Perception, Analysis and Synthesis of Speaker Age

Schötz, Susanne

2006

TRAVAUX DE L'INSTITUT DE LINGUISTIQUE DE LUND 47

Perception, Analysis and Synthesis of Speaker Age

Susanne Schötz

GSLT
The Graduate School
of Language Technology

LUND
UNIVERSITY

Department of Linguistics and Phonetics
Centre for Languages and Literature
Lund University
Box 201
SE-221 00 Lund

ISSN 0347-2558
ISBN 91-974116-4-7

Printed in Sweden
Media-Tryck
Lund 2006

,,Am Heimcomputer sitz' ich hier
Und programmier' die Zukunft mir!“
“I program my home computer
Beam myself into the future!”

Kraftwerk, 1981

Abstract

Speaker age is an important paralinguistic feature in speech which has to be considered in the study of phonetic variation. Knowledge about this feature may be used to improve speech technology applications, e.g. automatic speech recognition and speech synthesis. The present thesis describes six studies of several phonetic aspects of age-related variation in speech.

As the speech production mechanism changes from young adulthood to old age, speech is affected in numerous ways. Human perception of speaker age is based on cues such as pitch, speech rate and voice quality, and is fairly accurate. However, it is still unclear which cues are the most important ones. The first study included in this thesis investigated the role of F_0 and speech rate (word duration) in age perception. It was found that while these cues may be less important than spectral ones (e.g. formant frequencies), they still correlate with chronological as well as perceived age. In the second study, two stimulus types of various lengths were compared. Results indicated that while longer stimulus duration (regardless of speech type) seems to improve the age estimation of females, spontaneous speech (regardless of duration) appears to contain more important cues for perception of male speaker age.

In the next two studies, several automatic estimators of speaker age were built, none of which reached the same accuracy as humans. Important features in machine perception of age were also investigated. It was found that prosodic features seem to be more important in the estimation of female age, while spectral features (e.g. F_2) appear to be more important for male age.

Although several acoustic correlates of speaker age are known, their relative importance has not yet been established. The next study analysed 161 features, automatically extracted from segments in six words produced by 527 speakers. Normalised means were used to ensure that the features could be compared directly. The most important acoustic correlates of speaker age were identified to be speech rate (segment duration) and intensity range. However, F_0 and some spectral measures (e.g. F_1 and F_2) may also, if used in combination with other features, be important correlates of age.

Synthetic speech may sound more natural if speaker age is included as a parameter. The final study developed a research tool which used data-driven formant synthesis and age-weighted linear interpolation to simulate an age between the ages of any two of four female differently aged reference speakers. Evaluation of the tool showed that speaker age may in fact be simulated using formant synthesis. The tool will be used in further studies of analysis by synthesis of speaker age.

KEY WORDS: speaker age, perceptual cues, automatic speaker recognition, acoustic analysis, acoustic correlates, data-driven, formant synthesis.

Sammanfattning

Talarålder är en viktig paralingvistisk egenskap i tal, som bör beaktas vid studiet av fonetisk variation. Kunskap om talarålder kan användas för att förbättra talteknologiska tillämpningar såsom automatisk taligenkänning och talsyntes. Föreliggande doktorsavhandling beskriver sex studier som undersökt ett flertal aspekter av den åldersrelaterade variationen i tal.

När talapparaten förändras från tidig vuxen till hög ålder, påverkas talet på flera vis. Människor kan ganska väl bedöma talarålder med hjälp av ledtrådar i bl.a. röstläget, taltempot och röstkvaliteten. Det är dock ännu oklart vilka ledtrådar som är viktigast. Den första studien i denna avhandling undersökte hur grundtonsfrekvens (F_0) och taltempo (ordduration) påverkar lyssnares perception av talarålder. Resultaten visade att dessa drag verkar vara mindre viktiga än spektrala drag (t.ex. formantfrekvenser), men också att båda dessa drag ändå korrelerade med både kronologisk och bedömd ålder. I den andra studien jämfördes två olika stimulustyper (ord och spontantal) av olika längd. Det visade sig att större stimulusdurationer (oavsett typ) verkar förbättra bedömning av kvinnlig talarålder, medan spontantal (oavsett duration) verkar innehålla viktigare ledtrådar för perception av manlig ålder.

I de två följande studierna konstruerades flera automatiska bedömare av talarålder. Med dessa undersöktes en mängd akustiska drag som kan vara relevanta vid maskinbedömning av ålder, där prosodiska drag verkade vara viktigare för uppskattning av kvinnlig ålder, men spektrala drag (t.ex. F_2) för manlig. De automatiska åldersbedömarna uppnådde dock inte samma prestanda som mänskliga lyssnare.

Även om åtskilliga akustiska korrelat till talarålder är kända, har deras relativa betydelse ännu inte fastställts. I nästa studie analyserades 161 akustiska drag, som mättes automatiskt i sex ord uttalade av 547 talare. Normaliserade medelvärden användes för att göra direkta jämförelser av de olika dragen. Taltempo (segmentduration) och intensitetsomfång identifierades som de viktigaste akustiska korrelaten till talarålder. F_0 och en del spektrala drag (t.ex. F_1 och F_2) verkar dock också kunna användas som åldersledtrådar – åtminstone tillsammans med andra drag.

Syntetiskt tal skulle kunna låta mer naturligt om talarålder ingick som en parameter. I den sista studien utvecklades ett forskningsverktyg för simulering av talarålder med datadriven formantsyntes och åldersviktad linjär interpolation mellan åldrarna hos fyra kvinnliga referenstalare. En utvärdering av verktyget visade att syntetiska röster med simulerad ålder bedömdes som ungefär lika gamla som naturliga röster i samma ålder. Verktyget kommer att användas i vidare studier för analys genom syntes av talarålder.

Acknowledgements

First of all, I would like to thank my three supervisors for being the best supervisors I could possibly wish for. Thank you Per Lindblad, my main supervisor, for your deep understanding and contagious enthusiasm for the field of phonetics, for our many inspiring discussions, for suggesting countless improvements to my manuscript, and for remembering what it is like to be a PhD student. Thank you Gösta Bruce, my assistant supervisor, for your confidence in me, for sharing your deep knowledge of phonetics (especially prosody) with me, and for being there whenever I had any questions or problems. Thank you Rolf Carlson, my other assistant supervisor, for your intellectual support and excellent guidance in the field of speech technology, for our many fruitful discussions of my work, and for all your help with GLOVE.

Second, I would like to thank my "unofficial" supervisor Johan Frid for encouraging me to work with the Praat and Wagon toolkits (and for rescuing me when I encountered problems), for writing the first version of the feature extraction script, for letting me use his *Mix Prosody!* script and aligner, and for always patiently answering my endless questions.

Third, I would like to thank Bengt Sigurd, who read my manuscript carefully and suggested numerous improvements.

I am also especially grateful to Christian Müller, for letting me use the m3iCat toolkit and for all his help with generating the analysis results.

A very warm thanks also to Malgorzata Andreasson, Johan Dahl, Birgitta Lastow, Britt Nordbeck and Anders Sjöström for their help with all that technical and bureaucratic stuff.

I would also especially like to thank Joost van de Weijer for his patience and invaluable help with my statistical analyses.

Thank you Johan Segerbäck for scrupulously correcting my English and for providing numerous useful comments on and corrections to my manuscript.

Thanks go to everyone involved in the SweDia 2000 project, especially the Lund group. If it had not been for them, I would probably never have become interested in speaker age. Thank you Mechtild Tronnier for suggesting that it might be fun for me to enter the field of speech technology. It was (and still is)!

I owe a considerable debt of gratitude to my subjects whom I recorded for my synthesis experiments as well as to supplement the SweDia corpus, and to my students who participated in my perception experiments. Without their help, I would never have completed the thesis.

Thanks also go to my colleagues at the Department of Linguistics and Phonetics for the warm and stimulating academic environment which they create. Special thank-yous go to Elisabeth Zetterholm, Anastasia Karlsson

and Victoria Johansson for helping me with lots of practical stuff.

I would also like to extend a posthumous thanks to Thore Pettersson, for manipulating me into finishing my first term paper for the introductory course in linguistics in 1988, and for predicting that I would get this far. Thore, the road was longer and had more turns than I could imagine, but somehow I managed to stay on it.

A warm thanks to all my other colleagues – phoneticians and speech technologists – in Sweden and the rest of Europe. Thank you Hartmut Traunmüller, Anders Eriksson and Inger Karlsson, just to mention a few with whom I have had fruitful discussions about my work.

I would also like to extend my gratitude to everyone involved in the Swedish National Graduate School of Language Technology (GSLT), who taught me so much about speech and language technology.

And thank you everyone else who has helped me in my work!

Finally, I would like to thank my family: my mother for always believing in me and for encouraging and supporting me, my father for never giving up on me (Papa, ich weiss dass Du jetzt stolz auf mich bist!), my sisters Madeleine and Liselotte for their encouragement and support whenever I needed it, my grandmother Elly for patiently putting up with recording sessions, my cat Vincent for keeping me company all those hours by the computer and most of all Lars for his endless patience with me throughout these last couple of years, for never complaining about all the time I spent working on my thesis, for always listening to me, for being my own personal computer consultant, and for making me feel like Kraftwerk...

Contents

List of Abbreviations

Abbreviation	*Description*
Amp SD	amplitude standard deviation (see p. 84)
ANN	artificial neural networks (see p. 51)
ASR	automatic speaker recognition (see p. 49)
ATRI	intensity of the strongest amplitude modulation (see p. 87)
B_1, B_2, B_3, B_4, B_5	the first five formant bandwidths (after Fant, 1960)
BN	Bayesian networks (see p. 55)
CA	chronological age (see p. 8)
CART	classification and regression trees (see p. 52)
ΔMFCC	first-order time derivative of a MFCC (see p. 50)
DT	decision trees (see p. 52)
F_0	fundamental frequency (see p. 10)
F_0 SD	F_0 standard deviation (see p. 83)
F_1, F_2, F_3, F_4, F_5	the first five formant frequencies (after Fant, 1960)
FFT	fast Fourier transform (see p. 49, see also FT)
FT	Fourier transform (see p. 49)
FTRI	intensity of the strongest frequency modulation (see p. 87)
GMM	Gaussian mixture models (see p. 51)
HMM	hidden Markov models (see p. 50)
HNR	harmonics-to-noise ratio (see p. 86)
ISB	inverse filtered spectral balance (see p. 118)
kNN	k-nearest neighbours (see p. 51)
L_1, L_2, L_3, L_4, L_5	the first five formant levels (after Fant, 1960)
LDA	linear discriminant analysis (see p. 51)
LPC	linear prediction coefficients (see p. 50) or coding (see p. 139)
LTAS	long-term average spectrum (see p. 86)
MDVP	Multi-Dimensional Voice Program (see p. 87)
MFCC	mel frequency cepstral coefficients (see p. 49)
NB	naive Bayes (see p. 55)
ND	normal distribution (see p. 53)
NHR	noise-to-harmonics ratio (see p. 87)
PA	perceived age (by human listeners) (see p. 8)
SAMPA	speech assessment methods phonetic alphabet (Wells, 2006)
SB	spectral balance (see p. 86)
SE	spectral emphasis (see p. 86)
SPI	soft phonation index (see p. 87)
ST	spectral tilt (see p. 86)
SVM	support vector machines (see p. 51)
SweDia 2000	Swedish dialect project (see p. 30)
VA	vocal age (see p. 8)
VOT	voice onset time (see p. 11)
VTI	voice turbulence index (see p. 87)

Chapter 1

General introduction

Every living being goes through the process of ageing. This is a very complex process, which affects an individual in numerous ways. Therefore it is not strange that the concept of age is addressed in most natural sciences and humanities disciplines.

In humans, ageing also involves changes in the way we speak. Our voices and speech patterns change from early childhood to old age. Although most changes occur in childhood and puberty, age-related variation can also be observed throughout our adult lives into old age. Consequently, our age is reflected in our speech, and speaker age can be – and has been – studied using several methodological approaches, mainly acoustic analysis and perception experiments. The present thesis is a perceptual and acoustic-phonetic study of mainly adult speaker age, which also includes some speech technology approaches.

This chapter offers a general introduction to the thesis. After a background and motivation, a brief review is given of the age-related changes in the speech production mechanism. It is followed by definitions of several central terms used in the thesis. The chapter ends with a description of the focus and scope of the thesis and a general outline.

1.1 Background and motivation

Age-related variation in adult speech has been studied extensively since the 1960s, and our knowledge is continuously increasing. The majority of the studies have concerned perceptual and acoustic aspects, though some speech technology approaches have been followed as well. However, owing to the complexity of the ageing process, more research is needed to fully understand age-related variation in speech.

1.1.1 Human perception of age

Human listeners are able to judge speaker age at accuracy levels much better than chance. Apparently, we rely on numerous perceptual cues, including pitch, speech rate, loudness and voice quality. In addition, a large number of other factors may influence age perception. These can be related to (1) the speaker, e.g. gender, physiological condition and language spoken, (2) the listener, e.g. age, culture and motivation, (3) the speech sample, e.g. stimulus type (such as read or spontaneous speech) and length, and (4) the task, e.g. whether it involves classifying speakers into two or more age groups or making an exact estimation of age. The numerous studies with listening experiments are difficult to compare because of differences in subjects, speech material and method. Owing to the probable influence of these factors, there is no single answer to the question of how accurate listeners' judgements of speaker age actually are. Furthermore, although most studies carried out so far have found pitch and speech rate to be the most important perceptual cues to speaker age, some recent studies have suggested that spectral qualities may also be important.

1.1.2 Acoustic analysis of age

Acoustic analysis of potential correlates of speaker age is imperative in order to understand what aspects of the speech signal are affected by speaker age. Numerous studies have investigated several acoustic features, including F_0 and F_0 stability, duration, resonance and correlates of voice quality. However, it is still unclear which acoustic features constitute the most important cues to speaker age and how they relate to each other. As in the perceptual studies, factors such as subjects, speech material and method may affect the findings of the analysis. Identifying the features in speech which constitute the best correlates of speaker age is thus a very complex task which awaits further investigation.

1.1.3 Speech technology approaches

Machine recognition of speaker age has been studied only to a limited extent, even though there are numerous applications where automatic age classification can be useful. For instance, objective age recognisers could be used in forensic phonetics to obtain good age estimates of perpetrators from recordings, thus facilitating the elimination of suspects of other ages. Spoken dialogue systems could benefit from age recognition, for instance in the provision of user-adapted information to a particular age group, such as the elderly or teenagers. Approaches involving automatic age recognition have

mainly used cepstral and perturbation features in combination with machine learning techniques, including hidden Markov models (HMM) and artificial neural networks (ANN).

So far, very little research on synthesis of speaker age has been carried out. Further studies in this field could lead to increased naturalness in synthetic speech and be used to personalise reading aids and voice prostheses.

1.1.4 General purpose and aim

Up till now, there have been few attempts to integrate classical phonetic and speech technology approaches in order to investigate some of the remaining problems in speaker age research. Such problems include the identification of the most important acoustic age correlates as well as the integration of age characteristics in speech recognition and synthesis in order to increase the naturalness of human–machine communication. The purpose of the present thesis is to use perceptual and acoustic-phonetic as well as speech technology methods to study female and male speaker age from several perspectives.

The first specific purpose is to find out whether prosodic or non-prosodic features are more important in human age perception, and the second one is to investigate how the accuracy of human age perception varies with the type and duration of the speech sample (stimulus). A third purpose is to study the automatic recognition of speaker age using a large number of acoustic features and to determine which individual feature is the single most important one, as well as which combinations of features are important, in machine perception of age. The fourth purpose is to analyse a large speech material and make direct comparisons among numerous acoustic features in order to identify the most important acoustic correlates of speaker age. The final specific purpose is to analyse and simulate speaker age using formant synthesis.

The general aim is to contribute to knowledge about perceptual as well as acoustic aspects of speaker age. In the world today, the population aged 65 years and older is growing. It has increased from 117 million in 1950 to 390 million in 2005 (Statistiska centralbyrån, 2005). According to the United Nations, this number is expected to increase to 1.5 billion people aged 65 or older in 2050; this corresponds to twice the population of Europe in 2005. A related issue is the increasing number of elderly people and young children who use computers and similar technology (Clements, 1999). Only with increased knowledge about how speech varies with age can we develop speech technology applications which adapt to the age of the user. For instance, spoken dialogue systems (e.g. tourist guides and car or pedestrian navigation systems) could adjust their functions to suit the age of the user. The performance of speech recognition is also likely to improve if it is ad-

justed to age-related variation, especially to children's and elderly people's voices. Moreover, it could become possible for the vocally handicapped to personalise their voice prosthesis to suit their current age.

1.1.5 Why both phonetics and speech technology?

Research carried out by phoneticians has sometimes been criticised by speech technologists for studying too narrowly defined problems and abstract forms which are far from real spoken language, and for being too subjective to allow any general conclusions to be drawn about speech. On the other hand, phoneticians have criticised speech technologists for being ignorant about phonetic theory, for being too technical and for building too complicated tools (Öhman, 2001; Greenberg, 2001; Fant, 2005). Fortunately, it looks as if the knowledge gap between phonetics and technology is beginning to narrow (see e.g. Frid, 2003 and Schaeffler, 2005). This thesis aims at contributing further to bridging the gap between these related disciplines, for instance by applying speech technology methods to several phonetic problems. It addresses several aspects – both phonetic and speech technological – within the scope of speaker age, including human perception, automatic segmentation and feature extraction from large speech corpora, automatic recognition of speaker-specific qualities and speech synthesis. In fact, one goal has been to prove that this can be done without the phonetician first having to become an engineer.

1.1.6 Speaker age in information theory

Speaker age has been defined by researchers either as a *symptom* (*Kundgabe*) (Bühler, 1934; Trubetzkoy, 1958) or as an *organic* (Traunmüller, 2000, 2005), *paralinguistic* (Lindblad, 1992; Traunmüller, 2005; Müller, 2006), *extralinguistic* (Laver, 1980, 1991, 1994; Marasek, 1997) or *non-linguistic* (Roach et al., 1998; Fujisaki, 2004) feature. Though defined somewhat differently, such features have one thing in common: they are perceptually easily distinguished from linguistic features as they do not, for the most part, alter or obscure the identity of linguistic elements.

For example, Fujisaki (2004, p. 1) defined *linguistic* information as "the symbolic information that is represented by a set of discrete symbols and rules for their combination". This type of information is discrete and categorical. On the other hand, *paralinguistic* information is defined by him as "the information that is not inferable from the written counterpart but is deliberately added by the speaker to modify or supplement the linguistic information", and can be both discrete and continuous. According to Fujisaki

Table 1.1: Information conveyed in speech, according to Fujisaki (2004)

Category	*Examples*	*Discrete/Continuous*
linguistic	lexical (word, accent, etc.) syntactic (phrase structure, etc.) pragmatic (discourse, focus, etc.)	discrete (symbolic) controlled by speaker
paralinguistic	intentional (exhortation, etc.) attitudinal (politeness, etc.) stylistic (fast, slow, etc.)	discrete and continuous can be controlled by speaker
non-linguistic	physical (age, gender, etc.) emotional (joy, sorrow, etc.) idiosyncratic	discrete and/or continuous generally cannot be controlled, but can be simulated

(2004), *non-linguistic* information "concerns such factors as the age, gender, idiosyncracy, and physical and emotional states of the speaker, etc.". Although the speaker may control his or her way of speaking to simulate e.g. an emotion, these features cannot generally be controlled. *Non-linguistic* features can be both discrete and continuous. Table 1.1 summarises the information conveyed in speech, according to Fujisaki.

In contrast to Fujisaki, in this thesis the term *paralinguistic* will be used to denote all aspects of speech – including speaker age – which are not considered linguistic.

After this brief presentation of the theme, the purpose and the aim of the thesis, the following section addresses one important aspect of speaker age, namely what actually happens to our speech mechanisms as we grow older.

1.2 Ageing of the speech production mechanism

From young adulthood to old age, the speech production mechanism undergoes numerous anatomical and physiological changes, which have not all been fully explored. For instance, there are substantial gender differences in the extent and timing of the ageing process (Beck, 1997; Linville, 2001). Moreover, the physiological differences between individuals seem to grow with advancing age (Ramig and Ringel, 1983). It is also important, but sometimes difficult, to distinguish among age-related, disease-related and environment-related changes in speech. Linville (2000, 2001, 2004) has provided excellent reviews of the numerous changes occurring in speech as we grow older. This section is mainly based on her work.

1.2.1 Respiratory system

Changes in the respiratory system affect speech breathing as well as the voice. The respiratory system reaches its full size after puberty but continues to change throughout adulthood to old age. Changes include decreased lung capacity (mainly due to loss of elasticity in lung tissue), stiffening of the thorax and weakening of respiratory muscles.

1.2.2 Larynx

The age-related changes of the larynx after it has reached its full size in puberty are numerous, and they affect mainly fundamental frequency and voice quality. Ossification of cartilages occurs later and is less extensive in females (fourth decade) than in males (third decade), while calcification probably occurs later than ossification in both females and males (cf. Jurik, 1984; Lindblad, 1992; Dedivitis et al., 2004; Mupparapu and Vuppalapati, 2005).

Muscle atrophy occurs in all intrinsic laryngeal muscles. As research has focused on the vocal folds, we do not know to which extent other intrinsic muscles are affected. Whether there are any gender differences is also still unclear. The changes in the complex structure of the vocal folds with increased speaker age are substantial. Besides general degeneration and atrophy, the folds shorten in males (particularly after age 70). Also, the epithelium (the thin outer protective layer of tissue) thickens progressively in females, especially after age 70, while it thickens in males up to age 70 but then grows thinner again. The mucous glands reduce their secretions, leading to less hydrated vocal folds, particularly in males. There also seems to be some evidence of laryngeal nerve degeneration, as well as some changes in the blood supply to the laryngeal muscles.

1.2.3 Supralaryngeal system

Changes in the supralaryngeal system may also affect speech. The craniofacial skeleton grows continuously by about 3–5% from young adulthood to old age. Muscle atrophy occurs in the facial, mastication and pharyngeal muscles. A slight lowering of the larynx in the neck increases the length of the vocal tract. Extensive degenerative changes occur in the temporomandibular joint, including a gradual reduction in size and reductions in blood supply. In the oral cavity, the mucosa grow thinner and lose elasticity, which is most apparent after age 70, and the mucosal surface roughens. Changes in the pharynx and soft palate include thinning of the epithelium, muscle atrophy and decreased sensation. The tongue surface becomes thinner and fissured,

while the tongue muscles suffer from atrophy and fatty infiltration, beginning as early as in the second or third decade.

1.2.4 Neuromuscular control

The effects of ageing on motor function can be observed in both the peripheral and the central nervous system. They may affect speech rate, coordination of articulators and breath support as well as the regulation of F_0. Peripheral changes include a type of "dying back" neuropathy, where the distal ends of the nerve fibres are affected earlier. Also, the number of motor units declines and conduction velocity slows down slightly.

