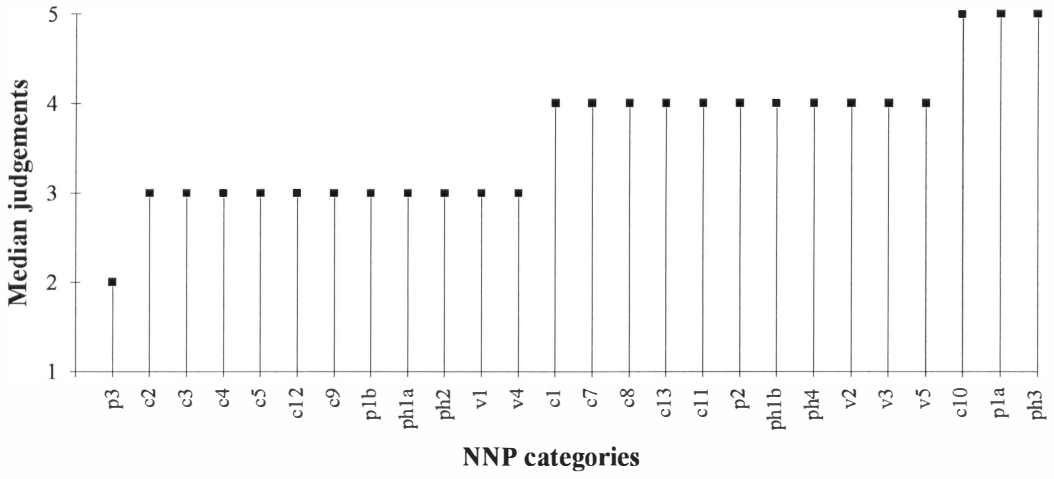


Figure 2. Median judgements of importance of each of the NNP categories.

differences between the listener groups we used, and that a single speaker can be judged as being less friendly and less highly educated for some NNPs than for others. It was also found that the same NNP used by different speakers in the same phonetic context will not be judged similarly, which suggests that even very short stimuli permit phonetically conditioned judgements about speaker personality to be made.

Experiment 1 showed that a single speaker can be judged differently as regards how friendly and educated he sounds when he uses some non-native pronunciations than others. This was interpreted as evidence that attitudes to non-native pronunciations are separate from attitudes to speaker groups. Evidence was also found that a single NNP in the same phonetic material produced by different speakers will be judged differently. This shows that the particular non-native pronunciation being used is not the only factor which influenced the listeners in their judgement of the stimuli. Other candidates are speaker characteristics, such as the perceived personality, accentedness, voice quality and sex of the speaker, as well as the assumed ethnic origin of the speaker.

When we compared judgements of single stimuli presented more than once with different information about the speaker's mother tongue, this information was found to have no significant effect on the listeners in one of our naive listener groups, T3, who represent higher socio-economic and educational status, while the other, lower status listener group, P2, was found to be significantly influenced in their judgements of the speakers by their beliefs about the speakers' backgrounds.

Five NNPs were sufficient to test the hypotheses of the first experiment, concerning how different kinds of listeners judge different NNPs uttered by different speakers, but this tells us nothing about other commonly occurring NNPs. Which kinds of NNPs are associated with the most negative listener responses? To answer this question, a large number of NNPs were tested for their perceptual effect on native listeners in the second experiment. Again, the stimulus NNPs were presented to the informants in short phrases. The second experiment resulted in a list of the 94 NNPs occurring in an extended material showing the median judgements they elicited from 91 new listeners on the three judgement dimensions of NNP importance, speaker friendliness and speaker education. This list should be of interest to teachers of Swedish as a second language.

Acknowledgements

I want to thank all the secondary school students and their teachers for participating in these experiments. I am grateful to Olle Engstrand and Renée van Bezooijen for their comments on an earlier version of this paper. This research was carried out in the project Attitudes to Immigrant Swedish, at the Department of Linguistics,

Stockholm University, led by Olle Engstrand and financed by the Swedish Council for Research in the Humanities and Social Sciences.