Central changes include a decline in brain weight from age 20 to 90 by about 10% as well as a decrease in brain size. There are reports of decreases in the number of nerve cells in the cortex as well as age-related changes in these cells, which may slow down motor movements. In addition, dopamine levels in the brain may decline by up to 50%, leading to slower sensorimotor processes.

1.2.5 Female and male ageing

In addition to what has already been mentioned, a few more words deserve to be said about the differences between female and male ageing. These are often related to the timing and extent of age-related changes throughout life. One obvious difference is the dramatic changes occurring in males at puberty; another is that females experience greater changes around menopause. Nevertheless, the age-related changes in adults are generally greater in men than in women as regards (1) the extent of laryngeal structure change, (2) fine-motor control of laryngeal abductory and adductory movements, (3) tongue movements and (4) speech rate. It has also been noted that the mucous membranes in the larynx are more sensitive in females than in males and that females may thus be more vulnerable to age-related changes in this respect (P. Kitzing, personal communication, 31 January 2006). On the other hand, men and women display similar age-related changes in speech breathing.

1.3 Definitions

Brief definitions or descriptions of terms are generally given the first time that each term appears in the text. For convenience, definitions of several central concepts used throughout the thesis have been collected in this section. Concepts related to speaker age are treated first, and then some other general phonetic and speech technology terms are defined.

1.3.1 Definitions of age and speaker age

The concept of age may – at first glance – seem to be easy both to define and to measure. However, there are several problems involved in this task, for instance whether to begin measuring at birth or at the time of conception. Human age is usually measured in years, but what are we measuring exactly? Several definitions of age have been proposed. Some are briefly described here, and the definitions used in this thesis are summarised in Table 1.2.

One very important measure is *chronological age* (CA), sometimes called calendar age. CA is often defined as the time from birth to the present (Cavanaugh, 1999). Another term, *biological age*, has sometimes been used as a synonym of CA. However, since CA per se is not a good predictor of biological (functional or physiological) age (Sprott and Roth, 1992), owing to factors connected with the psychological, physiological, neurological and biochemical manifestations of the ageing process, it is more appropriate to regard them as two separate concepts (Hollien, 1987). Moreover, old age should not be confused with pathology. The ageing process in itself is no disease, as most older people are in fact healthy.

Since psychological and behavioural (cultural and social) factors also exert a strong influence on the ageing process, alternative definitions of age have been proposed. *Psychological age* is defined as the sum of long-term changes in the personality, identity and cognitive systems (Linville, 2001). *Socio-cultural age* is defined as the extent to which a person demonstrates the age-dependent behaviours and habits expected in a culture, including customs, language and interpersonal style (Linville, 2001; Cavanaugh, 1999). The concept of age thus has to do not only with chronology and biological changes, but also with how old a person feels or is felt to be and even how old a person sounds (Hollien, 1987; Mulac and Giles, 1996).

Research concerned with speaker age has yielded two definitions of the term *speaker age*, namely *perceived age* (PA) and *vocal age* (VA). In perception studies, *perceived* (or *estimated*) *age* is defined as the age of a speaker as subjectively perceived by listeners. In many cases, the mean value of the estimates made by a group of perceivers is used as a measure of PA. The CA and PA of the same speaker may differ more or less depending on, among other things, the above-mentioned effects of physiological and behavioural factors on the speech signal. In previous research, PA has sometimes been used to denote *perceived age class*, for instance when speakers have been classified as either old or young. When estimated in exact years, PA is often referred to as *direct age*.

Vocal age (VA) is defined as the sum of the effects of long-term changes in the speech apparatus at a certain stage in life, as observable in the acoustic speech signal (Linville, 2001; Brückl, 2002). VA is affected by physiological,

psychological and socio-cultural factors. A problem with this term is that it could be taken to imply that only voiced (vocal) sounds are concerned. As this thesis investigates voiced as well as voiceless speech, the term *vocal age* is avoided.

Table 1.2: Definitions of (speaker) age used in this thesis

Term	*Definition*
chronological age (CA)	the time elapsed from the date of birth to the present day (often measured in years)
perceived age (PA)	the age of a speaker as subjectively perceived by a listener (or the mean value estimated by a group of listeners)
speaker age	the age (CA or PA) of a speaker or a group of speakers

In this thesis, the term *speaker age* is used in a general sense to denote either the CA or PA of an individual or a group of speakers, i.e. the age of one or several individuals who have produced a speech sample of some kind. Whenever a distinction is needed, the more specific term *chronological age* or *perceived age* is used.

1.3.2 Other general definitions

In addition to the concepts defined in the previous section, several other technical terms are used in the thesis. Some of the more important ones are described here. The definitions of the two distinct concepts of *perceptual cues* and *acoustic correlates* follow the ones proposed by Heldner (2001). Unfortunately, these two notions are sometimes confused. In this thesis, the term *perceptual cues* refers to the acoustic-phonetic information which the listener is able to use when perceiving a quality in speech, such as speaker age, while the term *acoustic correlates* is used to mean the acoustic-phonetic features of the speech signal which can be measured objectively. When reference is made at the same time to both acoustic correlates and perceptual cues, the term *phonetic cues* is used – for instance when acoustic measures are used in the automatic recognition of age and the results are then compared with human perception of age.

Throughout the thesis, the term *acoustic* is used for acoustic as well as temporal characteristics, including features such as the number of syllables per second.

In most parts of the thesis, the term *speech rate* is used in a general phonetic sense. In Chapter 4, however, the meaning of the term is used in a general *acoustic* sense, and also includes temporal measures.

The term *duration* is generally defined as the time – measured in seconds (s) or milliseconds (ms) – that a phonetic segment (e.g. a word or a phoneme) lasts. However, *duration* has been given two special definitions in the two

perceptual studies presented in Chapter 2. In Study 1 (see Section 2.6), it is defined as the actual acoustic duration of the same stimulus type, namely the word *rasa*. A slightly different definition is used in Study 2 (see Section 2.7). Here, *duration* denotes stimulus duration, i.e. the time (in seconds) during which listeners are exposed to a certain stimulus.

In Chapter 2, the term *spectral* is used as a synonym of *non-prosodic*, defined here as everything but F_0 and duration, while in Chapter 3 *spectral* is used to denote a group of acoustic features comprising all resonance and inverse filtered features.

Throughout the thesis, the term *relative intensity* refers to the intensity of a segment relative to other parts of the sample (e.g. word). All speech samples used in the studies presented here have been *normalised for intensity* by setting the maximum intensity of all samples to the exact same value using a built-in function in the speech analysis tool Praat. This was done to reduce analysis errors. For instance, it is hard to tell if an overall low intensity is due to a speaker talking softly or just being far away from the microphone. As a consequence, only segment intensity relative to other parts of the word can be measured.

The word *intensity* is used throughout this thesis in the sense of *sound pressure level* or SPL (also referred to as *intensity level*), which is measured in decibels (dB) and not in watts per square metre (W/m^2). This is in line with a common usage in phonetics and related disciplines. However, in the acoustic sciences and also often in phonetics, the distinction between these two concepts is made clear through use of the terms *intensity* (measured in W/m^2) and *sound pressure level*/*intensity level* (measured in dB). This usage is generally considered to be preferable. The word *intensity* should therefore be read as *sound pressure level* throughout the thesis. Moreover, the term *LTAS amplitude* is used in the sense of *LTAS level*, measured in dB.

When reference is made to the fundamental and formant frequencies, the notation F_0, F_1, F_2, etc., will be used (cf. Fant 1960; Lindblad 1992), while the notation F0, F1, F2, etc., used in Chapter 3 will denote feature groups related to the fundamental component and the formants, including mean and median frequency as well as frequency range and standard deviation. In Study 4, formant bandwidths and levels are also included in the formant feature groups.

The abbreviation ASR will be used for *automatic speaker recognition*, not for the more common *automatic speech recognition*.

Table 1.3 contains short definitions (in alphabetical order) of some general terms often appearing in the thesis. Additional abbreviations are explained in the list of abbreviations (see p. ix).

Table 1.3: Some general definitions of words and concepts used in the thesis

Term	*Definition*
acoustic correlates	the objectively measurable features of the speech signal which relate to a specific quality in speech, e.g. speaker age or breathy voice quality
acoustic features	the objectively measurable features of the speech signal (including temporal features, e.g. the number of syllables per second)
duration	the acoustic length of a speech sample, often measured in seconds (s) or milliseconds (ms)
harmonics-to-noise ratio (HNR)	a measure of the relative amount of noise in the voice signal, calculated as the average ratio (in dB) of the overall harmonic spectral energy to the overall inharmonic spectral energy
jitter	cycle-to-cycle frequency variations in the fundamental period of vocal fold vibration (see Figure 4.2, p. 85)
noise-to-harmonics ratio (NHR)	a measure of the relative amount of noise in the speech signal, calculated as the average ratio (in dB) of the inharmonic spectral energy (in 1.5–4.5 kHz) to the harmonic spectral energy (in 0.07–4.5 kHz)
perceptual cues	the acoustic-phonetic information used by listeners when perceiving a quality (e.g. age) in speech
phonetic cues	acoustic correlates as well as perceptual cues
plosive closure	the period of time between the cessation of formants for the preceding vowel and the plosive release (except in phrase-initial position after silence)
relative intensity	the intensity of a segment relative to other parts of the sample (the only measure of mean and median intensity used in this thesis; see p. 10)
resonance	the filtering of speech sounds in the supralaryngeal cavities owing to the size and shape of the vocal tract
shimmer	cycle-to-cycle amplitude variations in the fundamental period of vocal fold vibration (see Figure 4.2, p. 85)
spectral	that which is measured or calculated from the spectrum of a speech sample
spectral features	features which are measured or calculated from the spectrum of a speech sample; used in this thesis as a synonym of "non-prosodic features"
speech rate	the tempo in speech, measured e.g. in syllables per second or in segment duration
voice onset time (VOT)	the interval between the release of a plosive and the start of voicing
voice quality	"the characteristic auditory coloring of an individual speaker's voice" (Laver, 1980, p. 1)

1.4 Scope and limitations of the thesis

Speaker age is a broad area of research. It includes aspects of production, acoustic analysis and perception of speech. As reflected in its title, the focus of the present thesis is on perception, acoustic analysis and synthesis. The studies presented here investigate perceptual and acoustic features in relation to age, and they also apply a few speech technology methods for the automatic recognition and synthesis of speaker age. Production, speech pathology and (bio)medical (including hormonal) issues are not addressed.

Speaker age may also be recognised by means of several linguistic cues, including choice of words and semantic content. Such linguistic factors are only briefly discussed in relation to non-phonetic factors which influence the impression of speaker age.

Moreover, not everything encompassed in the concepts focused upon is treated. In fact, almost only variation in adult speaker age is studied. The drastic changes occurring in children are beyond the scope of the thesis. However, one child speaker is studied in Chapter 5, which concerns synthesis of speaker age. Furthermore, the only language studied is Swedish, although some previous experiments involving other languages are reviewed. Since the focus is on perception and acoustic aspects, only two features which may be classified as not strictly acoustic are studied, namely the number of syllables and phonemes per second. Several researchers have included closely related features, such as intra-oral air pressure as well as measures of electroglottography (EGG) and airflow, in their studies. In this thesis, such features are described only when reference is made to previous studies.

As for the speech technology aspects of this thesis, only one pattern matching technique (classification and regression trees) and one synthesis approach (formant synthesis) are investigated. However, some other methods are described briefly.

1.5 Thesis outline

The main part of the thesis is divided into four chapters, each addressing a specific aspect of speaker age.

Chapter 2 concerns human perception of speaker age. It provides a review of previous research and discusses known measures of accuracy, perceptual cues and non-phonetic factors which influence perception. Moreover, the chapter presents two studies of human age perception, which are based on the research first described in Schötz (2004) and Schötz (2005b). The first investigates how F_0 and speech rate (measured as word duration) influence age perception, while the second concerns effects of stimulus type and duration in the perception of speaker age.

Chapter 3, on machine perception of speaker age, briefly describes some common methods used in the field of automatic speaker recognition, and it also reviews some known research done on machine classification of speaker age. In addition, it presents two experiments concerning recognition of speaker age using the CART (classification and regression trees) technique which were first described in Schötz (2005a) and Schötz (2006b).

Chapter 4 concerns acoustic analysis of speaker age, and begins by discussing known acoustic correlates of age and by giving an overview of

previous acoustic studies. It also describes a study where a large number of acoustic features were compared in order to identify the most important acoustic correlates of age.

Chapter 5 addresses synthesis (simulation) of speaker age. After a short introduction to speech synthesis in general and formant synthesis in particular, an experiment with data-driven formant synthesis of speaker age using age-weighted linear interpolation, based on the research described in Schötz (2006a), is presented.

The thesis ends with a concluding discussion and some ideas for future work in **Chapter 6**.

Chapter 2

Human perception of speaker age

2.1 Introduction

Most people are able to estimate an individual's age from speech samples alone at accuracy levels significantly better than chance (Ptacek and Sander, 1966; Ryan and Burk, 1974; Huntley et al., 1987; Linville, 2001), perhaps because of constant confrontation with this task throughout our lives, e.g. when listening to someone on the telephone or radio (Shipp and Hollien, 1969). However, the numerous perception studies of speaker age carried out so far have varied considerably in method and speech material, as well as in speaker and listener characteristics. Different studies are often difficult to compare. Therefore, we are still unable to tell exactly how well listeners are able to judge speaker age. The choice of cues and the accuracy obtained seem to depend on the type and length of the speech samples (Ramig, 1986). Moreover, the relationship of the perceptual cues used by listeners in age estimation with the acoustic correlates of chronological as well as perceived age has still not been fully established. In fact, the cues used by listeners to estimate speaker age do not always correspond to age-related changes which can be measured acoustically (Linville, 2000).

This chapter covers several aspects of age perception. It begins by presenting some known perceptual cues, measures of accuracy and factors which may influence age perception, and continues with a brief overview of the previous related research. The next part of the chapter describes two perceptual studies which try to answer some of the remaining questions in this field. The first study investigates the role of F_0 and word duration in age perception and their relation to non-prosodic cues, while the second study concerns the effects of stimulus type and length in the perception of speaker age.

2.2 Perceptual cues to speaker age

Perceptual cues to speaker age are auditory cues, i.e. voice and speech characteristics, which listeners have reported that they use when judging the age of an individual speaker. The less acoustic information present in a speech sample, the more difficult the age estimation task becomes; but even with very little information, we are still not reduced to random guessing (Linville, 2001). We are likely to use different (sets of) cues depending on the type and duration of the speech sample. For instance, when listening to one minute of spontaneous speech, we are often able to find age cues in the speech rate, dialect, semantic content and choice of words. In a single sustained vowel such cues are absent, and we have to rely on other cues, including pitch and voice quality.

In several studies, listeners have been asked to mention cues which they believe have influenced their judgements of speaker age. Table 2.1 shows an overview of perceptual cues reported in previous studies.

Table 2.1: Perceptual cues to increased speaker age

Perceptual dimension	*Feature and manner of change*
pitch	lower and less variable pitch
length, speech rate	longer segments/pauses, slower speaking/reading rate, increased hesitancy
loudness	reduced loudness
voice quality	increased hoarseness, harshness, strain, vocal tremor and breathiness
other dimensions affecting the speaking skill	less precise articulation, different phrasing (shorter breath groups), lack of smoothness, less vitality/vigour/liveliness, "sociolect of age"

Ptacek and Sander (1966) had subjects casually name perceptual indicators of old age. The subjects mentioned different phrasing, reduced reading rate, lack of smoothness, increased hesitancy, increased hoarseness and strain, lower and less variable pitch and reduced intensity and vitality. Ryan and Burk (1974) showed that air loss, slow articulation rate, voice tremor, laryngeal tension and imprecise consonants may be strong perceptual cues to old age. Hollien (1987) listed the following perceptual age cues reported by previous studies: changes in pitch, vocal loudness, vocal fry and tremor, increased breathiness and slower speech production. Linville (2001) reported ten vocal features which have been considered as typical of "old" speakers in previous research: lower vocal pitch (regardless of speaker gender), increased hoarseness or harshness, increased strain, vocal tremor, increased breathiness, reduced loudness, slower speech rate, greater hesitancy, less precise articulation and longer duration of pauses. Linville (2001) also suggests that, apart from listeners mentioning "imprecise articulation" as a cue to

old age, little is still known about how resonance information affects the perception of age. Minematsu et al. (2002a,b) found that listeners mentioned low speech rate, quavering voices, little vigour and loudness, and inarticulate sounds as being among the features which influence age perception.

2.3 Non-phonetic factors which may influence perception of speaker age

Besides the phonetic information used by listeners as perceptual age cues, a large number of other factors may influence the perception of speaker age. These factors are discussed in this section. Figure 2.1 shows four groups of contributing factors, where the individual factors have been tentatively ranked within each group according to their possible importance, based on the findings of previous research.

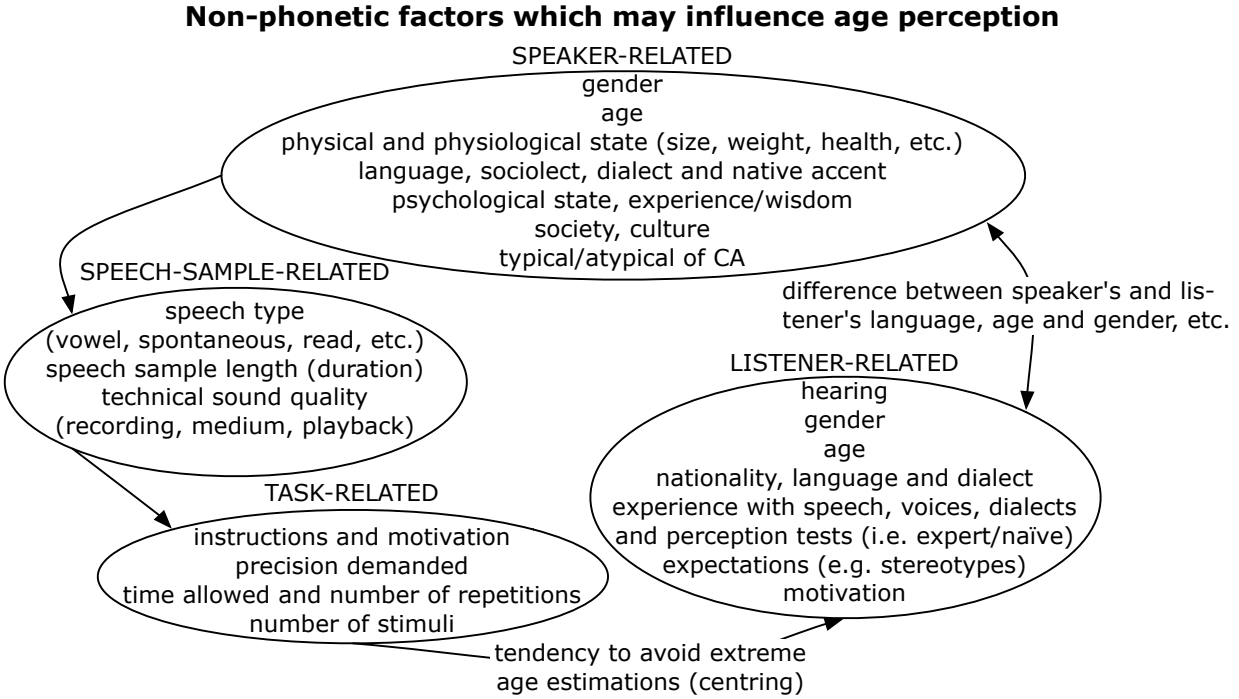


Figure 2.1: Some non-phonetic factors which may influence age perception

2.3.1 Speaker-related factors

Gender may determine the strategy and (set of) age cues used by listeners, as women and men age differently; a large number of gender-related acoustic changes with advancing age have been reported, for instance in F_0 (Higgins and Saxman, 1991; Linville, 2001).

Age group and age range seem to have a strong effect on age perception. A large number of studies have reported that listeners tend to overestimate the age of young voices while underestimating that of older voices (Shipp and Hollien, 1969; Huntley et al., 1987; Braun and Cerrato, 1999; Minematsu et al., 2003). This phenomenon is known as *centring* (Braun and Cerrato, 1999). However, not all studies agree. Hartman (1979) reported a general tendency for listeners to underestimate speaker age. Moreover, young and old speakers are often judged with different precision. Jacques and Rastatter (1990) reported that elderly speakers were less correctly identified as old (by both young and old listeners) than young speakers as young. Brückl and Sendlmeier (2003) found that listeners significantly underestimated middle-aged female speakers, and they also suggested that listeners try to avoid overestimating the age of older speakers.

Physical and physiological state, i.e. size, weight and the physiological condition and health of the body, including the speech organs (Ramig and Ringel, 1983; Ramig et al., 1985; Ringel and Chodzko-Zajko, 1987), such as the effects of cigarette smoking (Braun and Rietveld, 1995), is another large group of factors which affect age perception. Non-smoking and non-drinking healthy speakers who take regular exercise generally sound younger than smokers, alcoholics and overweight, tired and sick speakers. Increased physiological variation in elderly speakers has been reported by Linville (2001), Ramig and Ringel (1983) and Ramig (1983, 1986).

Language, sociolect, dialect and degree of native accent may also affect the perception of age (Linville, 2001). If the language is known to the listener, semantic and dialectal age cues can be used, while listeners have to rely on other cues when judging a speech sample of an unknown language (Brückl and Sendlmeier, 2003). However, Braun and Cerrato (1999) found no significant difference in performance for listeners familiar and unfamiliar with the language spoken. Ramig (1986) suggested that a "sociolect of age", i.e. a learned speech production pattern which is seen as typical of a certain age group, may be used by some speakers. Stölten and Engstrand (2003) found that estimated speaker age is likely to have a large effect on estimations of dialectal strength, and they suggested that listener judgements of dialect strength must be corrected for age effects.

Psychological state, including emotion, attitude, mood, alertness (or tiredness), etc., is included in the phonetic information of the speech sample. It is possible that a tired speaker may sound older than an alert one, although this needs to be verified by research. Other factors related to psychological state are e.g. experience, wisdom and authority. An individual who has ex-

perienced a great deal may sound older than less experienced people of the same age. Similarly, wisdom and authority may also add to the perceived age of a voice.

Social and cultural factors could also influence age perception. If the speaker comes from a cultural and social background which is familiar to the listener, the accuracy of age estimation may increase.

Typicalness or atypicalness of speakers relates to the difference between speakers whose age seems easy to estimate and speakers who seem to be more difficult to judge, probably owing mainly to the physiological and environmental factors mentioned in this section. There seems to be a subdivision into typical and atypical speakers, the former displaying a stronger correlation between PA and CA (Morris and Brown, 1987; Ringel and Chodzko-Zajko, 1987; Schötz, 2001a). In Linville (1987), several old speakers were identified as young by the majority of the listeners. Braun (1996) suggested that one reason for this may be discrepancies between biological and chronological age. Ringel and Chodzko-Zajko (1987) argued that such inter- and intra-individual variation in age-related changes depends on the life style (e.g. exercise, nutrition and health) of the individual. Atypical speakers – sometimes referred to as *difficult voices* (Braun, 1996) – are speakers whose age is difficult to assess, either in that estimates vary greatly among listeners or in that they are consistently judged as being much older or younger than their CA. This category has been defined as having a standard deviation of nine or more years in listening experiments (Shipp and Hollien, 1969).

2.3.2 Listener-related factors

Obvious listener-related factors include hearing impairment, illness, the effects of medical or narcotic drugs, etc. For instance, hearing-impaired listeners cannot make use of the age cues in the same way as normally hearing listeners. There are numerous other factors of this type which may influence age perception.

Gender influence on age estimation is an area which is still unclear. Some studies have found females to perform better than males (Hartman, 1979), while others have found male listeners to perform somewhat better (Braun and Cerrato, 1999).

Listener age seems to influence age perception, although reports vary with respect to the nature and extent of this influence. Several studies suggest that the listener's age as well as the difference in age between the listener

and the speaker do influence estimation accuracy (Shipp and Hollien, 1969; Huntley et al., 1987; Braun, 1996; Brückl and Sendlmeier, 2003). Younger listeners have often performed better than older listeners (Kreiman and Papcun, 1985; Linville, 1987), especially when estimating the age of younger speakers. Shipp and Hollien (1969) noted that young listeners agreed more on what constituted a young than an old speaker, and they offer two possible explanations: (1) young listeners are more familiar with young voices, and (2) "young voices have more distinctly identifiable characteristics than old ones". Kreiman and Papcun (1985) found that the age of younger speakers was difficult to identify for older listeners. They explained this by reference to older individuals' decrease in ability to discriminate speech as well as their lack of experience with younger speakers. Huntley et al. (1987) observed that young and middle-aged listeners were better judges of age than adolescent and old ones, perhaps because they were assisted by their familiarity and personal experience. Jacques and Rastatter (1990) suggested that elderly listeners become less able to determine speaker age, perhaps owing to speech processing alterations which occur at older ages, as F_0 and resonance information seemed to be harder to process for elderly listeners in their study.

Nationality, language and dialect may influence age perception. Nagao and Kewley-Port (2005) observed that Japanese listeners were better at judging the age of both Japanese and English speakers than were English listeners. Cultural factors may emphasise the importance for a Japanese listener to be able to recognise the age of a speaker. For instance, the Japanese may use different levels of politeness depending on the age of the interlocutor (Daikuhara, 1986). As regards dialect, Stölten and Engstrand (2002) found a clear relationship between estimated age and estimated degree of dialect strength.

Experience or familiarity with similar speakers (age group, culture, dialect, etc.) may influence age estimation accuracy. Admittedly, Braun (1996) found no significant difference in the correlation between PA and CA for expert and naïve listeners in her experiment with a standardised text using high-quality recordings. However, the difference between the trained and untrained listeners also concerns the degree of agreement within each group, as evidenced in standard deviations. A standard deviation of up to six years is considered to indicate a close listener consensus (Horii and Ryan, 1981). In Braun (1996), only the expert group had a high level of agreement among listeners. She also investigated listener certainty, finding that untrained listeners were somewhat more certain than trained ones, and that men were more certain than women.

Expectations of what a young or old voice should sound like may also affect age perception accuracy (Huntley et al., 1987). Elderly individuals may use a "sociolect of age", meaning that learned patterns of speech production are part of what causes listeners' expectations about elderly speech to be met.

Motivation to perform well in an age perception experiment is also likely to influence judgement accuracy. However, it is hard to know if a listener is having a "bad day" or is preoccupied with matters other than the age estimation task. Ways of increasing motivation may be to pay the subjects for their efforts or to organise an informal contest (Which one of you is the best judge of age?).

2.3.3 Speech-sample-related factors

The speech types of stimuli used in previous studies include phonated, low-pass filtered or whispered vowels of various durations as well as isolated words (mono-, bi- or multisyllabic), phrases (in carrier sentences), read sentences (played forward or backward) and spontaneous monologues. Age estimates from vowel stimuli are generally less accurate than those from read or spontaneous speech (Ptacek and Sander, 1966; Brückl and Sendlmeier, 2003), as shown in Table 2.3. Brückl and Sendlmeier (2003) obtained the best results with spontaneous speech, probably because this speech type was produced in the most natural way by the speakers and should therefore contain more age information, for instance in semantic content, choice of words and sociolect. Brückl and Sendlmeier (2003) also observed that in age estimation based on vowel stimuli, accuracy increased when the vowel onset was included. This is consistent with the notion of articulatory dynamics in speech being an important age cue.

Speech sample length has not been directly addressed as a variable influencing the accuracy of speaker age perception, although some studies have used stimuli of varying duration (Ptacek and Sander, 1966; Brückl and Sendlmeier, 2003). More research is needed to establish the minimal and optimal stimulus duration for age estimation tasks, as well as the duration beyond which no improvement is obtained.

Technical sound quality constitutes a major influence on age estimation performance. The better the sound quality of a speech sample (depending on both recording and playback conditions), the easier the age estimation task. For instance, telephone-transmitted speech has yielded less accurate results than HiFi speech (Braun and Cerrato, 1999). In addition, the recording and playback equipment (microphone, loudspeakers or headphones) and medium

(cassette, DAT, hard disk) as well as the recording situation may influence the result in some respect, and so may the type and amount (or absence) of background noise (anechoic chamber or home environment), etc. Moreover, the playback volume and – if loudspeakers are used – the acoustics of the test room may influence judgements.

2.3.4 Task-related factors

The instructions and motivation given to the participants of an age perception experiment may greatly influence results. Written instructions may sometimes lead to better results than oral ones, as they can be read several times and be better understood. Performance may also be enhanced if the subjects are paid or if they are encouraged to take part in an age estimation contest, where the best judge receives some kind of prize.

The precision demanded in previous age estimation experiments has varied from judging the exact number of years to classification into 2, 3, 5, 7 or 18 age groups. The precision requested has to be stated clearly in the instructions given to the participants before the test; otherwise listeners may choose their own level of precision, e.g. by decade, even if direct age estimations were intended by the experimenters. It is highly likely that the higher precision is demanded in a test, the lower the age estimation accuracy will be. A related problem concerns the boundaries between age classes. For instance, if the boundary between the two age classes of 30 and 40 was set at age 40, a listener who perceived a 39-year-old speaker as 30 years old would appear to be more correct than a listener perceiving the same speaker as 40 years old (Shipp and Hollien, 1969).

Time allowed and number of repetitions vary among listening tests. The time in which listeners must make their judgements (before the next stimulus is played) may influence performance. In addition, stimuli may be played once, twice or any number of times.

The number of stimuli may also influence age estimation performance. If too many stimuli are used in the test, listeners may become bored or tired. And if too few stimuli are used, errors due to the listeners' unfamiliarity with the test situation may occur, especially for the first few stimuli. This effect is sometimes referred to as the *learning effect*. To minimise such errors, a set of practice stimuli at the beginning of the test may be used.

2.4 Measures of accuracy

As previously mentioned, age estimation accuracy is highly task-dependent, with tasks ranging from distinguishing between two age groups of speakers to making direct age estimations in years. Depending on the task, different measures are used, and different accuracy rates are obtained. Table 2.2 shows some measures of accuracy used in previous studies of age perception.

Table 2.2: Measures of accuracy used in age perception studies (*PA*: perceived age, *CA*: chronological age)

Measure
Correlation between PA and CA
Percentage of correctly identified age groups
Percentage of correspondence (quotient of CA and PA, lower divided by higher)
Mean absolute error (absolute difference in years between mean PA and CA)

Correlation between perceived age (PA) and chronological age (CA) is one commonly used measure in perceptual studies (e.g. Shipp and Hollien, 1969; Braun, 1996). High correlations have been interpreted as highly accurate age estimations by listeners (Neiman and Applegate, 1990). However, a high correlation may just as well be the result of a constant error, e.g. an overestimation by 10 years (Braun, 1996). This should be taken into consideration when using correlation as a measure.

The percentage of correctly identified age groups is used when the task is to classify speakers into age groups (e.g. Ptacek and Sander, 1966; Jacques and Rastatter, 1990). This measure may be less reliable than others, since the percentage of correct answers depends on the number of age groups there are to choose from (Cerrato et al., 2000).

Neiman and Applegate (1990) and Cerrato et al. (2000) measured accuracy as the correspondence between PA and CA. The percentage of correspondence between CA and PA is calculated by dividing the lower of these two values by the higher one. This is a non-linear score, as it allows for much larger differences between PA and CA for older speakers than for younger ones.

Mean absolute error, calculated as the absolute difference between mean PA and CA, has been used in some recent studies (e.g. Braun, 1996; Braun and Cerrato, 1999; Nagao and Kewley-Port, 2005). Braun and Cerrato (1999) conclude that absolute error is the best way to establish listener accuracy in a realistic manner, both when used as an average over listeners per speaker and when averaged over different speakers for each listener.

Another measure of accuracy is the variation (or standard deviation) of the estimates made by different listeners. If only mean PA is taken into consideration, and if half the group underestimate a stimulus by 10 years

and the other half equally overestimate the same stimulus, the average PA value wil be very close to the CA of the stimulus. Also, if the majority of the listeners make very accurate judgements, a small number of listeners who completely misjudge the age of the stimulus may lower the overall accuracy value. This problem has been recognised, and solutions have been suggested where atypical speakers can be identified using the standard deviation (SD) for PA. Shipp and Hollien (1969) defined atypical speakers as having an SD of nine or more years in listening experiments, while Ryan and Burk (1974) selected typical speakers for an acoustic study from among those who had an SD for PA of six years or less. Horii and Ryan (1981) also considered standard deviations of up to six years to indicate close agreement among listeners. Minematsu et al. (2003) found that the SD for PA was larger for adult speakers than for children, as could be expected – i.e. "perceptual age has a sharper distribution for younger speakers and a broader for older speakers".

2.5 Previous related studies

Perceptual studies of speaker age began to emerge in the first half of the 20th century. Since then, researchers from various parts of the world have carried out numerous experiments with listeners judging the age of speakers of varying chronological age. Most studies so far have used healthy speakers and listeners with normal hearing.

This section contains a brief review of a selection of previous perceptual studies, in order to give the reader a sense of the large variation among the different investigations carried out. Table 2.3 summarises the speech data, subjects, precision demanded and findings of these experiments.

1. McGehee (1944) carried out a voice recognition test, where 30 judges (24 female, 6 male) estimated the age of five male speakers (aged 25–37) from read speech samples (56 words). Mean values were calculated for the judgements of each speaker, and the results were also ranked on the basis of perceived age. The youngest speaker (25 years) was judged to be the oldest (36 years). McGehee found a tendency for the listeners to base their estimations on pitch and rate. Since no measure of accuracy was given by McGehee, the author has calculated the mean error between PA and CA to be 4.2 years.

2. Ptacek and Sander (1966) asked ten speech pathology students to differentiate healthy young adult speakers (under age 35) from older ones (over age 65) based on three types of stimuli: (1) a sustained vowel (/ɑ/)

Table 2.3: Data and results from 14 perceptual studies of speaker age (*rev.*: reversed, *?*: no explicit information given in the reference article, *phon.*: phonated, *whisp*: whispered, *spont.*: spontaneous)

No	*Authors*	*Year*	*Type of speech*			*Speakers*		
			language	*type*	*dur. (s)*	*N*	*gender*	*age*
1	McGehee	1944	English	read	56 words	5	♂	25–37
2	Ptacek & Sander	1966		1 vowel	4	72	♀ ♂	18–87
			English	read	53 words	36	♀ ♂	19–87
				rev. read	53 words	36	♀ ♂	19–87
3	Shipp & Hollien	1969		read	≈ 8?[a]	175	♂	20–89
			English	read	≈ 8?[a]	175	♂	20–89
				read	≈ 8?[a]	175	♂	20–89
4	Huntley et al.	1987	English	read	≈ 8?[a]	105	♂	20–90
5	Linville	1987	English /æ/	phon.	?	75	♀	25–80
				whisp.	?			
6	Jacques & Rastatter	1990	English 10 vowels	phon.	5	40	♀ ♂	21–74[b]
				F_0 only				
				whisp.				
7	Braun	1996	German	read	≈ 45	40[c]	♂	25–58
			German	read	≈ 45	40[c]	♂	25–58
8	Decoster	1998	Dutch	[a]	2	32	♀ ♂	21–94
9	Cerrato et al.	1998	Italian	read[d]	≈ 60	42	♀ ♂	18–59?
			Italian	read[d]	≈ 60	42	♀ ♂	18–59?
10	Braun & Cerrato	1999	German	read[d]	≈ 45	20	♂	21–61
			German[e]	read[d]	≈ 45	20	♂	21–61
			Italian[e]	read[d]	≈ 45	20	♂	21–61
			Italian	read[d]	≈ 45	20	♂	21–61
11	Stölten & Engstrand	2002	Swedish	spontan.	≈ 30	12	♀ ♂	22–27+ 55–76
12	Stölten & Engstrand	2003	Swedish	spontan.	≈ 30	12	♀ ♂	22–27+ 55–76
				spontan.[f]				
13	Brückl & Sendlmeier	2003	German?	3 vowels	2.2	56	♀	20–87
				read	?			
				spont.	?			
14	Nagao & Kewley-Port	2005	English/ Japanese	vowel /i/	0.500	30/ 30	♀ ♂	24–86
				phrase	?			
				read	22/17 words			

[a] No duration is mentioned, but the author read the same 22-word sentence in about 8 seconds.

[b] Mean ages for each age group (older, younger).

[c] 20 smokers, 20 non-smokers.

[d] Telephone-transmitted speech.

[e] However, the listeners had no knowledge of this language.

[f] However, the stimuli had been speech rate and F_0 manipulated.

Data and results from 14 perceptual studies of speaker age (continued) (*?*: no explicit information given in the reference article, *%corct*: % correct, *corl.*: correlation, *μ.err.*: mean error, *%corsp.*: % correspondence)

Listeners				*Task*	*Results*			
N	*gender*	*age*	*nationality*		*%corct*	*corl.*	*μ.err.*	*%corsp.*
30	♀ ♂	?[b c d]	American	direct	-	-	4.2[a]	-
10	♀ ♂	?[b]	?		78%	-	-	-
				2 groups	99%			
					87%			
30	?	?[b]	?	3 groups	(PA directly related to CA)			
40	?	?[b]		7 groups	(PA rises directly with CA)			
25	?	?[b]		direct	-	0.88	-	-
120	?	9–84	?	direct	-	> 0.88	-	-
16	♀	elderly	American?	3 groups	45%	-	-	-
15		young			51%	-	-	-
16		elderly			38%	-	-	-
32	♀ ♂	20–70[b]	?	2 groups	50–80%	-	-	-
					48–77%	-	-	-
					38–59%	-	-	-
12[c]	♀ ♂	29–62	German?	direct	-	0.70	5.9	85.5%
19[d]	♀ ♂	20–32	German?	direct	-	0.68	6.5	84.1%
7[b]	♀	young	Belgian?	2 groups	-	> ?	-	-
4[c]	♂	?	Italian	7 groups	≈ 27%	-	-	≈ 85%
13[c]	♀ ♂	?[b]	Italian	7 groups	≈ 27%	-	-	≈ 85%
20	♀ ♂	?[b]	German	direct	-	.578	7.9	-
20	♀ ♂	?[b]	Italian	direct	-	.300	9.8	-
20	♀ ♂	?[b]	Italian	direct	-	.751	7.4	-
20	♀ ♂	?[b]	German	direct	-	.745	7.8	-
19[b]	♀ ♂	20–28	Swedish	4 groups	-	0.92	-	-
20[b]	♀ ♂	20–29	Swedish	4 groups	-	0.96	-	-
					-	0.86	-	-
15	♀ ♂	22–35	German?	direct	-	0.344–0.738	-	-
					-	0.862	-	-
					-	0.864	-	-
24/24	♀ ♂	young	English/Japanese	direct	-	0.38	≈ 22	-
					-	0.62	≈ 16	-
					-	0.76	≈ 14	-

[a] Calculated by the author.

[b] University or college students.

[c] Trained phoneticians or speech therapists.

[d] Naïve listeners.

produced by 72 female and male speakers (aged 18–34 and 67–87 years), (2) read speech (53 words read by 36 female and male speakers aged 19–23 and 67–87 years) and (3) the same read speech played backwards. It was found that the listeners were impressively accurate at differentiating young speakers from old ones. The mean percentages of correctly identified age groups were 78% for the vowel, 87% for the reading samples played backwards, and 99% for the reading samples played forwards.

3. Shipp and Hollien (1969) had three groups of young adult college students judge a 22-word read sentence produced by 175 male speakers (aged 20–89 years), divided evenly into seven age groups with a ten-year span (20–29, etc.). One group of 30 students were asked to judge whether a stimulus belonged to a young voice, an old one or one that was neither young nor old; a second group of 40 listeners were asked to classify the stimuli into the seven age groups; and a third group of 25 were asked to perform direct age estimations. For the the first two groups, the mean values of the perceived age groups were directly related to the chronological age decade. For the group who made direct age estimations, the correlation between PA and CA was 0.88. When comparing the direct age estimations with the judgments along the seven-point scale, Shipp and Hollien found a high level of agreement.

4. Huntley et al. (1987) studied the span between speaker and listener age as a factor in deciding the relationship between PA and CA. Listeners from four age groups (9–15, 20–30, 40–50 and 60–80 years) judged direct age from the third sentence of the *Rainbow Passage* read by 105 normal, healthy men (aged 20–90 years, 15 speakers in each decade group). Although younger speakers were overestimated and older speakers underestimated, a strong positive correlation between CA and PA was obtained ($> 0.88\%$). Young (20–30) and middle-aged (40–50) adults performed somewhat better than the other groups. Statistical testing revealed that both listener and speaker age group had a significant effect on mean perceived age.

5. Linville (1987) gave 15 young and 16 old women the task of identifying the age group of phonated and whispered vowel (/æ/) stimuli produced by 75 female speakers from three age groups (25–35, 45–55 and 70–80). The phonated vowels yielded 45% correct age classifications by the old women, and 51% by the young ones. For the whispered vowels, the old females made a correct identification in 38% of cases while the young females were correct in 43% of cases. The young women were more accurate than the old ones in their judgements, but both listener groups performed statistically significantly better than chance.

6. Jacques and Rastatter (1990) used phonated, filtered and whispered five-second stimuli from ten vowels, produced by 20 young and 20 old speakers of both genders. The listeners – 16 students and 16 older adults of both genders – were asked to categorise the speakers as under 25 or over 70. The whispered stimuli generated the lowest, and the phonated stimuli the highest, percentages of correct speaker judgements. Significant main effects for speaker age (but not speaker gender or vowel condition) as well as interactions for speaker age and stimulus type were found, but no significant differences between the phonated and the filtered stimuli.

7. Braun (1996) had one group of 12 trained forensic phoneticians and one group of 19 university students listen to about 45 seconds of read speech produced by 40 male speakers aged 25–58, and then make direct age estimations and indicate the confidence level of their estimates on a five-point scale (from 1 for very uncertain to 5 for absolutely certain). Results showed high correlations between CA and PA for both the expert group (0.70) and the naïve listeners (0.68). The only difference between the two groups was that only the expert group had close agreement in terms of standard deviations. The mean difference between PA and CA for all 40 speakers averaged over all listeners was found to be 5.9 years for the expert group and 6.5 years for the student group.

8. Decoster (1998) invited seven young female speech therapists to place speech samples consisting of two seconds of sustained vowels in one of two categories (old and young). Results showed almost perfect perception.

9. Cerrato et al. (1998) played 60-second samples of read speech from 21 female and 21 male speakers to 4 male experts and 13 naïve university students of both genders and asked them to estimate the age group (among seven groups: 18–24, 25–31, 32–38, 39–45, 46–52, 53–59) of each speaker. The overall percentage of correct answers for age was 27%, and the percentage of correspondence between CA and PA was 85%. Experts did not perform statistically significantly better than naïve listeners.

10. Braun and Cerrato (1999) investigated the accuracy of subjective age estimation of voices speaking one known and one unknown language using 20 native Italian and 20 native German speakers and the same number of Italian and German listeners. The stimuli sets were telephone-transmitted read texts of about 45 seconds' duration, and the task was to estimate speaker age to the exact year. Results showed that listeners did not benefit substantially from their knowledge of the respective language, as there was no statistically significant difference between tasks in the performance of the

various listener groups. Overall, listeners were not able to estimate speaker age from telephone-transmitted samples as accurately as had been suggested by previous research.

11 and 12. Stölten and Engstrand (2002, 2003) asked 9 male and 10 female students aged 20–28 years to estimate speaker age on a four-point scale (about 20, 40, 60 or 80 years) from 30 seconds of spontaneous monologue produced by 6 old and 6 young Swedish female and male speakers of the Arjeplog dialect. A strong correlation between perceived and chronological age (0.92) showed that listeners were well able to tell younger from older speakers.

In a second age estimation test, manipulated monologues (lower speaking rate and raised F_0 for all except younger women, for whom F_0 was lowered) were judged by 19 students using the same procedure. Results showed that the age judgements of the manipulated samples were about 10 years higher than those of the original samples, suggesting that both speaking rate and F_0 may be cues to perceived age. The correlation between perceived and chronological age was found to be 0.96 for the original samples and 0.86 for the manipulated samples.

13. Brückl and Sendlmeier (2003) analysed speech samples of read speech, sustained vowels and spontaneous speech from 56 female speakers aged 20–87, and then carried out a direct age estimation test using 15 adult listeners (6 females and 9 males, aged 22–35). They found the highest correlation between PA and CA for the more naturally produced spontaneous stimuli (0.864), perhaps because they contained the most age information (e.g. in the semantic content, choice of words and sociolect), followed by read speech (0.862) and vowels (0.330–0.738).

14. Nagao and Kewley-Port (2005) asked 24 English and 24 Japanese young adult listeners to judge direct age from (1) a sustained vowel (/i/), (2) the phrase *BCC* (/bisisi/) and (3) the second sentence of *The North Wind and the Sun*, spoken by 30 English and 30 Japanese young, middle-aged and old speakers. Sentences yielded the highest correlation between PA and CA (0.76), followed by phrases (0.62) and vowels (0.38). The mean error was also lowest for the sentences (< 15 years). The Japanese listeners were generally better judges of age than the English ones.