References

Bannert, R. (1990): *På väg mot svenskt uttal.* Lund: Studentlitteratur.

Bourhis, R., Giles, H., Tajfel, H. (1973): "Language as a determinant of Welsh identity". *European Journal of Social Psychology*, **3,** 447–60 .

Carranza, M., Ryan, E. (1975): "Evaluative reactions of bilingual Anglo and Mexican American adolescents toward speakers of English and Spanish". *International Journal of the Sociology of Language*, **6,** 83–104 .

Cunningham-Andersson, U., Engstrand, O. (1988): "Attitudes to immigrant Swedish — A literature review and preparatory experiments". *Phonetic Experimental Research at the Institute of Linguistics, University of Stockholm (PERILUS),* **8,** 103–152.

Cunningham-Andersson, U., Engstrand, O. (1989): "Perceived strength and identity of foreign accent in Swedish". *Phonetica*, **46**, 138–154 .

Edwards, J. (1977): "Students' reactions to Irish regional accents". *Language and Speech,* **20 (3)**, 280–286 .

Engstrand, O., Cunningham-Andersson, U. (1988): *IRIS — A data base for cross-linguistic phonetic research. Slutrapport från projektet "Invandrarröster i Sverige — fonetiska modeller (IRIS)* (Final report from the project "Immigrant voices in Sweden — phonetic models (IRIS)) . Manuscript, Dept. of Linguistics, University of Uppsala .

Faringer, G. (1982): "Språk och social identifikation — en undersökning bland gymnasieelever i Stockholmsområdet". *Inst. för Nordiska Språk vid Uppsala Universitet, FUMS rapport,* **106**.

Giles, H. (1970): "Evaluative reactions to accents". *Educational Review*, **22,** 211–27

Labov, W. (1966): *The social stratification of English in New York City* . Washington DC.: Center for Applied Linguistics.

Lambert, W., Giles, H., Picard, O. (1975): "Language attitudes in a French-American community". *International Journal of the Sociology of Language,* **4,** 127–152 .

Lambert, W., Hodgson, R., Gardener, R., Fillenbaum, S. (1960): "Evaluational reactions to spoken languages". *Journal of Abnormal and Social Psychology*, **60**, 44–51 .

Romaine, S. (1980): "Stylistic variation and evaluative reactions to speech problems in the investigation of linguistic attitudes in Scotland". *Language and Speech,* **23 (3)**, 213–232 .

Ryan, E., Carranza, M., Moffie, R. (1977): "Reactions toward varying degree of accentedness in the speech of Spanish-English bilinguals". *Language and Speech*, **20**, 267–273.

Sebastian, R., Ryan, E., Corso, L. (1978): "Social judgements of speakers with differing degrees of accentedness". *Paper presented at the meeting of the Ninth World Congress of Sociology,* Uppsala.

Westin, C. (1984): *Majoritet om minoritet, en studie i etnisk tolerans i 80-talets Sverige. En rapport från Diskrimineringsutredningen.* Stockholm: LiberFörlag.

Williams, F. (1974): "The identification of linguistic attitudes". *International Journal of the Sociology of Language,* **3,** 21–32 .

Appendix

The following list shows the 94 NNPs grouped according to the median judgement of how important each NNP is. The columns represent, from left to right, the speaker, the NNP category involved, the orthographical representation of the word(s) in which the NNP occurs, a phonetic transcription of the native pronunciation (NP), the NNP and the median judgements of how important each NNP sounds and how friendly and how well educated each speaker sounds while uttering each NNP are also shown.