2.6 Study 1: F_0 and word duration in age perception

Numerous studies of speaker age have found that F_0 (including F_0 variation) is a very important phonetic cue, and that speech rate also seems important (Hollien, 1987; Jacques and Rastatter, 1990; Linville, 1987, 2001). However, less is known about how spectral or non-prosodic[1] features, e.g. formant frequencies, affect age perception. A pilot study (Schötz, 2003b) indicated that spectral cues may be important too. Moreover, some recent studies have either failed to find strong correlations between measurements of F_0 and age, or have suggested that other source characteristics, including shimmer and energy peaks in long-term average spectra (LTAS), may be important in age perception (Minematsu et al., 2002b; Linville, 2002; Brückl and Sendlmeier, 2003; Winkler et al., 2003). One reason for the recent contradictory findings may be differences in speech material, e.g. the gender of the speakers. Study 1 was an attempt to find out more about the relationship between prosodic and non-prosodic cues to both female and male speaker age. In this study, the term *duration* is defined as *word duration*, i.e. the actual phonetic duration of one stimulus type, namely the word *rasa*.

2.6.1 Purpose and aim

The primary purpose of this study was to investigate if perceptual cues found in F_0 and duration are more important than spectral (or non-prosodic) cues in human perception of speaker age. A second purpose was to compare the age perception of female speakers with that of male speakers, based on which perceptual cues are preferred for each gender. An additional purpose was to study whether acoustic features related to F_0 and duration (as measured in a single word) correlate with chronological and perceptual age. The aim was to obtain more information about the acoustic cues found in F_0, duration and non-prosodic features, and also to learn more about listener strategies in the perception of female and male speaker age.

2.6.2 Questions and hypotheses

The following two research questions were asked:

1. Are F_0 and duration more important to age perception than spectral (non-prosodic) cues?

[1]The two terms *spectral* and *non-prosodic* are used in this chapter as synonyms, and are defined as everything but F_0 and duration.

2. Are there differences between the perception of female and male speaker age as far as F_0 and duration are concerned?

It was hypothesised that F_0 and duration are less important than spectral cues, but that their importance may differ between the genders.

2.6.3 Speech material and preparations

The speech material encompassed 24 natural, 4 synthesised and 96 resynthesised versions of the isolated Swedish word *rasa* ['ʁɑ̀ːsa][2] ('collapse'). Six older women (age: 60–82), six younger women (age: 18–31), six older men (age: 60–76) and six younger men (age: 20–29) had produced the natural versions by elicitation. They were healthy speakers of the same dialect (Småländska) taken from the SweDia 2000 dialect corpus, which contains material from speakers belonging to an old as well as a young generation (Bruce et al., 1999). In order to facilitate analysis of inter-speaker variation, the natural stimuli were named using abbreviations representing village, age group, gender and speaker number, as exemplified in Table 2.4.

Table 2.4: Two examples of stimulus name abbreviations: *aow1* is the first old woman from village *a*; *sym3* is the third young man from village *s*

	Village (a/s)	*Age (o/y)*	*Gender (w/m)*	*Speaker number (1/2/3)*
example 1	a	o	w	1
example 2	s	y	m	3

Four synthesised versions with monotonous F_0 values were created using two young-speaker MBROLA-based concatenative synthesisers. Lucas (Filipsson and Bruce, 1997) generated two synthesised male stimuli (age: 30; F_0 values 80 & 110 Hz), and Ofelia (Svensson, 2001) generated two synthesised female stimuli (age: 24; F_0 values 160 & 220 Hz). The 4 synthesised and the 24 natural versions were then used to generate new PSOLA-resynthesised stimuli using a "prosody switching" script called *Mix-Prosody!* (developed by Johan Frid[3]) for the speech analysis software Praat (Boersma and Weenink, 2005). *Mix-Prosody!* takes two input sound files A and B. It first extracts the F_0 and duration from each sound file. Then it "mixes" (i.e. switches) the prosody of the two sound files and generates two new output sound files AB and BA, where AB has the non-prosodic features (i.e. everything but F_0 and duration) of A, but the F_0 and duration of B, and BA has the non-prosodic features of B but the F_0 and duration of A, as shown in Figure 2.2.

[2]The majority of the speakers spoke a South Swedish dialect variety where the velar or uvular [ʁ] is used. However, there were also some speakers from other dialectal areas where the alveolar [ɹ] is used instead.

[3]Department of Linguistics and Phonetics, Centre for Languages and Literature, Lund University.

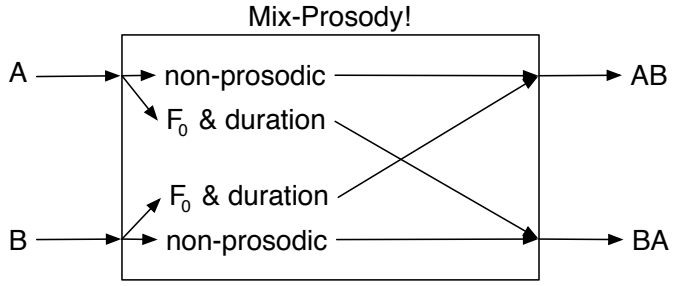


Figure 2.2: Schematic overview of the *Mix-Prosody!* script

2.6.4 Method

The method used included acoustic analysis of F_0 and duration as well as listening experiments.

Tools

The main tool used in this study was a Praat script, which automatically extracted word duration and F_0 features. *Praat* is a free, scriptable software tool for acoustic analysis, developed by Paul Boersma and David Weenink at the Department of Phonetics, University of Amsterdam. The tool is continuously being enhanced, and it currently comprises a considerable number of analysis functions and algorithms for "doing phonetics by computer" (Boersma and Weenink, 2005). A scripting function allows the development of new functions. Praat scripts can be used to automatise new as well as built-in functions.

Procedure

First, word duration and F_0 (mean, range and SD) were measured automatically for the natural and synthesised stimuli using the Praat script. Then, two listening tests – one with female, one with male stimuli – were created. Each test consisted of five different parts based on the various stimuli types used, as described in Table 2.5. Part 1 contained the six old natural and the two synthesised versions. This part was used to establish whether the listeners would perceive the old natural voices as older than the synthesised stimuli; knowing this would be important when interpreting the results of Parts 2–4. Part 2 included resynthesised "prosody-switched" stimuli from the synthesised versions and the six old speakers. In Part 3, one old speaker switched prosody with all young speakers, and in Part 4 one young speaker switched prosody with all old speakers. Part 5 comprised only natural and synthesised versions, and was used to obtain direct age estimations for all the speakers used in the tests.

Table 2.5: The five parts and stimulus types of the listening tests

Part	*Stimulus description*	*N*
1	6 old natural, 2 synthesised	12 pairs
2	resynthesised (from 6 old natural, 2 synthesised)	12 pairs
3	resynthesised (from 1 old, 6 young)	6 pairs
4	resynthesised (from 1 young, 6 old)	6 pairs
5	12 natural, 2 synthesised (presented twice)	28

All stimuli were normalised for intensity (see p. 10) and presented in random order at a constant and comfortable sound level using an Apple PowerBook G4 with Harman Kardon's SoundSticks loudspeakers. Pairs of stimuli were played to the listeners (students of phonetics at Lund University) in Parts 1 to 4. In Part 1, their task was to judge if the the old natural stimuli sounded older than the synthesised ones or not. In Parts 2 to 4, the listeners were asked to judge which member of each pair of resynthesised stimuli sounded older – the stimulus containing the F_0 and duration of an old speaker, or the stimulus containing the spectral features of an old speaker. In Part 5, the listeners were asked to make direct age estimations (in years) of the stimuli. In the test with female stimuli, 31 listeners (24 female and 7 male; age: 18–36, mean age: 21.6) took part; and 29 listeners (24 female and 5 male; age: 18–28, mean age: 22.3) participated in the male speaker test.

2.6.5 Results

The results from each part of the listening tests are described in this section, and also summarised in Tables 2.6 to 2.9.

Part 1

Part 1 comprised stimulus pairs of six old natural and two synthesised voices. For female speakers, 88% of the old women were found to sound older than the corresponding synthesised stimuli (Ofelia). When the speaker "responsible" for most of the errors (aow3) was excluded from the results, this percentage increased to 92%. In the 12% of the cases where Ofelia was judged to sound older, there was no large difference between the high F_0 version (5%) and the low F_0 version (7%). In the male test, the listeners judged 87% of the natural old stimuli to sound older than the two synthesised stimuli (Lucas). Out of the 13% of cases where Lucas was judged to sound older, the low F_0 version accounted for 9% and the high F_0 version for 4%. Two old natural speakers (aom2 and aom3) were judged to be younger than Lucas's low F_0 version, but older than "his" high F_0 version.

Table 2.6: Results for the female and male tests, Part 1 (natural and synthesised stimuli). Numbers and percentages for the listeners' judgements as to which sounded *older*: *old* natural or *young* synthesised stimuli

Part	*Test*	*Stimulus pairs*	*Total*	*Old natural*		*Young synthesised*	
		(natural, synthesised)	*N*	*N*	*%*	*N*	*%*
1	female	old natural, Ofelia	372	329	88	43	12
	male	old natural, Lucas	348	304	87	44	13

Part 2

In Part 2, which comprised resynthesised stimuli from natural old and synthesised versions, the listeners judged 69% of the female stimuli with old spectral (non-prosodic) features to sound older than the stimuli with old F_0 and duration. For the male speakers, old spectral features were judged to sound older than old F_0 and duration in 93% of cases.

Part 3

Listeners perceived 66% of the old spectral features as older than old F_0 and duration for the female speakers in Part 3, which contained resynthesised stimuli from one old and six young natural speakers. For the male speakers, three of the stimulus pairs (aom2 "mixed" with aym1, sym1 and sym2) yielded results close to 50%. As a consequence, the overall result for this part is down to 69%.

Part 4

Part 4, which contained resynthesised stimuli from one young and six old speakers, yielded the lowest result for the female speakers. Only 57% of the stimuli with older spectral features were judged to be older. Somewhat higher results were found for the male speakers as compared with Part 3: 80% of the stimuli consisting of older spectral features in combination with younger duration and F_0 were judged to be older.

Summary of Parts 1 to 4

In Part 1, the old natural speakers were much more often judged to sound older than the synthesised versions. The results for Parts 2 to 4 revealed that the resynthesised stimuli with spectral features of old speakers combined with F_0 and duration of young speakers were more often judged to sound older than the opposite combination (i.e. young spectral features and old F_0 and duration). Comparison of the results shows that the differences were

Table 2.7: Results for the female and male tests, Parts 2, 3 and 4 (resynthesised stimuli). Numbers and percentages of the listeners' judgements as to which sounded *older – old* spectral quality or *old* F_0 and duration

Part	*Test*	*Stimulus pairs (resynthesised)*	*Total*	*Spectral quality*		*F_0 and duration*	
			N	*N*	*%*	*N*	*%*
2	female	old + Ofelia (mix)	372	258	69	114	31
	male	old + Lucas (mix)	348	322	93	25	7
3	female	one old + young (mix)	186	121	66	62	34
	male	one old + young (mix)	174	120	69	54	31
4	female	old + one young (mix)	186	104	57	77	43
	male	old + one young (mix)	174	139	80	35	20

statistically significant for both female and male speakers, though slightly better for the male speakers, except for Part 1 (see Table 2.8).

Table 2.8: χ^2 results for Parts 1 to 4 of the female and male tests

Test	*Part 1*		*Part 2*		*Part 3*		*Part 4*	
	$\chi^2(1)$	$p <$	$\chi^2(1)$	$p <$	$\chi^2(1)$	$p <$	$\chi^2(1)$	$p <$
female	219.882	0.001	55.742	0.001	19.022	0.001	4.028	0.045
male	194.253	0.001	254.205	0.001	25.034	0.001	62.161	0.001

Part 5

As regards the female speakers, the estimation accuracy of direct age for the old women ranged from nearly perfect (aow1, sow1) to fairly good (aow2, sow2, sow3). The mean absolute difference between PA and CA (mean absolute error) for all natural female speakers was 7.2 years. As was expected from the results of the previous parts, one atypical speaker (aow3) was judged to be much younger than her CA. The two synthesised voices (named *hofe* (Ofelia with high F_0) and *lofe* (Ofelia with low F_0)), which are found in the middle of Figure 2.3, were judged to be much older than their "CA", as were two of the young voices. However, the other young voices showed results very close to the CA of the speakers. Correlation of the mean PA with CA shows that the result was statistically significant for the old women ($r = 0.825$; $p < 0.05$), but not for the young ones ($r = 0.097$; ns).

The accuracy of the direct age estimates for the male speakers ranged from almost perfect (som3) to underestimations by about 10–15 years (aom2, aom3, som1) for the old men (to the left in Figure 2.4). The mean absolute error for all natural male speakers was 10.6 years. Many of the young speakers (right-hand side of Figure 2.4) were perceived as much older than their CA, but some estimations came very close (aym2, sym2). The two synthesised versions (named *hluc* (Lucas with high F_0) and *lluc* (Lucas with

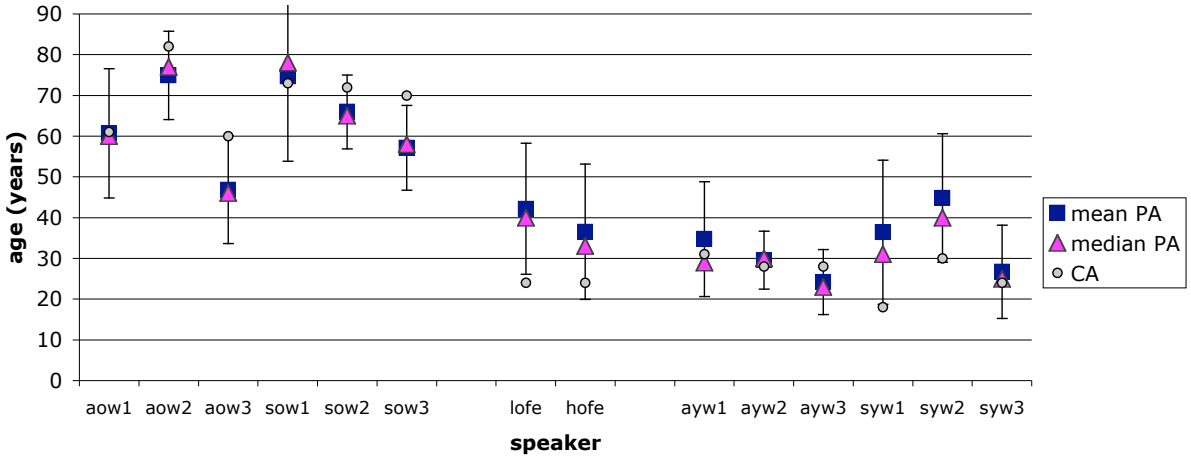


Figure 2.3: Chronological age (CA) and mean, median and standard deviation of perceived age (PA) for each female speaker stimulus in Part 5

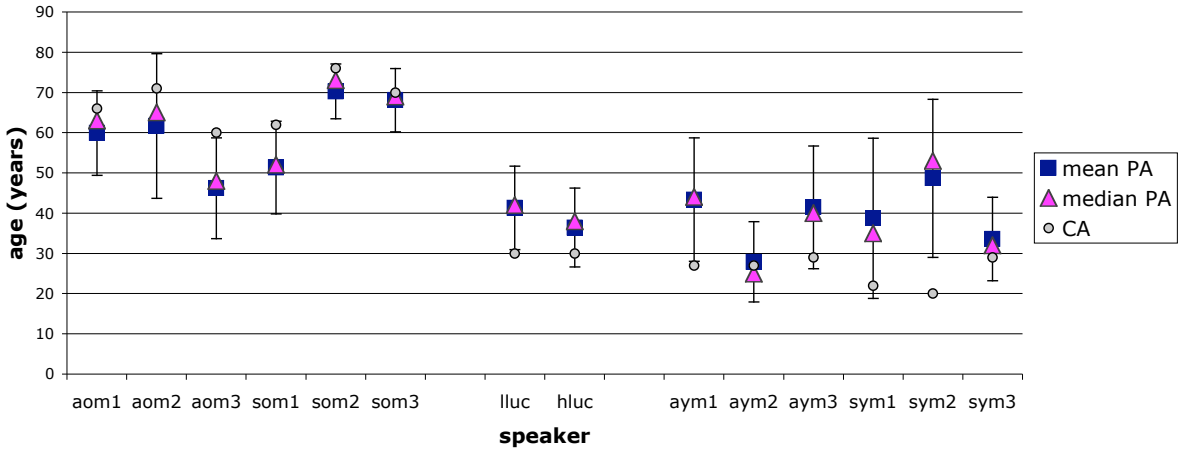


Figure 2.4: Chronological age (CA) and mean, median and standard deviation of perceived age (PA) for each male speaker stimulus in Part 5

low F_0)), found in the middle of Figure 2.4, were both overestimated, and the low F_0 version was perceived as older than the high F_0 version. The correlation of mean PA and CA was statistically significant for the old men ($r = 0.944$; $p < 0.05$), but not for the young ones ($r = 0.522$; ns).

To sum up, the listeners did make fairly good age judgements. Nevertheless, younger speakers were frequently overestimated, while older speakers were underestimated. No statistically significant difference between the listeners' ability to judge the age of female and male speakers was found ($t(30) = 0.3$; $p > 0.05$). Comparison of the results for female and male speakers shows that the male ones (old and young) yielded a somewhat better correlation between PA and CA (see Table 2.9 and Figure 2.5). Statistically significant results were found only for the older speakers of both genders, but the younger speaker results were slightly better for the males.

Table 2.9: Correlations between mean perceived age and chronological age of young and old female and male speaker age (statistically significant values in boldface)

	Old		*Young*	
	r	$p <$	r	$p <$
female	**0.825**	0.05	0.097	ns
male	**0.944**	0.05	0.362	ns

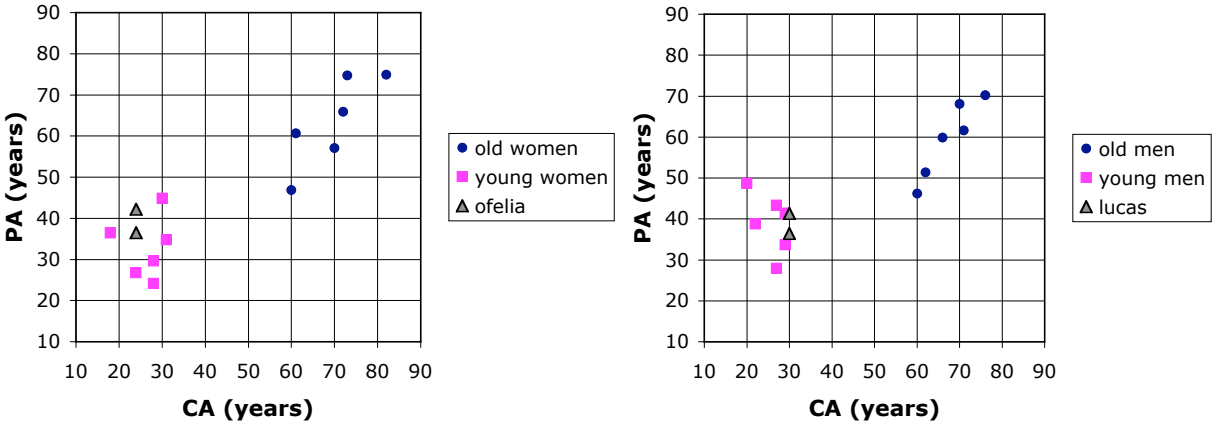


Figure 2.5: Scatter plots of perceived age (PA) with chronological age (CA) in Part 5 of the female (left) and male (right) tests

Duration and F_0

Word duration correlated with chronological age (CA) and mean perceived age (PA) for both male and female speakers, but it correlated with the standard deviation for PA (PA SD) only for female speakers (see Table 2.10).

Table 2.10: Correlation of duration with chronological age (CA), mean perceived age (PA) and the standard deviation for PA (PA SD) for female and male speakers (statistically significant r values in boldface)

Duration vs.	*CA*		*Mean PA*		*PA SD*	
	r	$p <$	r	$p <$	r	$p <$
female	**0.431**	0.001	**0.525**	0.001	**0.394**	0.005
male	**0.415**	0.005	**0.522**	0.001	−0.220	ns

The acoustic values for mean F_0, F_0 range and F_0 SD were also correlated with CA, mean PA and PA SD. Statistically significant results were obtained only for female speakers. Although mean F_0 and F_0 SD in semitones (ST) correlated neither with CA nor with mean PA for the female speakers, both F_0 range and F_0 SD in Hz did. The strongest correlation was found between F_0 SD in Hz and female CA, as shown in Table 2.11.

Table 2.11: Correlation of mean F_0, F_0 range and F_0 SD with CA, mean PA and the standard deviation for PA (PA SD) for female (A) and male (B) speakers (statistically significant r values in boldface)

A	*CA*		*Mean PA*		*PA SD*	
Female	r	$p <$	r	$p <$	r	$p <$
Mean F_0	0.179	ns	0.172	ns	0.140	ns
F_0 range	**0.442**	0.005	**0.378**	0.005	−0.052	ns
F_0 SD (Hz)	**0.465**	0.001	**0.412**	0.005	−0.027	ns
F_0 SD (ST)	0.284	ns	0.140	ns	−0.277	ns

B	*CA*		*Mean PA*		*PA SD*	
Male	r	$p <$	r	$p <$	r	$p <$
Mean F0	0.020	ns	0.128	ns	0.142	ns
F_0 range	−0.113	ns	−0.239	ns	0.032	ns
F_0 SD (Hz)	0.166	ns	0.197	ns	0.205	ns
F_0 SD (ST)	0.278	ns	0.310	ns	−0.218	ns

2.6.6 Discussion and conclusions

This study was mainly perceptual, although acoustic measures of F_0 and duration were analysed in order to shed further light on how these two acoustic features affect the perception of female and male speaker age.

F_0 and duration vs. spectral (non-prosodic) cues

The results from Parts 2 to 4 in the perceptual part of the study clearly suggested that duration and F_0 may not be as important as non-prosodic cues in the perception of speaker age, at least not when the speech samples are single words. For male speakers, this experiment confirmed the results of the pilot study Schötz, 2003b. It also showed that the same relationship exists for female speakers. The hypothesis of F_0 and duration being less important than spectral cues was thus confirmed. However, in the acoustic part of the study, higher correlations of word duration with both male and female CA and PA were found than with the standard deviation of mean perceived age. This may indicate that duration is an important age cue, but that inter-listener variation cannot be explained in terms of word duration. As for F_0, it seems that mean F_0 may be a less important age cue than F_0 range and F_0 SD, judging from the results displayed in Table 2.11. The relationship between duration and F_0 has to be studied further using a larger and more varied speech material in order for any firm conclusions to be drawn.