Speaker	NNP type	Word	NP	NNP	NNP importance	Speaker friendliness	Speaker education
ro	ph1a	starkast	[stàrkast]	[estarkast]	1	3	3
ce	ph2	vandraren	[vàndrarn]	[vanrarən]	1	3	4
ce	ph2	andra	[àndra]	[anrə]	1	3	4
ro	c5	dom	[dɔm]	[ðɔm]	2	3	3
ce	v3	kappan	[kàpan]	['kapən]	2	3	3
na	v4	insvept	[ìnsve:pt]	[ĩsfept]	2	3	2
na	v4	anses	[ànse:s]	[ãses]	2	3	2
ed	p1b	till sist	[tɪl sɪst]	[ti:l si:st]	2	3	3
ot	p1b	till sist	[tɪl sɪst]	[ti:l si:st]	2	3	3
hi	p3	starkast	[stàrkast]	['starkast]	2	3	4
mi	p3	kappa	[kàpa]	['kapa]	2	3	3
ed	ph4	ta av	[tɑ: ɑ:v]	[tɑʔav]	2	3	4
je	c1	skina	[ɧì:na]	[xi:na]	3	3	3
mi	c2	varm	[varm]	[vaɽm]	3	4	3
bw	c2	överens	[ø:vər'ɛns]	[øveʁens]	3	3	3
ed	c3	gav	[gɑ:v]	[gaf]	3	3	3
ts	c3	medveten	[mè:dve:ten]	[metveten]	3	3	3
ju	c3	gav	[gɑ:v]	[gɑf]	3	3	3
fm	c4	tvungen	[tvèŋen]	[dvuŋen]	3	3	2
ib	c4	just då	[jəst do:]	[juzdɔ]	3	3	3
ed	c4	just då	[jəst do:]	[juzdɔ]	3	3	3
je	c5	dom	[dɔm]	[ðɔm]	3	3	2
fm	c7	hårdare	[hɔ̀:ɖare]	[hɔrdare]	3	3	2
fm	c8	nordan	[nù:ɖan]	[nɔdan]	3	3	3
fg	c9	nordan	[nù:ɖan]	[nɔɽɖan]	3	4	3
sa	c9	hårt	[ho:ʈ]	[hɔɽʈ]	3	3	2
fg	c12	vinden	['vɪndn]	[wɪndən]	3	3	3
fg	c12	vandraren	[vàndrarən]	[wan...]	3	3	3
je	v1	ta av	[tɑ: ɑ:v]	[ta av]	3	3	3
ed	v1	gav	[gɑ:v]	[gav]	3	3	3
ce	v3	tvistade	[tvìstade]	[tvɪstəde]	3	3	3
na	v4	nånsin	[nɔ̀nsin]	[nõsĩ]	3	3	2