Perception of female and male speaker age

Several interesting results were found regarding different listener strategies used when judging female and male speaker age. The listeners were equally good at estimating female and male age. Speech rate (duration) may be equally important as a cue to the age of females and males, as indicated by the statistically significant correlations between duration and female and male PA. However, gender differences were found in F_0. Statistically significant correlations were obtained only for female age with CA and PA. This might suggest that F_0 is more important in the perception of female than of male age.

Influencing factors

Numerous other factors may, however, have influenced the results of this study. In Part 5, some speakers were judged to have a PA very different from their CA. It is possible that these "atypical" speakers affected the results of Parts 1 to 4. Moreover, old speakers were often underestimated, while young ones were overestimated. This may be explained by a general reluctance to make extreme age judgements. Further, the range of CA was wider for the old speakers than for the young ones, so the above direct comparison between the groups as to correlation is likely to be a bit misleading.

Several studies have reported that female F_0 contains more variation than male F_0 (Klatt and Klatt, 1990). However, since F_0 in semitones (ST) seems to be closer to perception than F_0 in Hz (Traunmüller and Eriksson, 1995), it is possible that the importance of F_0 to age perception is as small for female speakers as for male speakers. In this study, only female (but not male) F_0 SD in Hz (but not in ST) correlated with age. This indicates that the distinction between acoustics and perception also involves differences between female and male voices. Not every acoustic variation is perceived. Moreover, in the measurement of F_0 range in the acoustic part of the study, creaky segments were not excluded. Speech with abrupt frequency halvings is more often perceived as having a coarser voice quality than as being lower in pitch (Traunmüller and Eriksson, 2000). This could have affected the listeners' age estimates of male speakers and may be one of the reasons why the results for male speakers were not statistically significant.

Conclusions

The following conclusions may be drawn from the study:

1. Spectral features seem more important in age perception than F_0 and duration. Duration may still be important, but the role of F_0 remains

unclear, although it is probably minor.

2. While word duration seems to be equally important in the perception of female and male speaker age, F_0 seems to be more important in the age perception of female speakers than of male ones.

2.7 Study 2: Effects of stimulus type and duration

This study addressed the question of how much and what kind of speech information we need to make as good estimates of speaker age as possible, and whether listeners use different strategies when estimating female and male speaker age. Recent studies have begun to find differences between female and male speakers in both human and machine age recognition (Müller et al., 2003; Schötz, 2006b). While F_0 seems to be a better cue to female age, formants seem to constitute better cues for judging male age. One possible explanation is that the characteristics of female voices appear to be perceived as more complex than those of male speech (Murry and Singh, 1980), suggesting that listeners would need either a partly different set or a larger number of phonetic cues when judging female age. In this study, the term *duration* is used to denote *stimulus duration*, i.e. the time (in seconds) during which listeners are exposed to a stimulus.

2.7.1 Purpose and aim

The purpose of the present study was to determine to what extent two different stimulus types (isolated words and spontaneous speech) and stimulus duration (exposure time) influence the estimation of female and male speaker age. The aim was to shed further light on how much information in speech is needed in order to make accurate estimations of speaker age.

2.7.2 Questions and hypotheses

In this study, the following research questions were posed:

1. In what way does stimulus duration and type affect the accuracy of listeners' perception of speaker age?

2. Is there a difference in the perception of female and male speaker age with respect to stimulus duration and type?

It was hypothesised that longer stimulus durations and stimuli containing spontaneous speech would lead to better estimation accuracy. A further hypothesis suggested that there would be differences between female and male speakers.

2.7.3 Speech material and preparations

Figure 2.6: The southern part of Sweden (Götaland – below the line) from where the speakers were selected

Six speakers from each of four different groups – older women (aged 63–82), older men (aged 60–75), younger women (aged 24–32) and younger men (aged 21–30), all from the southern part of Sweden (see Figure 2.6) – were selected semi-randomly from the SweDia 2000 database (Bruce et al., 1999), which contains elicited isolated words and spontaneous narratives of non-pathological native speakers of Swedish, recorded in their homes. For each of the 24 speakers, four different speech samples were extracted, normalised for intensity and used in the following listening tests:

- Test 1: about 10 seconds of spontaneous speech
- Test 2: about 3 seconds of spontaneous speech
- Test 3: a concatenation of six isolated words: *käke* [ˈɕɛ̀ːkə] ('jaw'), *saker* [ˈsɑ̀ːkəʁ][4] ('things'), *själen* [ˈɧɛːlən] ('the soul'), *sot* [suːt] ('soot'), *typ* [tyːp] ('type' [noun]) and *tack* [tak] ('thanks') – duration ≈ 4 seconds
- Test 4: an isolated word: *rasa* [ˈʁɑ̀ːsa] ('collapse') – duration ≈ 0.65 seconds

[4]The majority of the speakers spoke a south Swedish dialect variety where the velar or uvular [ʁ] is used. However, there were also some speakers from other dialectal areas where the alveolar [ɹ] is used instead.

2.7.4 Method

Four separate perception tests – one for each of the four sets of stimuli – were carried out. Two listener groups participated in one test each, while a third group took part in two of the tests. The gender and age distributions of the three groups are shown in Table 2.12, along with information on which test(s) and set(s) of stimuli each group was presented with. All subjects were students of phonetics at Lund University. Their task was to make direct age estimations based on first impressions of the 24 stimuli, which were presented free-field at a constant and comfortable volume using an Apple PowerBook G4 with Harman Kardon's SoundSticks loudspeakers. In all four tests, the stimuli were played only once in the same random order. The listeners were also asked to indicate the cues which they believed had affected their judgements.

Table 2.12: Test number, stimulus set, number of listeners, and gender and age distributions of the listener groups in the four tests

Test (stimuli)	*Total (N)*	*Female*	*Male*	*Age range (mean/median)*
1 (10 s)	29	18	11	19–65 (27/22)
2 (3 s)	33	22	11	19–57 (25/23)
3 (6 words)	37	33	4	19–55 (26/24)
4 (1 word)	37	33	4	19–55 (26/24)

2.7.5 Results

Accuracy

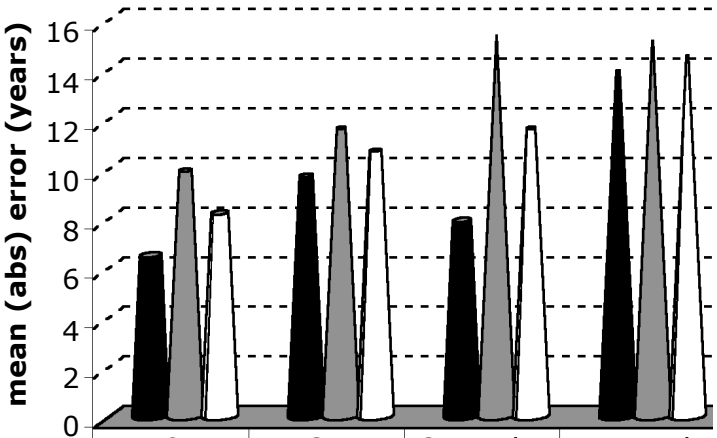
Figure 2.7 displays the mean absolute error, i.e. the average of the absolute difference between perceived age (PA) and chronological age (CA) in years, for female, male and all speakers in the four tests, while Figure 2.8 shows the corresponding correlations between CA and PA. Since the graphs of the two measures display similar results (a lower mean error in Figure 2.7 corresponds to a higher correlation in Figure 2.8), and since it has been suggested that mean absolute error is a more realistic measure of listener accuracy (Cerrato et al., 2000), only this measure will be mentioned and discussed below.

In the first three tests, the listeners' judgements were about twice as accurate as the ones made by a simple "baseline" estimator, which judged all stimuli to be 47.5 years old (the mean CA of all speakers). In Test 4, the shortest (1 word) stimuli yielded results at levels approximately half-way between the baseline and the results of the other tests. The sum, mean and median values of the errors for all speakers in the four tests as well as for the baseline are shown in Table 2.13.

In all four tests, the listeners' judgements of women were more accurate than those of men. The highest accuracy was obtained for the female speaker ten-second stimuli (6.5), while the male speaker six-word stimuli had the lowest accuracy (15.3). The listeners also tended to overestimate the younger speakers and to underestimate the older ones.

Table 2.13: Sum, mean and median values of the mean absolute error for all speakers in the four tests as well as the baseline (calculated as always guessing 47.5 years, the mean age of all speakers)

Test	*1 (10 s)*	*2 (3 s)*	*3 (6 words)*	*4 (1 word)*	*Baseline*
sum	196.5	256.1	277.6	348.7	497.0
mean	8.2	10.7	11.6	14.5	20.7
median	7.2	10.0	10.0	16.7	19.5



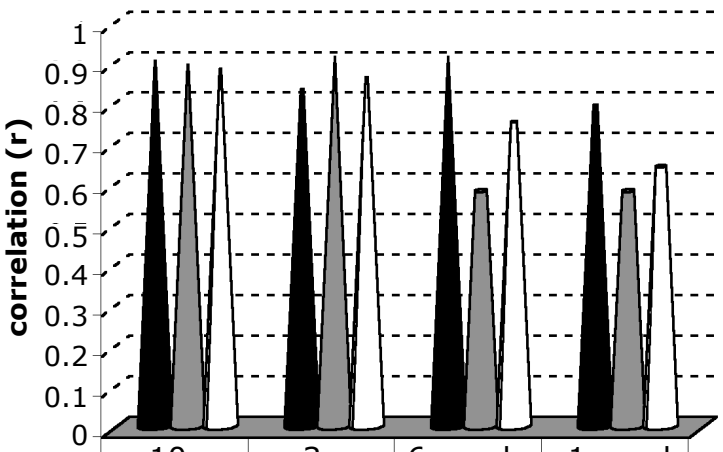
	10 s	3 s	6 words	1 word
Female	6.5	9.7	7.9	13.9
Male	9.9	11.6	15.3	15.1
All	8.2	10.7	11.6	14.5

Figure 2.7: Mean absolute error for the four sets of stimuli for female, male and all speakers

Stimulus and speaker gender effects

The listeners' mean absolute errors were subjected to two separate analyses of variance. In the first analysis, speaker gender and speaker age – old or young – were within-subject factors, and stimulus duration – short (1 word), medium (6 words and 3 s) or long (10 s) – was the between-subjects factor. In the second analysis, the between-subjects factor was stimulus type – spontaneous or word stimuli – instead of stimulus duration.

Longer stimulus durations led to higher accuracy, an effect which was significant ($F(2,100) = 71.059$, $p < 0.05$). A difference between the judgements



	10 s	3 s	6 words	1 word
Female	0.90	0.83	0.91	0.79
Male	0.89	0.91	0.58	0.58
All	0.88	0.86	0.75	0.64

Figure 2.8: Correlations between perceived age (PA) and chronological age (CA) for the four sets of stimuli for female, male and all speakers

for female and male speakers was also observed. Accuracy for increasingly longer stimuli improved more for the female than for the male speakers. For the female speakers, a lower mean absolute error was observed for the ten-second stimuli (6.5) than for the three-second stimuli (9.7), and the error for the six-word stimuli (7.9) was lower than for the one-word stimuli (13.9). The difference between longer and shorter stimulus durations was much smaller for the male speakers, with a mean absolute error of 9.9 for the longest (10 second) stimuli, higher errors for the medium-length three-second and six-word stimuli (11.6 and 15.3) and a similar error for the one-word stimuli (15.1). The interaction of speaker gender and stimulus duration was, however, not significant ($F(2,100) = 2.171$, ns).

In the second analysis, the interaction of speaker gender and stimulus type was significant ($F(1,68) = 39.296$, $p < 0.05$). For the male speakers, the listeners' judgments were more accurate for the spontaneous stimuli than for the word stimuli. Lower mean absolute errors were obtained for the two sets of spontaneous stimuli (9.9 and 11.6) than for the two sets of word stimuli (15.3 and 15.1). This effect, however, was not observed for the female speakers. For them, the mean absolute error for the six-word stimuli (7.9) was lower than that for the three-second spontaneous stimuli (9.7), but slightly higher than that for the longer spontaneous stimuli (6.5).

Listeners' cues

Most of the listeners indicated several cues which they believed had influenced their age judgements. The cues mentioned by the most speakers are

presented in Table 2.14. Dialect, pitch and voice quality affected the listeners' estimates in all four tests, while semantic content influenced their judgements in the tests with spontaneous stimuli. A common listener's remark in the tests with spontaneous stimuli concerned speakers talking about the past. Such speakers were often judged as being old, regardless of other cues. Additional listeners' cues included speech rate, choice of words or phrases and experience or familiarity with similar speakers (age group, dialect, etc.).

Table 2.14: Cues said by the listeners to have affected their judgements (absolute numbers of listeners)

Test	*Dialect*	*Pitch*	*Voice quality*	*Semantic content*
10 seconds	24	23	25	22
3 seconds	25	26	32	17
1 & 6 word(s)	33	31	34	0

2.7.6 Discussion and conclusions

Although only a limited number of stimulus durations and types were investigated in this study, a few interesting results were found. These are discussed below, along with a few suggestions for future work.

Accuracy

The listeners were significantly better (about twice as good) judges of age than the baseline estimator in three of the tests. This is in line with previous work. However, it remains unclear what accuracy levels can be expected from listeners' judgements of age. When dealing with the accuracy of perceived age, differences in speakers' chronological age have to be taken into account as well. A mean absolute error of 10 years could be considered less accurate for a 30-year-old speaker (a PA of 20 could be regarded as $20/30 = 66.7\%$ correct) than for an 80-year-old speaker (a PA of 70 could be regarded as $70/80 = 87.5\%$ correct). Obviously, there is a need for a better measure of accuracy for age estimation tasks. The fact that three different listener groups participated in the tests may also have influenced accuracy, as no check was made of the degree of consensus among listener groups.

In all four tests, the listeners' judgements of women were more accurate than those of men, perhaps because the listeners were mainly women themselves. However, the influence of listener gender on performance in age estimation tasks is still unclear and has to be further investigated. Although most researchers have not reported any difference in performance between male and female listeners, some studies have found females to perform better than males, while others have found male listeners to perform somewhat

better (Braun and Cerrato, 1999). Another explanation could be that the male speaker group contained a larger number of atypical speakers and consequently would be more difficult to judge than the female speaker group. Shipp and Hollien (1969) found that speakers whose age was difficult to estimate had standard deviations of nine years or more. Perhaps such a measure can be used to decide whether speakers are typical representatives of their CAs or not.

Stimulus effects

In this study, longer durations for the most part yielded higher accuracy for the listeners' age estimates. This raises the question of optimal durations for age estimation tasks. When does a further increase in duration for a specific speech or stimulus type no longer result in higher accuracy? Further studies with a larger and more systematic variation of stimulus duration for each stimulus type are needed to answer this question.

Significant effects for both accuracy and speaker gender differences were found for the two stimulus types compared in this study. However, elicited isolated words and spontaneous speech can be difficult to compare in a study of speaker age. Several listeners mentioned that they were strongly influenced in their judgements by the linguistic (semantic) content of the spontaneous stimuli, which may explain why the male speaker spontaneous stimuli yielded higher accuracy than the word stimuli. Besides providing more information about the speaker (dialect, choice of words, etc.), spontaneous speech is also likely to contain more prosodic and spectral variation than isolated words. However, for the female speakers, the lower accuracy obtained for the three-second spontaneous stimuli than for the – only slightly longer – six-word stimuli cannot be explained by stimulus type effects alone. It would be interesting to compare a larger number of speech types in order to find the types best suited for estimating the ages of both female and male speakers. Future work should include studies where several different speech types are compared and varied more systematically with respect to phonetic content and quality as well as phonetic variation and dynamics.

Speaker gender effects

As already mentioned in the previous paragraph, there were differences between female and male speakers with respect to which stimulus type and durations yielded higher age estimation accuracy. One explanation for the differences between female and male speakers may be that listeners use different strategies when judging female and male speaker age. As suggested in Schötz (2006b), it is possible that listeners use more prosodic cues (mainly

F_0) when judging a female speaker's age, but that spectral cues (i.e. formants, spectral balance, etc.) are preferred when judging a male speaker's age. Consequently, the results from this study may indicate the importance of the fact that for male speakers, spontaneous stimuli provide listeners with more spectral information, while longer stimuli contain more prosodic information needed to estimate female speaker age more accurately. The differences in the perception of female and male speaker age have to be studied further, and speaker gender has to be taken into consideration in future research investigating acoustic correlates of and perceptual cues to speaker age.

Conclusions

Although more research is needed to verify and further explore the results of this study, two tentative conclusions are drawn:

1. Longer stimulus durations and spontaneous speech samples (which contain more speaker-specific information) seem to improve the accuracy of listeners' perception of speaker age.

2. There are differences between perception of female and male speaker age, possibly explained by differences in listeners' strategies and choice of perceptual cues.

2.8 Summary

From a large number of previous studies concerning perception of speaker age, we have learned that (1) human listeners are fairly good at estimating the age of an unknown (and unseen) speaker, (2) perceptual cues to speaker age include variation in pitch, speech rate, voice quality, articulation and phrasing, and (3) several non-phonetic factors influence listeners' judgements.

The two studies presented here provide answers to some of the questions about human age perception not yet answered by the large amount of research carried out in this field.

The first study addressed the role of F_0 and duration in age perception. A series of listening tests were conducted and acoustic measures of F_0 and duration were obtained automatically. The results of the listening tests indicate that F_0 and duration may be less important than spectral or non-prosodic cues (defined as everything but F_0 and duration), and that although duration seems equally important for both genders, F_0 seems somewhat more important for the age perception of female speakers than of male ones.

The second experiment investigated age perception based on different stimulus types and lengths using another series of listening tests. From the results we may assume that it is likely that listeners use different acoustic cues and listening strategies when estimating the age of female and male speakers. Stimulus duration (i.e. longer speech samples, regardless of speech type) seems to be important when judging female speakers, while stimulus type (i.e. spontaneous speech, regardless of duration) seems to be more important in the case of male speakers.

Chapter 3

Machine perception of speaker age

3.1 Introduction

Machine perception concerns the building of software or machines which can sense and interpret their environment. This chapter addresses one such aspect, namely automatic recognition of speaker age. Why would anyone want to build a machine that can judge a speaker's age? Could you not just ask people how old they are? There are a few situations where this would be difficult. In forensic speaker identification, when recorded speech (e.g. from an answering machine) of an unknown perpetrator is all that is available, an automatic age estimator could judge his or her age so that suspects of other ages could be eliminated. Moreover, some people might be offended when asked how old they are. Not even computers should ask a lady her age.

Automatic recognition of age could improve human–machine communication. If user age could be identified automatically, spoken dialogue systems could adapt their communication behaviour to the age of the user. For instance, the system could use more youthful language when talking to a teenager. It could also suggest age-adapted information, such as tourist attractions or directions. As the number of children and elderly people who use computers in their daily lives increases, age-adapted speech recognition is becoming more important. Still, research on automatic age perception is relatively scarce (Shafran et al., 2003). One explanation is that it certainly is not an easy task. Age cues are present in every phonetic dimension, and they are hard to separate from other speaker variation characteristics, such as physiological condition and dialect.

This chapter begins with an overview of automatic speaker recognition and previous related research. It then describes two studies concerning automatic age estimation, and ends with a general discussion of the two studies.

3.2 Automatic speaker recognition (ASR)

Unlike automatic *speech* recognition, automatic *speaker* recognition (ASR) does not aim to identify the phonetic and linguistic content of speech samples. Instead, this field of research is about distinguishing speaker-specific characteristics of speech in order to recognise who is speaking. However, similar features and methods are used in both fields. The basic idea is to have a model of each target (phoneme, word, speaker), and to compare those models with the patterns found in an unknown speech sample. Two major tasks in ASR are speaker verification and speaker identification. In speaker verification, the task is to accept or reject the identity claim of a speaker by comparing the unknown speech sample against a database of enrolled users. The task in speaker identification is to identify one speaker among a group of known (enrolled) users. Identification performance deteriorates when the number of speakers increases.

Recognisers usually consist of three parts: signal pre-processing, feature extraction and pattern matching. They are described in the following sections.

3.2.1 Signal pre-processing

The first step in ASR is usually sampling the speech signal to a computer and processing it in some way. The sampling frequency and resolution of the signal are often optimised to facilitate the next step, which is feature extraction. Other preparations may include segmentation of the speech sample into linguistic or phonetic units (utterances, words, phonemes).

3.2.2 Feature extraction

A large variety of acoustic features have been used in ASR. In this section, only a few common features are described. Vectors consisting of spectral features extracted once every time frame of about 5–20 ms are often used. They encompass both anatomical (size of the speech organs) and learned (dialect, speaking style) features.

Spectra and spectrograms are well-known records of speech analysis. They are computed using Fourier transform (FT) or fast Fourier transform (FFT, a faster method for calculating the FT). FT decomposes a complex signal such as speech into its individual components (different frequencies and their amplitudes) for each time frame (Smith, 1999). Spectra and spectrograms are used to extract features such as F_0 and formant frequencies.

Very common features are mel frequency cepstral coefficients (MFCC). They are derived from cepstra, where the spectra computed from the FFT are

first transformed using the mel scale. A cepstrum is calculated using the FFT of the log of the FFT. Normally 8–16 coefficients are computed. Sometimes, delta coefficients are calculated as well, i.e. the first-order (ΔMFCC) and second-order time derivatives of the MFCCs (Bogert et al., 1963).

Other common features are linear prediction coefficients (LPC) derived from linear prediction analysis. This method separates a sampled waveform of speech into its source and filter components by inverse filtering, and predicts the current value of the waveform using a linear combination of earlier values (Markel and Gray, 1976).

3.2.3 Pattern matching

Methods used for pattern recognition in ASR are mainly knowledge-based or statistical. The knowledge-based methods use acoustic-phonetic and linguistic knowledge about speech and speaker variation to identify a speaker. Statistical approaches use models derived from training with known speech data to calculate the probability of correspondence between an unknown speech sample and the models (Ainsworth, 1997).

In knowledge-based systems, the recognisers rely on rules based upon expert knowledge about the variation of features such as F_0 and formant frequencies to carry out pattern matching. They are considered less successful than statistical methods (Ainsworth, 1997) because they do not model the statistical differences in the data and are less able to handle speech (and speaker) variation.

Two of the most widely used statistical methods are hidden Markov models (HMM) and artificial neural networks (ANN). An HMM consists of a number of states, which may represent a phoneme or a word. Transition probabilities determine the run from state to state through the model (Rabiner, 1989). Figure 3.1 shows an example of a three-state HMM.

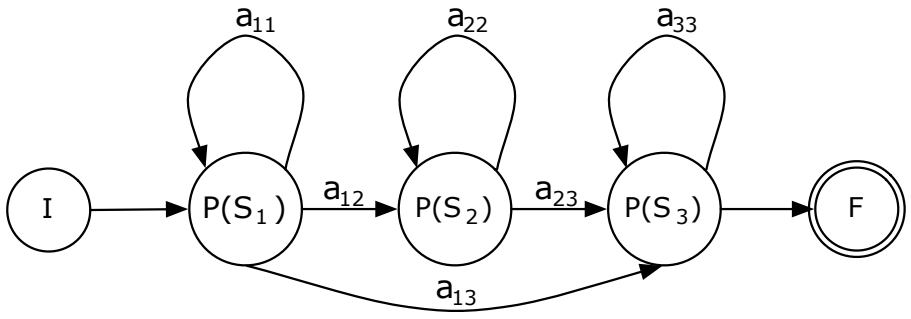


Figure 3.1: An example of a three-state hidden Markov model with initial (I) and final (F) states, observation probabilities $P(S_i)$ and transition probabilities a_{ij}

Every state in the HMM also carries observation probabilities for the features observed in that state, which are often modelled with Gaussian mixture models (GMM). GMMs are mixtures of Gaussian (normal) distributions with means, variances and weights (Jelinek, 1976). State duration is modelled as the number of time frames spent in the current state before a move to another one. HMMs are described by reference to their number of states and their initial probabilities, observation probabilities and transition probabilities. In the training of an HMM, the probabilities are estimated and optimised for the training data. Greater speaker variation requires more training material. In pattern matching, the HMM models the sequence of states and the corresponding model sequence which would be the most likely to generate the input speech sample.

An ANN is a large collection of interconnected simple processing units (neurons or nodes), organised in several layers. The processing ability of the network is stored in the weights of the inter-node connections. Training the ANN usually involves adjusting the weights to minimise the difference between the output of the ANN and the true target values (Callan, 1999). This method is often used to discriminate between classes, such as phonemes.

Other pattern recognition methods include linear discriminant analysis (LDA), support vector machines (SVM), k-nearest neighbours (kNN) and Bayesian networks (BN). It would be beyond the scope of this thesis to describe them all. The reader is kindly referred to Huang et al. (2001) and Müller (2005) for detailed descriptions. However, the next section presents another common pattern recognition method: cluster analysis.

Cluster analysis

Data clustering or cluster analysis is used to classify data items (or feature vectors) with similar patterns into groups or clusters (Jain et al., 1999). In machine learning, supervised learning is used when the group affiliation of the objects is known in advance, while unsupervised learning is used to find structures in an unknown material. The degree of similarity between objects is often determined by means of a distance measure of some kind. One popular measure is the Euclidean distance, which is the straight-line distance between two points x and y in an N-dimensional space:

$$D_{Eu}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.1}$$

Some clustering methods are hierarchical. They can be divided into agglomerative (builds the hierarchy from the individual elements by the progressive

merger of clusters) and partitional (divides the data stepwise into increasingly smaller groups) hierarchical clustering methods. The result of hierarchical clustering is typically represented by a decision tree (DT). A decision tree consists of nodes and branches, where the root node is usually located at the top and the leaf nodes at the bottom. Each leaf node represents a decision.

Classification and regression trees – CART

Classification and regression trees (CART) is a machine learning technique which uses both statistical learning and expert knowledge to construct binary decision trees, formulated as a set of ordered yes/no questions about the features of the data. The best predictions based on the training data are stored in the leaf nodes of the CART. Its advantages over other pattern recognition methods include human-readable rules, compact storage, handling of incomplete and non-standard data structures, robustness to outliers and mislabelled data samples, and efficient prediction of categorical (classification) as well as continuous (regression) feature data (Breiman et al., 1984; Huang et al., 2001; Sigvardson, 2002). Figure 3.2 shows an example of a CART.

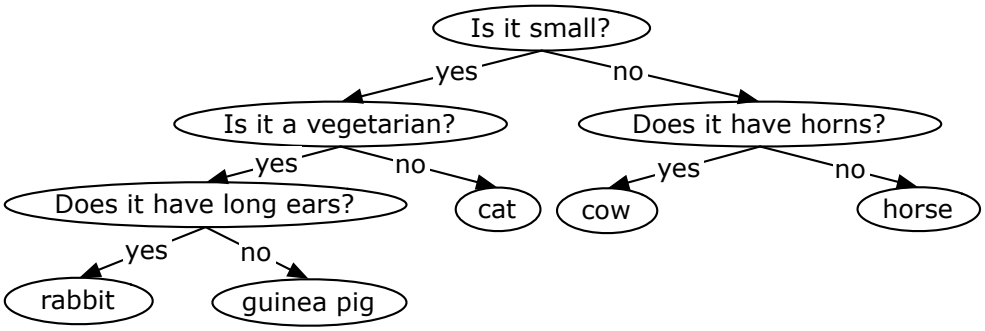


Figure 3.2: A simplified example of a CART for animal classification.

CARTs are trained using data material of the same type that will be used in the actual recognition task. Trees are built by successive partitioning or splitting of the material into smaller groups. A new node is created in each split, and the aim is to find, for each node, a question which partitions the material in the optimal way. Two important components, besides the construction of a question set, are measures of goodness and stop criteria. Measures of goodness are used in classification in order to find out how good (or bad) a split is. This is quantified by a measure of impurity. The objective of each new split is to reduce impurity for the whole tree.

Stop criteria are used to end the splitting at some point. A tree with only one item in each leaf node may be pure (as each node contains data samples from only one class), but it has become too specialised to be used for

anything but the training data. Therefore, the splitting is either stopped at an earlier stage of the tree building process, or the tree is pruned afterwards. Pre-pruning involves defining a stop criterion, e.g. when further splits do not reduce impurity, or when a pre-defined smallest number of items per node has been reached. Post-pruning is more complicated and involves finding the branch which can be collapsed into a single leaf node at the lowest cost.

The CART technique has previously been used to recognise phonetic qualities (Riley, 1991; Wang and Hirschberg, 1992; Frid, 2003). For instance, Frid (2003) automatically modelled rules for segmental as well as prosodic qualities in Swedish using CART. His letter-to-sound (LTS) conversion rules for 78,125 words resulted in 96.9% correct predictions for all letters. Frid also used CART learning to predict intonation (both by letter and by whole-word patterns) and Swedish word accent and dialect.

The next section presents previous research concerning automatic age estimation. Most studies have used features and methods described in the present section.

3.3 Previous related studies

So far, there have been relatively few attempts to build automatic age estimators. Table 3.1 provides an overview of previous studies. It is followed by a brief summary of the studies. Cepstral coefficients and perturbation measures have often been used as features. The number and age range of speakers vary among studies, as do the type of speech sample, the method used and the accuracy desired. Most studies have concerned age classification into two or three age groups.

Minematsu et al. (2002a,b) proposed a technique for automatic classification of perceived age (PA, judged by 12 students) using MFCC, ΔMFCC and amplitude derivatives (ΔPower) as acoustic features. Forty-three speakers previously judged as elderly and equally many speakers judged as non-elderly were modelled using GMM and normal distribution (ND). Two methods were used for classification: LDA and ANN. The first attempt correctly identified elderly speakers in 90.9% of cases using the LDA method. An attempt was then made to improve the classifier by including two additional features. These were speech rate, calculated as morae per time unit; and local perturbation of power, calculated as the number of power (amplitude) peaks per time unit satisfying the condition of differing by more than a threshold value from the previous peak. This increased the identification rate to 95.3%.

Table 3.1: Data and results from previous automatic age estimation studies (*pert.*: perturbation, *(?)*: no explicit information is given in the reference article, *correl.*: correlation, *CA*: chronological age, *PA*: perceived age)

Study	Minematsu et al., 2002	Shafran et al., 2003	Minematsu et al., 2003	Müller et al., 2003	Müller, 2005, 2006
Sample:					
language	Japanese	English	Japanese	German, English	German, English
type	read	spontaneous telephone calls	read	read, spontaneous	read, spontaneous
size/duration	15 sentences (training), 5 × 5 s (testing)	15 words (average)	60 s (training), 5 s (testing)	13 s (average)	single utterances
features	MFCC, ΔPower, ΔMFCC, speech rate, Power pert.	MFCC, F_0	MFCC, ΔMFCC, ΔPower	jitter, shimmer	F_0, speech rate, pauses, jitter, shimmer, HNR
Speakers:					
N	86	1,854(?)	407	439	1,164
gender	♀ ♂	♀ ♂	♂	♀ ♂	♀ ♂
age	20–60: 43, 60–90: 43	≤ 25: 93(?), 26–50: 1,075(?), ≥ 50: 686(?)	6–12: 123, 20–60: 141, 60–90: 143	> 60: 347, < 60: 46	10–97 (mostly aged 10–20 & 60–97)
Method:					
technique	GMM & ND, LDA & ANN	GMM & HMM	GMM & ND, LDA & ANN	5 different (ANN best)	6 different (ANN best)
task	PA: 2 groups	CA: 3 groups	PA: direct	CA: 2 groups	CA: 4 gender-specific groups
Accuracy	95.3% correct	70.2% correct	0.89 (correl. PA)	95.7% correct	63.5% correct

Shafran et al. (2003) used HMM-based classifiers on cepstral and F_0 features to recognise gender, age, dialect and emotion from a corpus consisting of 1,854 phone calls (65% female, 35% male callers) to a customer care system. The corpus contained a total of 5,147 utterances with an average length of 15 words divided into five age groups: (< 25, ≈ 25, 26–50, ≈ 50 and > 50). A trivial classifier assigning the most probable class label to all test points (33.3%) served as a "chance estimator" (baseline). Results for age were 68.4% correct classifications using only cepstral features, and 70.2% correct using cepstral as well as F_0 features.

Minematsu et al. (2003) conducted another study with 123 male speakers aged 6–12, 141 male speakers aged 20–60 and 143 male speakers aged 60–90. Thirty students in their early twenties estimated direct speaker age from single sentences. Each speaker was then modelled with GMM using MFCC, ΔMFCC and ΔPower as features. Two methods were used for the machine estimations. The first method modelled PA as discrete labels, while the second one was based on the normal distributions of PA. Both methods showed almost the same correlation between human judgements and machine estimation (0.89 for discrete labels and 0.88 for distributions).

Müller et al. (2003) compared six of the most common machine learning approaches for classification tasks – decision trees[1] (DT), ANN, kNN, naïve Bayes (NB) and SVM – in a study of automatic classification of age group and gender. Jitter and shimmer were extracted automatically and used as acoustic features. Two speech corpora, comprising 393 speakers (about 10,000 utterances from 347 speakers over 60 years, about 5,000 utterances from 46 speakers under 60 years; gender distribution: 162 females, 231 males), were used in the study. Results showed that all six methods performed significantly better than the baselines, which were simple classifiers always predicting the more frequently occurring class (elderly: 88%, male: 59%). ANN performed best with 81.09% correct gender judgements and 96.57% correct age group estimations.

Müller et al. also used Bayesian networks (BN) to integrate the gender classifier with the two age classifiers by first separately calculating the probability of a given speaker being female or male as well as being elderly or non-elderly, and then combining the results to obtain the most probable age and gender classification. This approach reduced errors likely to occur in a sequential classifier (gender first, then age), where failure to determine the correct gender strongly affects the performance of a gender-specific age classifier. For a more detailed description of BN, see Müller et al. (2003).

[1] C4.5 decision tree induction (Quinlan, 1993)

Müller (2005, 2006) further developed his approach for age and gender classification under the name AGENDER, with target applications such as mobile shopping and pedestrian navigation systems. First, the automatically extracted acoustic features were extended to include not only jitter and shimmer, but also F_0, HNR (harmonics-to-noise ratio) and speech rate (syllables per second) as well as pause duration and frequency. Second, the number of speakers was increased through the addition of another corpus to encompass a total of 507 female and 657 male speakers, divided into four age classes for each gender according to their CA (children: 10–12, teenagers: 13–19, adults: 20–65, seniors: 60–97). The majority of the speakers were children and seniors. Classification models were trained using the same five machine learning techniques as in Müller et al. (2003), i.e. DT, ANN, kNN, NB and SVM, as well as one additional method: GMM. The best accuracy for the four age classes was again obtained using ANN, with an overall accuracy of 63.5%, which was five times better than the baseline estimator. Adults were the most accurately classified (females 84.73%, males 87.87%).

To the author's knowledge, no attempt to estimate direct chronological age automatically has yet been made. Moreover, although the CART technique has been used to predict a number of phonetic qualities, it has not yet been used in age recognition experiments. The next two sections present two studies where the CART technique was used to estimate direct chronological speaker age. They distinguish themselves from previous automatic age recognition studies as regards the type of speech samples, the number and kind of acoustic features and the method used. In these studies, a single disyllabic word was used, since this type of stimulus had proved to yield fairly good human age estimates in a previous study (see Section 2.6). Moreover, the number of acoustic features investigated was significantly larger than in previous research.

In order to separate the CART method from the actual decision trees, henceforth the term *CART* will denote a single decision tree, while *CARTs* will be used about more than one tree. When reference is made to the method, the term will be used only in phrases, e.g. *the CART technique* or *CART learning*.

3.4 Study 3: CART estimation of direct age, age group and gender

Different phonemes contain different kinds of age-related phonetic cues. For instance, no F_0 is present in voiceless fricatives, and stressed vowels are known to generally hold more relevant phonetic information than unstressed

vowels or consonants (Lindblad, 1992; Bruce, 1998). Study 3 investigated which acoustic-phonetic features in word- and phoneme-sized segments are important in the estimation of age using the CART technique. Besides the estimation of direct CA, which is a continuous measure and therefore a matter of regression, this experiment also involved age group (old/young) and gender classification.

3.4.1 Purpose and aim

The main purpose of the study was to build automatic recognisers of direct chronological age, age group and gender from speech samples consisting of single words using the CART technique. Age group was included in order to allow comparison of the CART technique with methods used in previous studies. Gender was included to find out whether automatic gender classification with CART learning would be accurate enough to be used as an input feature of an age estimator. In addition, the study was designed to permit comparison between human and machine age perception, using the results from the human perception study described in Section 2.6.

The primary aim was not to build a state-of-the-art recogniser, but rather to identify important features in automatic age perception of adult age, age group and gender. Additional goals were to find out if automatic estimation of adult age would perform as accurately as human listeners, and whether automatic gender classification would be accurate enough to be used in gender-specific age recognition.

3.4.2 Questions and hypotheses

The following questions were addressed:

1. Which are the most important features in automatic age estimation from single words, and are different features important in different phoneme segments (e.g. vowels, fricatives)?

2. What accuracy could be reached by a CART estimator of age based on single words, and how would it compare with human perception of age?

3. Is gender a good input feature for automatic estimation of adult speaker age?

It was hypothesised that:

1. Automatic age estimators are likely to use different features in different phoneme segments (e.g. vowels, fricatives) when estimating age, as different phoneme types contain different kinds of phonetic information.
2. Automatic age estimation can reach the same accuracy as human age perception.
3. Gender is a good input feature for automatic estimation of adult speaker age.

3.4.3 Speech material

In order to be able to compare the results of this experiment with those of the study of human age perception described in Section 2.6, the same type of material was used here. It consisted of 2,048 isolated elicitations of *rasa* [ˈʁɑ̀ːsa] ('collapse') produced by 214 female and 214 male native speakers of Swedish, aged 17 to 84 years, from 36 villages in southern Sweden (Götaland). The material was taken from the SweDia 2000 speech corpus (Bruce et al., 1999). Each speaker had contributed between 3 and 14 elicitations of the word, and all of them were included in order for there to be some intra-speaker variation for the experiment. The 24 elicitations used in the human perception experiment were also included in the material. The mean word duration was about 0.4 seconds.

3.4.4 Method and procedure

The optimal method for this study would be one that is straightforward and easy to use. Combining statistical learning with expert (human) knowledge, the CART technique can use features that quite easily compare with the cues used by human listeners. In addition, the existence of a ready-to-use application which had been successfully used in previous phonetic recognition studies (see e.g. Frid, 2003), and the fact that the CART technique produces human-readable trees, made the choice of method an easy one.

Tools

Three major tools were used in this experiment: the speech analysis software *Praat* (see Section 2.6.4), *Align*, an automatic aligner, and *Wagon*, which is an implementation of the CART method.

The pre-processing stage included normalisation of the words for intensity (in Praat) and alignment. Automatic alignment of the phoneme segments was carried out with an aligner developed by Johan Frid[2] and used with his

[2]Department of Linguistics and Phonetics, Centre for Languages and Literature, Lund University.

kind permission. *Align*, which was called from a Praat script, generated a synthesised version of *rasa* to obtain known phoneme boundaries. Using an automatic alignment technique (Black and Lenzo, 2003; Malfrère and Dutoit, 1997), the aligner compared each frame of the synthesised version with the word to be aligned using Euclidian distances of MFCCs. The path with the shortest total acoustic distance was identified. This path would show how the synthetic version should be stretched or compressed to become as similar as possible to the word to be aligned, and how the two versions were temporally related. Based on the known boundaries in the synthesised version, the corresponding boundaries could be estimated in the word to be aligned.

Feature extraction was done using a Praat script. The script was originally developed for the author by Johan Frid. Most features were extracted using Praat's built-in functions and algorithms, namely values for duration, intensity, F_0, formant frequencies, jitter and shimmer. Some of the arguments to the functions had to be adjusted for female speakers. For instance, different min and max F_0 values were set for female and male speakers in Praat's pitch analysis. The script was adapted for this study by the author.

Pattern matching was carried out with Wagon, a CART implementation from the Edinburgh Speech Tools package (Taylor et al., 1999). It consists of two separate applications: *Wagon* for building the trees and *Wagon_test* for testing the trained trees with new data. Wagon supports discrete as well as continuous features in both input and output. It also contains a large number of options for controlling the tree-building processes. Only the three options controlled in the present study will be briefly explained here. A more detailed description of the Wagon tree building algorithm and its control options is given in Taylor et al. (1999). The *stop* value was used for fine-tuning the tree to the training set; the lower the value (i.e. the number of feature vectors in a node before considering a split), the more fine-tuned the tree and the larger the risk of it becoming overtrained. If a low *stop* value is used, the overtrained tree can be pruned using the *held_out* option, where part of the data set (e.g. 10%) is removed from the training set and then used for pruning to build smaller CARTs. All trees in this study were built with the *stepwise* option switched on, which means that the implementation, instead of considering all features, looked for and incrementally used the individual best features in order to build trees which were smaller and more general but involved a higher computational cost.

Pre-processing

The material was prepared for the CART experiments through signal normalisation and alignment. Just as in the human listener study, the words

were normalised for intensity so that only segment intensity relative to other parts of the word could be measured (see p. 10). The alignment tool was used to segment and transcribe all instances of the word into the SAMPA (Wells, 1997) phonetic alphabet – `rA:sa` – with fairly good accuracy, although some alignment errors did occur. Automatic alignment was preferred over manual in order to save time. Manual spot checks revealed that the script occasionally failed to segment a word correctly. An example of such a misalignment can be seen in Figure 3.8 (p. 78).

Feature extraction

A Praat script extracted 50 acoustic features from the four phoneme segments. The features were extracted only once per phoneme segment, often using mean values. This was preferred over feature extraction once every time frame (of e.g. 10 ms) to distinguish the phonetic information found in different phonemes, and also in order to keep the data size at a reasonable level. Features included a number of measurements (mean, median, range and SD) of F_0 and formant frequencies (F_1–F_5), as well as of relative intensity (see p. 10), segment duration, HNR (see p. 86), spectral emphasis (see p. 105) and spectral tilt (see p. 105). Appendix A, p. 168, shows a complete list of the features.

A description file containing all the feature names was created, and the feature vectors extracted were stored as two sets together with the following features:

- Segment label (as different phoneme segments contain different acoustic information)
- Chronological age (CA, in exact years, defined as a continuous feature because not every age was represented in the training material)
- Age group (a binary feature, where *old* was stipulated as 42 years or older – 42 being the youngest age defined as old in the SweDia 2000 corpus – and *young* was stipulated as younger than 42)
- Gender (a binary feature, which may influence age prediction)

One set was used only for comparison with the human listener study. It contained only the same 24 speakers and words (24 words × 4 segments = 96 vectors) that had been used in the human perception study. The other set comprised the other 404 speakers (1,924 words × 4 segments = 7,696 vectors), and was further split into a training set (90% = 6,157 vectors) and a test set (10% = 1,539 vectors). An example of an excerpt from the data vectors extracted is shown in Figure 3.3.

File	AgeGr	Age	Gend	Segm	Dur	F1_Mean	F1_Med.	F1_R	F1_R2	F1_SD	F2_Mean	F2_Med.
bre_yw_3_rasa_w4	y	23	w	r	0,085	521,067	538,095	116,788	128,024	51,44	1646,697	1619,231
bre_yw_3_rasa_w4	y	23	w	A:	0,15	649,592	654,847	66,991	58,074	21,2	1078,341	1056,504
bre_yw_3_rasa_w4	y	23	w	s	0,175	820,831	821,848	295,538	227,075	85,19	1781,646	1783,797
bre_yw_3_rasa_w4	y	23	w	a	0,12	643,357	624,743	189,988	184,572	79,83	1420,224	1410,155
bro_om_1_rasa_w1	o	57	m	r	0,175	835,897	865,693	319,708	248,448	108,34	1636,218	1816,945
bro_om_1_rasa_w1	o	57	m	A:	0,17	710,307	713,184	53,363	29,834	13,58	1268,937	1291,772
bro_om_1_rasa_w1	o	57	m	s	0,265	859,217	853,844	431,253	363,895	146,35	1982,889	2033,049
bro_om_1_rasa_w1	o	57	m	a	0,31	735,689	746,615	160,2	78,815	35,94	1376,886	1362,736
bro_om_1_rasa_w2	o	57	m	r	0,125	592,534	586,653	77,234	59,264	24,44	1162,478	1167,82
bro_om_1_rasa_w2	o	57	m	A:	0,185	682,304	678,015	52,013	44,132	18,05	1241,177	1247,226
bro_om_1_rasa_w2	o	57	m	s	0,22	800,153	815,537	410,55	284,958	107,32	1917,57	1914,37
bro_om_1_rasa_w2	o	57	m	a	0,315	721,863	730,687	148,731	108,622	40,99	1378,786	1368,209

Figure 3.3: Part of an example feature data set with metadata (file name, age, gender, etc.) and feature vectors (duration, mean and median F_1, etc.)

Pattern matching

The pattern matching procedure was somewhat tentative. A number of test runs were carried out in search of the best decision trees for each feature. Age and age group were estimated both with and without gender as an input feature. Gender was then predicted using neither age nor age group as input features. To reduce computation time, a subset of the data (489 words × 4 segments = 1,956 vectors) was used in an initial search for the option values that would generate the best trees. The *stop* value was in turn set to 2, 3, 4, 5, 10, 20, 50 and 100, and the *held_out* percentage value for pruning was variously set to 0, 10 and 20. These tests suggested that *stop* values of 3, 5 and 10 in combination with all three *held_out* values would generate the best prediction trees. In the remaining tests, the options were restricted to these values.