Speaker	NNP type	Word	NP	NNP	NNP importance	Speaker friendliness	Speaker education
ju	p1b	till sist	[tɪl sɪst]	[ti:l si:st]	3	3	3
ro	p1b	till sist	[tɪl sɪst]	[ti:l si:st]	3	3	2
pu	p2	vandraren	[vàndrarn]	[vanˈdraren]	3	3	3
mv	p2	medge	[mè:dje:]	[ˈmedje]	3	3	3
fg	ph1a	skulle	[skèl]	[eskule]	3	3	3
je	ph1a	starkaste	[stàrkast]	[estar...]	3	3	3
na	ph1b	tvistade	[tvìstade]	[təvɪstade]	3	3	2
ro	ph2	först	[fœʂt]	[føʃ]	3	3	3
na	ph2	först	[fœʂt]	[føʃ]	3	3	2
fm	c1	skina	[ɧì:na]	[xi:na]	4	3	3
fg	c1	skönt	[ɧø:nt]	[xunt]	4	3	2
ot	c3	gav	[gɑ:v]	[gaf]	4	2	2
fm	c4	svepte	[svè:pte]	[zvepte]	4	3	3
pu	c4	just då	[jøst do:]]	[juzdɔ]	4	3	2
ro	c6	hårdare	[hɔ̀:ɖare]	[hɔrθare]	4	3	2
ot	c7	hård	[hɔ:ɖ]	[hɔrd]	4	2	2
pl	c8	nordan	[nù:ɖan]	[nodan]	4	3	2
sh	c11	strålar	[strɔ̀lar]	[strɔtar]	4	3	2
sa	c12	vem	[vɛm]	[ɸɛm]	4	3	2
pl	c13	gång	[gɔŋ]	[gɔŋg]	4	3	2
mv	c13	tvungen	[tvèŋen]	[tvuŋgen]	4	3	2
na	c13	tvungen	[tvèŋen]	[tvuŋgen]	4	3	2
mi	c13	tvungen	[tvèŋen]	[tvuŋgen]	4	4	2
hs	c13	tvungen	[tvèŋen]	[tvuŋgen]	4	3	2
ed	c13	tvungen	[tvèŋen]	[tvuŋgen]	4	3	3
mv	c13/4	gång	[gɔŋ]	[gɔŋk]	4	3	2
ot	v1	gav	[gɑ:v]	[gaf]	4	2	2
fg	v2	överens	[ø:vərˈɛns]	[ɔʊverens]	4	3	3
hs	v2	började	[bœ̀rjade]	[bɔr...]	4	3	3
hs	v2	överens	[ø:vərˈɛns]	[ɔʊverens]	4	3	3
ib	v3	vandraren	[vàndrarən]	[vandrn̩]	4	2	2
ot	v4	vinden	[vìndən]	[vĭden]	4	2	2
mi	v5	fram	[fram]	[frɔm]	4	3	2
ib	v5	dom kom	[dɔm kɔm]	[dom kom]	4	3	2
hi	p1a	tätare	[tɛ̀:tare]	[tɛtare]	4	3	3
je	p1a	tätare	[tɛ̀:tare]	[tɛtare]	4	3	2
na	p2	vandraren	[vàndrarən]	[vanˈdraren]	4	3	2
pl	p2	tvistade	[tvìstade]	[tvisˈtade]	4	2	2
hs	p2	slutade	[slù:tade]	[sluˈtade]	4	3	2
hs	p2	slutade	[slù:tade]	[sluˈtade]	4	3	2
sa	ph1a	Sverige	[sverje]	[esverje]	4	3	2

Speaker	NNP type	Word	NP	NNP	NNP importance	Speaker friendliness	Speaker education
sa	ph1a	starkare	[stàrkare]	[estarkare]	4	3	2
fg	ph1b	tvistade	[tvìstade]	[təvɪstade]	4	3	3
ro	ph2	insvept	[ìnsve:pt]	[insvep]	4	3	2
pu	ph2	anses	[ànse:s]	[ases]	4	3	3
sh	ph4	försöket	[fœʂø:kət]	[...ktet]	4	2	2
hs	c1	sjuttio	[ɧʉ́ti]	[xɔti]	5	3	2
jh	c10	skina	[ɧì:na]	[skina]	5	3	2
sa	c13	tvungen	[tvèŋen]	[twaŋgen]	5	3	2
jh	c13	tvungen	[tvèŋen]	[twaŋgen]	5	3	1
sh	v2	kappa	[kàpa]	[kɔpa]	5	3	2
sh	v3	hårdare	[ho:ɖare]	[hɔrdre]	5	3	1
sa	p1a	tätare	[tɛ̀:tare]	[tɛtare]	5	3	2
fg	p1a	tätare	[tɛ̀:tare]	[tatare]	5	3	3
hs	ph1b	strålar	[strò:lar:]	[sɛtrɔlar]	5	3	2
jh	ph1b	genast	[jè:nast]	[jenasit]	5	3	1
sh	ph3	starkare	[stàrkare]	[strakare]	5	3	2
jh	ph3	vandrare	[vàndrarə]	[varandra]	5	3	1
sh	ph3	vandrare	[vàndrarə]	[varandra]	5	3	1
jh	ph3	starkast	[stàrkast]	[strakats]	5	3	1
jh	ph4	överens	[øvərˈɛns]	[ɔʊversens]	5	3	2
mi	ph4	insvept	[ìnsve:pt]	[insvempt]	5	3	2