Baselines representing accuracy at a chance level were not easy to calculate, especially for age. Not every age was represented in the material, and the ages included in the data were not equally distributed. Since there were 57 ages ranging from 17 to 89 in the data, a rough baseline for age might either be calculated as 1/57 (≈ 1.75%) or as 1/(89–17 + 1) = 1/73 (≈ 1.37%). However, these values are neither comparable with the correlation between PA and CA, nor do they take account of the ages missing in the material. Age group and gender were both binary features. Female speakers were found in 3,928 out of the 7,696 vectors, so while one possible baseline for gender would be 51.04% (3,928/7,696), another would be 50%, given an expected equal distribution in the population to be predicted. For age group, a rough baseline might be 50%, since there were equally many (3,848) vectors for older as for younger speakers. However, since the range of CA was 42 (distributed as 36 different ages) for the old group, but only 19 (each age from 17 to 35) for the young one, this is not really a representative value. The baselines suggested in the result tables below should thus be regarded only as rough estimates of the performance of a baseline predictor.

In the first actual test runs, the whole data set containing feature vectors for all phoneme segments was used. Additional tests were run using feature vectors of one segment at the time. This was done in order to obtain some idea of which phonemes contained the best information for age and gender prediction (i.e. generated the best trees). An additional goal was to find out if the CARTs preferred different features for different phoneme segments. Finally, tests were run with the data set containing the same 24 words as in the human listener study using the best CARTs for each segment. The first (i.e. best) features of the resulting trees were compared with the cues preferred by the human listeners (see Section 2.6.6).

3.4.5 Results

Tests with all segments

Table 3.2 displays the results for the best CARTs using all phoneme segments. The highest correlation between chronological and estimated age for direct (continuous) age recognition was 0.385 (with the input feature of gender). For age group (old/young), the best percentage of correct estimates was 66.80 (gender used as input feature), and for gender the best result was 83.63% correct estimates. All estimators performed significantly better than the baselines.

Table 3.2: Results of the best CARTs for age, age group and gender using all features

Continuous feature	*correlation*	*baseline*
Age (without gender)	0.344	0.019
Age (with gender)	0.385	0.019
Discrete feature	*correct (%)*	*baseline (%)*
Age group (without gender)	65.37	50
Age group (with gender)	66.80	50
Gender	83.63	51.04

Tests with one segment at a time

Table 3.3 shows the best results for each segment. The stressed vowel `A:` yielded the best results, with a correlation between chronological and estimated direct age of 0.454, 72.14% correct estimations for age group and 90.62% correct estimations for gender. Including gender as an input feature only marginally influenced the results of the trees.

The features used in the first three yes/no questions (i.e. in the root node and two levels down) of the best CARTs for each segment are shown in Table 3.4.

Table 3.3: Results for the best CART estimators of age, age group and gender for each segment (best values in boldface)

Segment	*Age (excl. gender)*	*Age (incl. gender)*	*Age group (excl. gender)*	*Age group (incl. gender)*	*Gender*
r	0.299	0.299	65.10	65.10	77.34
A:	**0.446**	**0.454**	**72.14**	**72.14**	**90.62**
s	0.406	0.393	64.06	64.84	80.99
a	0.273	0.286	63.28	63.28	87.50
baseline	0.019	0.019	50	50	51.04

Table 3.4: Top three features used by the best CARTs for each segment to estimate (a) age (without gender), (b) age (with gender), (c) age group (without gender), (d) age group (with gender) and (e) gender (*Rel. in.*: relative intensity)

(a)	r	A:	s	a
1st	F_3 (median)	F_4 (mean)	F_1 (range)	F_1 (range)
2nd	F_1 (range)	F_5 (range)	F_2 (range & mean)	F_4 (median)
3rd	F_2 (mean)	F_2 (mean)	F_4 (mean & median)	F_5 (median)
(b)	r	A:	s	a
1st	F_3 (median)	F_4 (median)	F_1 (range)	F_2 (median)
2nd	F_2 (median & SD)	HNR	Rel. in. (mean)	Shimmer
3rd	Rel. in. & F_0 (mean)	F_2 (mean)	F_4 (mean)	F_0 (mean & median)
(c)	r	A:	s	a
1st	F_3 (mean)	HNR	F_1 (range)	F_1 (range)
2nd	F_4 (range)	F_1 (median)	F_3 (mean)	F_2 (mean)
3rd	F_2 (median)	F_0 (range)	F_4 (median & range)	F_5 (mean)
(d)	r	A:	s	a
1st	F_3 (mean)	HNR	F_1 (SD & range)	F_1 (range)
2nd	F_4 (range)	F_1 (median)	Rel. in. (mean)	F_2 (mean)
3rd	F_2 (median)	F_0 (range)	Gender	F_5 (mean)
(e)	r	A:	s	a
1st	F_0 (median & mean)	F_0 (median)	F_5 (median)	F_0 (median & range)
2nd	F_1 (median)	F_1 (mean)	F_0 (mean)	F_1 (median)
3rd	Rel. in. (range)	F_4 (median)	F_2 (median)	F_2 (median)

The four segments often had different features in their top three postitions. For age, questions about the formant frequencies dominated, but HNR, relative intensity (Rel. in.), F_0 and shimmer were also used. Important cues for the age group CARTs were formant frequencies, and sometimes HNR, relative intensity, shimmer and gender. Often different features were used in the first questions when gender was included among the input features than when it was excluded, and the feature of gender occurred only once among all top-three questions. The trees for gender prediction depended on first questions about F_0 values, but also on questions about formant frequencies and relative intensity.

Comparison with human listeners

Table 3.5 compares the chronological age of the 24 speakers with the mean PA as judged by 30 human listeners (see Section 2.6) and the best CART estimations for each speaker. The human listeners judged the age of 13 speakers more accurately, while the CART judged 9 of the speakers with higher accuracy. Two speakers were estimated equally well by humans and the CART. Neither the human listeners nor the automatic estimator was clearly better than the other at judging the age of female or male speakers. A trivial baseline estimator judged all speakers to be 47.3 years old (the mean CA of all speakers), which yielded a mean absolute error of ±20.7 years. Both humans and the CART performed better than the baseline estimator.

Table 3.5: Chronological age (CA), human perceived age (PA) and CART (using the segment `A:`) age estimates of 24 speakers (closest estimates in boldface; speaker ID *(Sp.ID)*: village (a/s) + age group (o/y) + gender (m/w) + number (1–3))

Sp.ID	*syw1*	*sym2*	*sym1*	*syw3*	*aym1*	*aym2*	*ayw2*	*ayw3*	*aym3*	*sym3*	*syw2*	*ayw1*
CA	18	20	22	24	27	27	28	28	29	29	30	31
PA	36	49	39	27	**43**	**28**	**30**	**24**	41	**34**	45	**35**
CART	**24**	**48**	**25**	**24**	67	**26**	34	57	**28**	53	**32**	44
Sp.ID	*aom3*	*aow3*	*aow1*	*som1*	*aom1*	*som3*	*sow3*	*aom2*	*sow2*	*sow1*	*som2*	*aow2*
CA	60	60	61	62	66	70	70	71	72	73	76	82
PA	46	47	**61**	**51**	**60**	**68**	57	**62**	**66**	**75**	**70**	**75**
CART	**72**	**70**	55	**73**	45	51	**65**	26	64	65	48	64

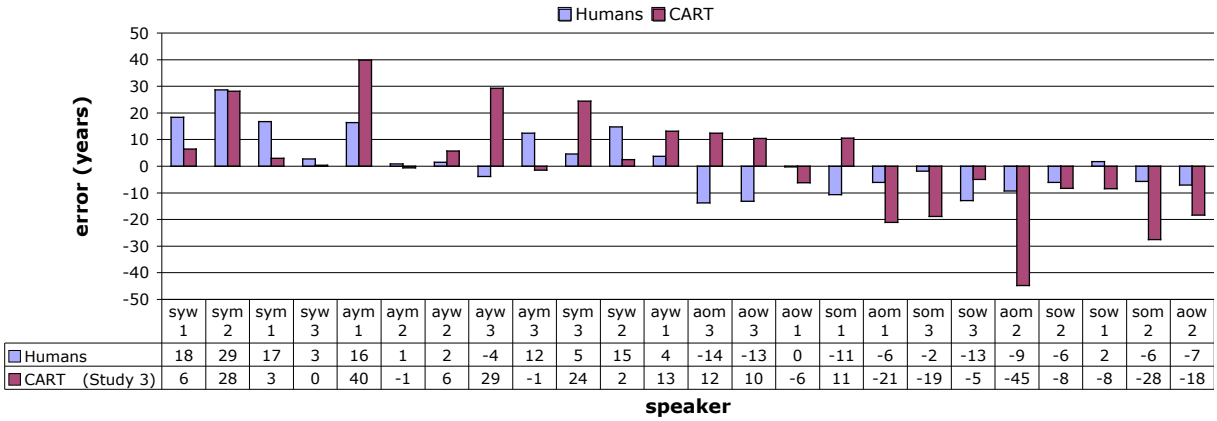


Figure 3.4: Deviation of human and CART age estimations from chronological age (24 speakers, youngest to the left, oldest to the right). For chronological age and explanations of speaker ID, see Table 3.5

A comparison of the misjudgements (in years) made by the humans and the best CART is shown in Figure 3.4, where the speakers are ordered according to their chronological age, with the youngest speaker (syw1) to the

left and the oldest speaker (aow2) to the right. The largest errors were made by the CART when trying to determine the age of one young (aym1) and one old (aom2) male speaker. For the CART estimations, the mean absolute error between the estimated and chronological age of the speakers was 14.45 years. The corresponding value for the human listeners was 8.89 years.

Comparison of the features used by the CARTs to predict age with the acoustic correlates of the cues used in the human listener study showed several similarities. Spectral cues (e.g. formant frequencies) were found to be more important than F_0 for both humans and CARTs. However, the human study also found duration to be an important phonetic cue to age, while the CARTs did not use duration in their first questions.

3.4.6 Discussion and conclusions

This experiment was a first attempt to build an automatic age estimator for single-word input using the CART technique. The following sections discuss possible explanations for the results.

Material and feature extraction

Speech produced by 428 speakers (214 females, 214 males) of southern Swedish dialects was used as material in the study. Factors such as the age distribution and dialect variation of the speakers are highly likely to have affected the results. Errors in the alignment and automatic feature extraction methods may also explain the poor performance of the CART estimators. This is discussed further in Section 3.6.

Owing to the small data set, one cannot be certain that the features used by the CARTs in this study actually mirror important age cues. More research with a larger material is needed to determine this.

Single-segment and all-segment CARTs

The hypothesis of automatic age estimators using different features in different phoneme segments was confirmed, as the top three features used by the best CARTs for each segment often varied.

The trees based on all phoneme segments did not perform as well as the ones using only the segments *A:* or *s*. This may be explained by limitations in the CART technique in terms of relating features from different segments, and by the fact that stressed vowels contain the most phonetic information. The relatively good performance of the CARTs for *s* is in line with a previous study (Schötz, 2003a), which found that the typical energy plateaus for [s] begin at higher frequencies for younger-sounding speakers. The segment *r*

displayed large allophonic variation among the speakers, which may explain the poor performance of the CART for *r*. Longer segment duration may be another reason why the predictors for *A:* and *s* outperformed those for the other segments.

3.4.6.1 Gender

The CART estimator for the segment *A:* was the best gender classifier (90.62% correct). However, the age estimators did not improve their accuracy when gender was included as an input feature, contradicting the hypothesis. The results suggest that it might be better to build separate estimators of age for female and male speakers.

CART and human age perception

Both the human and CART results were better than chance, and the human listeners (mean error: ± 8.89 years) were clearly better at predicting age than the CARTs (mean error: ± 14.45 years).

Although the age cues for the human listeners displayed similarities with the features used by the CARTs, this does not necessarily mean that humans and automatic estimators use similar strategies when estimating age. The features used by the CARTs may, however, give some indication about where to look for acoustic correlates of the cues used by humans in age perception. Studies using methods to extract more acoustic features (e.g. inverse filtered features, LTAS and formant bandwidths and levels) and a larger, more varied speech material are needed to further increase phonetic knowledge about human and machine perception of speaker age.

Conclusions

From this study, the following conclusions may be drawn:

1. Important features for automatic age estimation using the CART technique are mainly spectral features, including formant frequencies, HNR and intensity, though different features were important in different phoneme segments.

2. A CART estimator based on a single segment of an isolated word (mean error ± 14.45 years) does not reach the performance of human listeners (± 8.89 years).

3. Although gender was estimated with $> 90\%$ accuracy, information about gender does not considerably influence age estimation using the CART technique.

3.5 Study 4: Important features in CART estimation of age

This study was mainly an attempt to improve the age recogniser described in the previous section by increasing the number of speakers and features used in the training, by building separate estimators for female and male speakers and by combining different groups of features.

3.5.1 Purpose and aim

The purpose of this study was to find the most important features in automatic age recognition. By systematically varying a large number of features, optimal feature combinations could be identified for female and male age estimation. The aim was to gain more knowledge about the various acoustic cues to speaker age in machine age perception as well as contributing to the improvement of automatic recognition of speaker-specific qualities.

3.5.2 Questions and hypotheses

In this study, the following two research questions were asked:

1. Which is the most important single feature in automatic age estimation from a single word?

2. Which is the best combination of features?

Partially based on the human perception study described in Section 2.6, one hypothesis was tested:

1. Spectral features (both source and resonance) are more important than prosodic features (F_0, duration, relative intensity) in automatic age estimation using single words.

3.5.3 Speech material

The speech material was extended from the previous CART study (Study 3), and it now consisted of 3,026 elicitations of *rasa* produced semi-spontaneously in isolation by 748 adult native Swedish speakers, equally many females as males, aged 17 to 89 years. All speakers were from 66 villages in the southern parts of Sweden and Finland (428 speakers from Götaland, 320 from Svealand and Finland). Each speaker had contributed between 3 and 14 elicitations of the word, and all elicitaions were included so as to obtain some intra-speaker variation in the experiment. There was extensive dialectal variation,

including intonation differences for the Swedish grave word accent (accent II) as well as allophonic variation of the phonemes /r/, /ɑː/ and /a/, with pronunciations ranging from the central Swedish [ˈɹɑ̀ːsa] to [ˈʁɑ̀ːsa], [ˈʁɑ̀ːsə] and even [ˈwɑ̀ːsə].

3.5.4 Method and procedure

The method and tools employed for this study were almost exactly the same as the ones described in Chapter 3.4.

Tools

The same aligner was used to segment the words into phoneme segments (see Section 3.4.4). Praat was again used as a tool for intensity normalisation and feature extraction. The feature extraction script was based on the one used in Study 3, but it was further developed to extract features from the whole-word segments as well, and also to extract a larger number of features, including formant bandwidths and levels. Wagon was again used in the CART experiments.

Pre-processing

The same built-in Praat function as in Study 3 was used to normalise the words for intensity (see p. 10). Alignment was also done in exactly the same way as in the previous study.

Feature extraction

The expanded Praat script was used to extract features separately for female and male speakers. The number of features was rather large: 78 for each segment, adding up to a total of 390 (78 × 5) features for each word. Features included a number of measurements (mean, median, range and SD) of fundamental frequency (F_0), formant frequencies, bandwidths and levels (F_1–F_5, B_1–B_5 and L_1–L_5), as well as of relative intensity (see p. 10), segment duration, HNR (harmonics-to-noise ratio), spectral emphasis (see p. 105) and spectral tilt (see p. 105). Also, several measurements of jitter and shimmer were extracted as well as energy peaks in several frequency bands – as measurements of spectral balance – in both the original and the inverse filtered signal. Appendix A (p. 169) shows a complete list of the features. A total of 3,002 feature vectors were extracted from equally many words for 748 speakers.

Description files containing the names and type (categorical or continuous) of the features to be used in the CART experiments were prepared.

Several data sets were created. The 24 speakers who had produced the words used in the human perception test in Section 2.6 were excluded from the data sets used to train the CART. Instead they were used to compare the CART results with human age perception. Three data sets – one for females, one for males, one for all speakers – were created from the feature vectors extracted for the other speakers. Again, the feature vectors were stored together with the same non-acoustic parameters as in Study 3: segment label, chronological age and gender (see p. 60). The three data sets were further split into training sets (80%) and testing sets (20%).

Pattern matching

Although the pattern matching procedure was similar to the one used in the previous study, there were some differences as well. To reduce computation time, a number of test runs were carried out before the actual experiments using a subset of the data (489 words) to find the option values that would generate the best trees. The *stop* value was in turn set to 2, 3, 4, 5, 10, 20, 50 and 100, and the *held_out* percentage value for pruning was variably set to 0, 10 and 20. These tests suggested that a *stop* value of 4 in combination with a *held_out* value of 0 would generate the best prediction trees. In the remaining tests the options were restricted to these values. A total of 409 actual experiments for the three speaker groups of females, males and all speakers were carried out to build and train age estimation CARTs using a number of feature combinations, as shown in Table 3.6.

Table 3.6: Features and feature combinations used to build the CART estimators of age and gender for female, male and all speakers (*sp.*: spectral)

Feature(s)	*Predictee*	*CARTs (N)*
1 single feature	age	390
Feature groups (duration, F0, F1, F2, F3, F4, F5, relative intensity, jitter, shimmer, sp. balance, inverse filtered sp. balance and one group for sp. emphasis, sp. tilt and HNR)	age	13
All prosodic features (duration, F0 and intensity)	age	1
All resonance (formants, spectral distribution) features	age	1
All inverse filtered features along with sp. emphasis, sp. tilt and HNR	age	1
All spectral (resonance and inverse filtered) features	age	1
All features	age	1
All features	gender	1
Total No of CARTs		409

An attempt to determine gender using all features together was also made. In all, 1,227 (409 × 3) CARTs were trained and tested in the experiment. Finally, CARTs based on all features for gender as well as for female

and male age were used to estimate the gender and age of the 24 speakers used in the human perception study, and the results were compared.

3.5.5 Results

Single features

Wagon provided two types of results for the CARTs: correlation between estimated and chronological age as well as mean absolute error. As the two values yielded similar results, only mean error will be used here. Results for the ten best single features, presented as the absolute mean error in years, are shown in Table 3.7.

Table 3.7: Mean (absolute) error (ME) in years for the ten best single-feature CARTs for female, male and all speakers (prosodic features in boldface). For an explanation of features, see Appendix A (p. 169)

Female speakers			*Male speakers*			*All speakers*		
No	*Feature*	*ME*	*No*	*Feature*	*ME*	*No*	*Feature*	*ME*
1	**F0_Mean_r**	18.58	**1**	**Dur_A:**	18.38	**1**	**Dur_rAsa**	**19.08**
2	**F0_Mean_A:**	18.71	**2**	**F0_R_A:**	18.78	**2**	**Dur_A:**	**19.09**
3	**F0_Med_rA:sa**	18.87	**3**	**Dur_rA:sa**	18.81	**3**	**F0_Mean_r**	**19.30**
4	**F0_Med_r**	19.07	4	Skew_s	18.98	4	F1_SD_s	19.45
5	**F0_Med_s**	19.10	5	F1_R2_s	18.99	**5**	**F0_Med_r**	**19.47**
6	**F0_Mean_s**	19.33	**6**	**F0_SD_rA:sa**	19.01	**6**	**F0_SD_rAsa**	**19.54**
7	F2_R_s	19.39	7	InKurt_a	19.02	7	F1_R_s	19.60
8	**F0_Med_A:**	19.50	8	F2_Med_rA:sa	19.15	**8**	**F0_R_rAsa**	**19.68**
9	**Dur_s**	19.58	9	F1_R2_rA:sa	19.16	**9**	**F0_SD_A:**	**19.72**
10	**Dur_rA:sa**	19.59	**10**	**F0_R_rA:sa**	19.21	**10**	**Dur_s**	**19.73**

The results for the single-feature CARTs differed for female, male and all speakers. However, the prosodic features dominated all three top-ten feature lists. In the female speaker list, F_0 values were found in all positions from 1st to 6th and also in 8th place, while duration values were found in 9th and 10th position. The only non-prosodic feature was F_2 in 7th place.

Male speaker top-ten prosodic values comprised duration in 1st and 3rd place, and F_0 in 6th and 10th place. Spectral and inverse filtered features held the other five positions: spectral skewness (4th place), inverse filtered kurtosis (7th place), F_2 (8th place) and F_1 (9th place).

In the CART results for all speakers, duration and F_0 dominated the top-ten list, but with two values for F_1 in 4th and 7th place.

The overall single best feature was duration of [ɑː] for male speakers (mean error: 18.38 years), while the worst single-feature result was that for median F_5 (mean error: 23.40 years) for female speakers. The best scores for a few other, possibly age-related, features are listed in Table 3.8.

Table 3.8: Best ranking and mean error (ME) for some other features for female-, male- and all-speaker CARTs. For an explanation of features, see Appendix A (p. 169)

	Female speakers			*Male speakers*			*All speakers*		
Feature	*Value for*	*No*	*ME*	*Value for*	*No*	*ME*	*Value for*	*No*	*ME*
F3	B3_a	20	19.83	B3_a	11	19.23	F3_Med_s	20	20.09
F4	F4_Med_s	19	19.80	F4_Mean_s	30	19.54	L4_A:	40	20.31
Jitter	Jitt_local_r	18	19.79	Jitt_locabs_r	43	19.73	Jitt_rap_r	17	20.01
Shimm	Shi_loc_dB_r	39	20.21	Shi_apq5_a	53	19.82	Shi_loc_db_r	24	20.14
SpEmph	SpEmph_s	29	20.12	SpEmph_rA:sa	32	19.57	SpEmph_s	11	19.87
SpTilt	SpTilt_a	81	20.62	SpTilt_A:	95	20.02	SpTilt_rA:sa	250	21.16
HNR	HNR_r	21	19.89	HNR_r	92	20.01	HNR_r	39	20.30
SpBal	SpSD_s	24	20.05	Skew_s	4	18.98	SB1-2_a	22	20.13
InvSpBal	InCOG_a	30	20.13	InKurt_a	7	19.02	ISB2-4_rA:sa	44	20.33

Feature groups

The results for the 13 feature groups can be seen in Figure 3.5, where the y-axis scale begins at 16 years so that the difference between the groups can be seen more clearly.

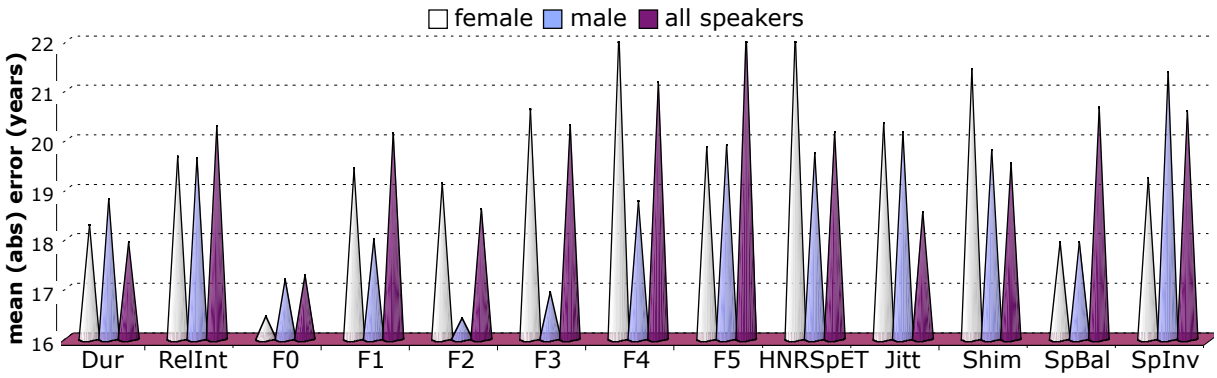


Figure 3.5: Mean absolute error of the CART age estimations for female, male and all speakers using the 13 feature groups. Note that the y-axis scale begins at 16 years (*Dur*: duration, *RelInt*: relative intensity, *HNRSpET*: feature group consisting of HNR, spectral tilt and spectral emphasis, *SpBal*: spectral balance, *SpInv*: feature group of inverse filtered spectral balance)

F0 shows the best results for females as well as for all speakers. The best male result was obtained for F2, followed by F3 and F0. Duration displayed quite a low mean error overall. Formant values resulted in much lower values for males than for females (except for F5) and all speakers. The mean error for spectral balance was quite low for the female and male groups, but not for the CART using all speakers.

Larger feature groups and all features

The results for the prosodic, resonance, inverse filtered and spectral (i.e. both resonance and inverse filtered) CARTs and the CARTs using all features are displayed in Figure 3.6. To show the difference between the different CARTs more clearly, the y-axis scale begins at 15 years.

The best estimator for female age was the one using all features, while for male and all speakers it was the one using only prosodic features. Generally, prosodic features performed better than the others. For male and all speakers, the all-feature estimator had the highest mean error of all five CARTs. The gender-specific estimators consistently performed better than the CARTs using both genders.

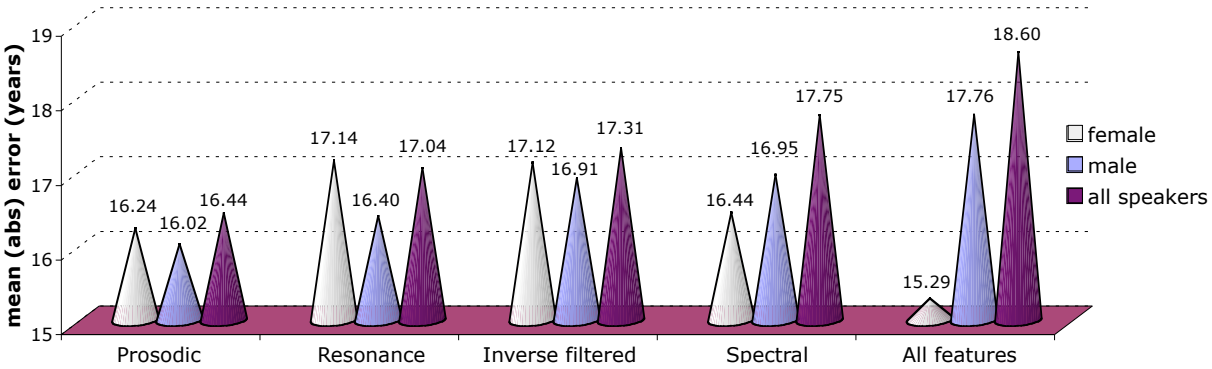


Figure 3.6: Mean absolute error of the CART age estimations for female, male and all speakers using prosodic, resonance, inverse filtered, spectral as well as all features. Note that the y-axis scale begins at 15 years

Gender

The CART experiment to determine gender using all features resulted in 285 out of 307 (92.8%) correct estimations for female speakers, and 276 out of 293 (94.2%) correct estimations for male speakers.

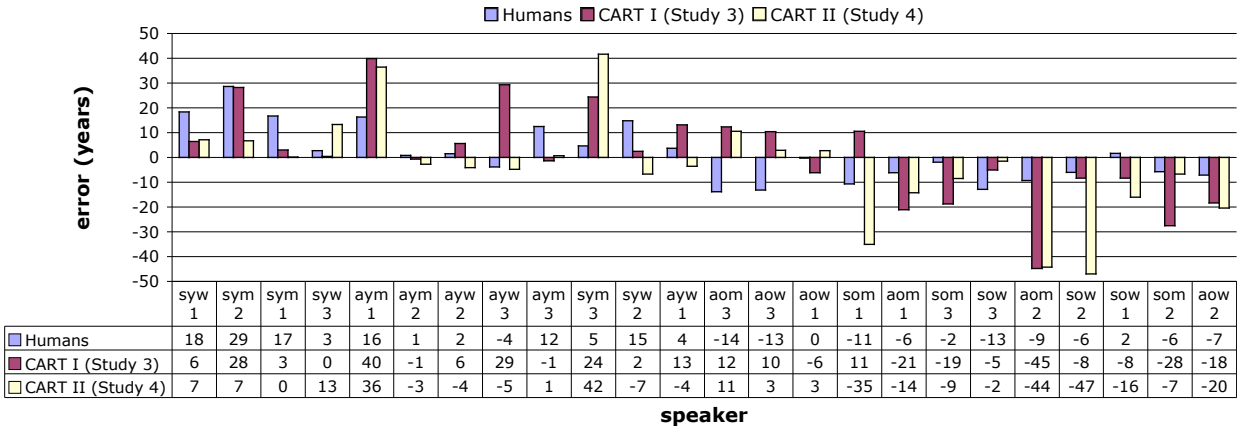
Comparison with human listeners and the previous CARTs

The gender CART correctly estimated the gender of all 24 speakers. As for age, Table 3.9 displays the CA as well as the mean PA for the 24 speakers as estimated by the 30 human listeners. It also shows the estimates made by the best CART from Study 3 (CART I) and by the CART in this study using all features for each gender (CART II). Although the estimates made by CART II were closer than the ones made by CART I for 13 speakers, human listeners still outperformed both CART estimators. Human estimates were better for 15 speakers, while one of the CARTs more accurately recognised

the age of 8 of the speakers. Two speakers were judged equally well by both human listeners and the CARTs.

Table 3.9: Chronological age (CA), human perceived age (PA) and CART (Studies 3 & 4) age estimates of 24 speakers (closest estimates in boldface, speaker ID *(Sp.ID)*: village (a/s) + age group (o/y) + gender (m/w) + number (1–3))

Sp.ID	*syw1*	*sym2*	*sym1*	*syw3*	*aym1*	*aym2*	*ayw2*	*ayw3*	*aym3*	*sym3*	*syw2*	*ayw1*
CA	18	20	22	24	27	27	28	28	29	29	30	31
PA	36	49	39	**27**	**43**	**28**	**30**	**24**	41	**34**	45	**35**
CART I (Study 3)	**24**	48	25	24	67	**26**	34	57	**28**	53	**32**	44
CART II (Study 4)	25	**27**	**22**	37	63	24	24	23	**30**	71	23	**27**
Sp.ID	*aom3*	*aow3*	*aow1*	*som1*	*aom1*	*som3*	*sow3*	*aom2*	*sow2*	*sow1*	*som2*	*aow2*
CA	60	60	61	62	66	70	70	71	72	73	76	82
PA	46	47	**61**	**51**	**60**	**68**	57	**62**	**66**	**75**	**70**	**75**
CART I (Study 3)	72	70	55	73	45	51	65	26	64	65	48	64
CART II (Study 4)	**71**	**63**	64	27	52	62	**68**	27	25	57	69	62



	syw 1	sym 2	sym 1	syw 3	aym 1	aym 2	ayw 2	ayw 3	aym 3	sym 3	syw 2	ayw 1	aom 3	aow 3	aow 1	som 1	aom 1	som 3	sow 3	aom 2	sow 2	sow 1	som 2	aow 2
Humans	18	29	17	3	16	1	2	-4	12	5	15	4	-14	-13	0	-11	-6	-2	-13	-9	-6	2	-6	-7
CART I (Study 3)	6	28	3	0	40	-1	6	29	-1	24	2	13	12	10	-6	11	-21	-19	-5	-45	-8	-8	-28	-18
CART II (Study 4)	7	7	0	13	36	-3	-4	-5	1	42	-7	-4	11	3	3	-35	-14	-9	-2	-44	-47	-16	-7	-20

Figure 3.7: Deviation of human and CART age estimates from chronological age in Studies 3 and 4 (24 speakers, youngest to the left, oldest to the right). For chronological age and explanations of speaker ID, see Table 3.5

A comparison of the misjudgements (in years) made by the humans, the best CART from Study 3 and the all-feature CART for each gender from this study is shown in Figure 3.7. The largest errors were made by the CARTs trying to predict the age of two young men (aym1, sym3), two old men (som1, aom2) and one old woman (sow2). The mean absolute error for the CART II estimates was 14.07 years, while the corresponding result for the human listening study was 8.89 years. This is an improvement of only 0.38 years, since the CART I estimates showed a mean error of 14.45 years.

Comparison of the features used by the CARTs to predict age with the acoustic correlates of the cues used in the human listener study showed several differences. Spectral cues (e.g. formant frequencies) were dominant over

F_0 for humans, while the CARTs relied more on F_0. However, although of no great concern in Study 3, both the human and the present CART studies found duration to be an important phonetic cue to age.

3.5.6 Discussion and conclusions

Study 4 was an attempt to improve the CART estimator from Study 3 in three ways: (1) by increasing the number of speakers and features, (2) by including features extracted from word segments and (3) by building separate estimators for female and male speakers. The improvement actually made was only marginal, but the experiment nevertheless provided some interesting results, which are discussed below.

Prosodic vs. spectral features

Prosodic features were found generally to perform better than spectral (non-prosodic) ones, i.e. both when single features, feature groups and larger groups of features were used in the CART estimators. These results are not in line with the findings of Studies 1 and 3. The hypothesis that spectral features are more important than prosodic ones was thus not confirmed. Interestingly, comparison of the 13 feature groups in Figure 3.5 for male speakers shows that the F2 and F3 groups outperformed the F0 and duration groups. However, comparison of the CARTs based on the larger feature groups in Figure 3.6 shows that the prosodic estimators performed slightly better than the spectral ones. This change in feature preferences is difficult to explain, and it is further discussed below.

Performance

Why did the new estimator perform only slightly better than the old estimator in the final test with the 24 speakers? There are a number of factors which may have contributed to this.

The speech material was extended to improve performance compared with the previous study. However, this may have had the opposite effect. Study 3 contained only speakers from Götaland. The extended material also comprised speakers from Svealand and Swedish-speaking southern Finland, thus covering a much larger dialectal area where other dialect types are spoken. This is likely to have increased inter-speaker variation. The 24 speakers of the final human comparison test, however, were all from Götaland, so possibly the estimator had little use for the increase in material when estimating the age of these speakers. A small test with the CART technique using a

subset with speakers from the same geographic region (Småland, a county in Götaland) showed a substantial improvement in performance.

Using 80% of the material for training may have been insufficient, considering the larger dialectal variation of the extended material. Results might improve if 90% were used. Other explanatory factors may be related to the unexpected shift in features of importance from spectral to prosodic ones for the male speakers. It is possible that the increased number of features could have contributed to this result. The validity of this outcome should be tested further, perhaps with a reduced number of feature values or with more compact features (e.g. MFCC) as substitutes for the ones used here.

Machine vs. human age perception

The slight improvement of CART performance obtained in Study 4 was far from enough to challenge human listeners. Humans still outperformed the CARTs, even when these used more features and were trained using a larger material, and when separate estimators were built for female and male speakers. Apart from the possible explanatory factors for the poor CART results mentioned above, the automatic feature extraction, further discussed in Section 3.6 below, may have contributed to the poor results for the CART estimators.

Conclusions

Based on the findings of this study, the following tentative conclusions were drawn:

1. In automatic recognition of speaker age, the most important single acoustic features appear to be mean F_0 (female speakers) and duration (male speakers). When all measures of one feature are combined into feature groups, F_0 measures (mean, range, etc.) seem to be the most important ones for females, while measures of F2 (F_2, B_2 and L_2) are likely to be the most important ones for males. However, prosodic features appear to be generally more important than spectral (resonance and inverse filtered) features for both female and male speakers.

2. CART estimators of speaker age use different features for female and male speakers.

3. Speaker gender can be successfully estimated with the CART technique.

3.6 Summarising discussion

The two studies described in this chapter differed from previous studies in several ways. They used a larger number of features, extracted from shorter speech segments (single words and phoneme segments) which were more acoustic-phonetic in nature than the more "typical" ASR features used in most previous studies. The CART method had never been used in age recognition studies before, and was chosen partly because of its ability to produce human-readable rules. Although the CART estimators did not manage to perform as well as human listeners, they nevertheless provided some interesting results which point towards a number of problems yet to be solved in the pursuit of a state-of-the-art estimator of speaker age. A number of these problems are discussed here along with some suggestions for future studies.

3.6.1 Speech material

When relating the results to the material, one should bear in mind that the two attempts to build automatic age estimators described here used only one single disyllabic word. One reason for the choice of material was the fairly good ability of human listeners to estimate age from single words (Schötz, 2001a). Better results would probably have been obtained with longer speech samples. Minematsu et al. (2002a) found that for their system, speech samples of 5 seconds were enough to generate stable identification performance, suggesting that single words are too short for reliable automatic age detection. Moreover, human listeners' ability to judge age improves if longer speech samples and spontaneous speech are used as stimuli (see Section 2.8).

Müller et al. (2003) found that to minimise the effect of random fluctuations in the data set, a large number of speech samples is better. One should, however, be careful when extending the speech material. In the studies presented here, dialectal variation increased as the material was extended. The extended material contained variation of the Swedish grave word accent as well as of the phonemes /r/ and unstressed /a/, with pronunciations ranging from the central Swedish [ˈɹɑ̀ːsa] to [ˈʁɑ̀ːsa], [ˈʁɑ̀ːsə] and even [ˈwɑ̀ːsə]. This is likely to have influenced the results. If possible, future studies should keep dialectal variation in the speech material at a minimum level.

Unbalanced data sets may also cause problems (Müller et al., 2003). As already mentioned, the speech samples for both studies were taken from the SweDia 2000 corpus, which has an even gender distribution. However, given that the aim of the project was to document only a younger and an older generation of adult speakers, far from every biological age was represented. In Study 4, no speakers were under 17 or over 89, and only three speakers

were between 36 and 51 years old. Most of the younger speakers were between 20 and 33 years, and most of the older ones between 55 and 77. This must have reduced CART estimation performance. Future errors related to age distribution can be avoided if more speakers of the under-represented ages are recorded and used in estimation experiments.

3.6.2 Method and procedure

Automatic extraction of speech parameters is problematic, owing to the large intra- and inter-speaker variation in voice type and register. No 100% reliable feature extraction methods exist today. Common errors in F_0 extractions are "jumps to the second harmonic indeterminacy of F_0 at aperiodicities such as voice creaks" Fant (2005). For formant frequencies and bandwidths, the task is even more difficult, and typical errors are "temporal jitter from one frame to the next, tendencies at high F_0 to synchronize on harmonics, and occasional jumps to a higher or a lower formant" (Fant, 2005). Künzel (2001) found automatic formant analysis to be very unreliable, i.e. about 50% error, which made him abandon automatic formant detection altogether. He found that female speakers were more difficult to analyse than males, and that most measurements failed because the algorithm picked a harmonic or a higher formant instead of the correct formant.

Still, the features were extracted automatically in the studies described in this chapter. While this saves time compared with manual feature extraction, one should always double-check automatic methods manually to reduce the influence of measurement errors. This was done here only to some extent. One effect observed was the use of F_0 as an important feature to predict gender in the voiceless segment *s*. One example of an alignment error is shown in Figure 3.8. Better alignment techniques will have to be used in future studies to eliminate such errors.

3.6.3 Prosodic vs. spectral features

Why did the prosodic features perform better in Study 4, while the spectral features seemed to be more important in Study 3 as well as in human perception of age (see Section 2.6.6)? A careful look at the CARTs from Study 4, which performed better when using prosodic features, reveals one striking detail: these CARTs had the smallest mean age errors for spectral features during the training (building) of the CARTs. There seems to be no immediate plausible explanation for this, except possibly methodological errors, implying that more studies with a different material (preferably longer than single words) and acoustic analysis as well as other machine learning techniques are needed to answer the question as to whether spectral or prosodic

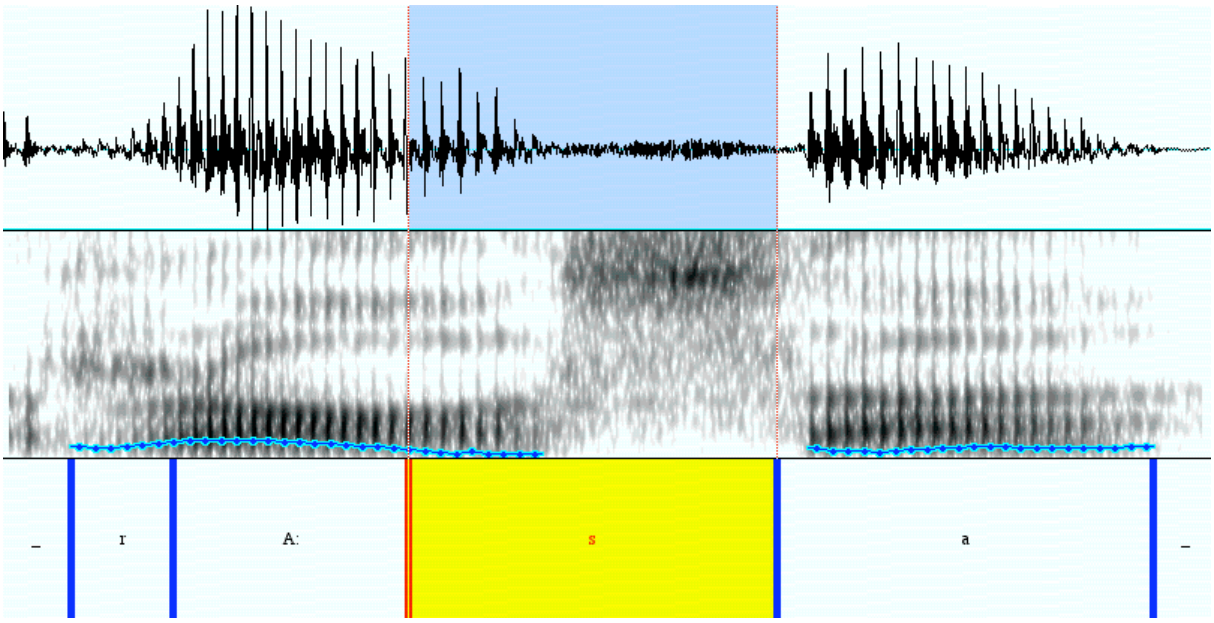


Figure 3.8: Example of an automatic alignment error where the segment identified as [s] is in fact part vowel, part fricative

features are more important in automatic age estimation.

3.7 Summary

Automatic speaker recognition (ASR) concerns the identification of speaker-specific characteristics (e.g. speaker age) from speech samples. In ASR, acoustic features are extracted from speech, and the characteristic is modelled (trained) – as well as recognised (tested) – using one or several pattern matching techniques. Common statistical methods are hidden Markov models (HMM), artificial neural networks (ANN) and cluster analysis with decision trees. One type of decision tree is classification and regression trees (CART), which uses both expert knowledge and statistical learning to build binary trees, based on a set of ordered yes/no questions about the features.

Little research on machine perception of speaker age exists today. The few existing previous studies have mainly extracted mel frequency cepstral coefficients (MFCC) or perturbation features from read or spontaneous utterances, and then used HMM or ANN to classify speakers into 2–4 age groups. Attempts to recognise perceived age (PA) – as estimated by a group of human listeners – have also been made.

In order to learn more about which acoustic-phonetic features are important in machine perception of speaker age, two studies were carried out using versions of the single Swedish disyllabic word *rasa* [ˈʁɑ̀ːsa] ('collapse') and the CART technique. The first study extracted 50 acoustic features (e.g. measures of F_0, duration and formant frequencies) from the phoneme

segments of 2,048 versions of *rasa*, produced by 428 speakers, equally many females and males. CART estimators of gender, age group (old/young) and direct chronological age were trained and tested. Results showed that the best CARTs made about 91% correct judgements for gender and about 72% for age group, while the correlation between chronological and estimated age was about 0.45. The estimators using the /ɑː/ and /s/ phonemes performed best, and spectral features (e.g. formant frequencies and HNR) seemed to be the most important ones in the estimation of direct age and age class, while F_0 was more important in gender estimation. Furthermore, different features seemed important in different phonemes. Estimation accuracy was compared with that of humans using the results of a previous listening test. It was found that although humans and CARTs used similar cues (features), the human listeners (mean error ± 8.89 years) were better judges of speaker age than the CART estimators (± 14.45 years).

The second study used an increased number of speakers (748) and acoustic features (78) – extracted for the four phonemes as well as for the whole word *rasa* – to construct separate estimators of direct age for female, male and all speakers. A total of 1,227 (409 × 3) CARTs were built for (1) 390 single features, (2) 13 feature groups (consisting of all features for one phonetic quality, e.g. F_1, B_1 and L_1) and (3) 5 larger feature groups of all prosodic, all resonance, all inverse filtered, all spectral (i.e. both resonance and inverse filtered) and all features. Results showed that F_0 and duration were the most important single features. Of the 13 feature groups, F0 and duration performed best for female speakers, while the formant groups of F2 and F3 were best for the male speakers. For the larger groups, the CART using all features was the best for female speakers, while the group with all prosodic features performed better for the male speakers. The best estimator of the second experiment (mean error ± 14.07 years) performed only marginally better than the one from the first study, perhaps owing to the larger dialectal variation of the extended material.

Although the performance of the automatic age estimators in the two studies did not reach the performance of human listeners, the results still provide information on which acoustic-phonetic features are the most important in machine perception of speaker age.

Chapter 4

Acoustic analysis of adult speaker age

4.1 Introduction

Age-related acoustic variation in temporal as well as in laryngeally and supralaryngeally conditioned aspects of speech have been well documented. However, the relationships among the numerous acoustic correlates of speaker age appears to be rather complex, and it is influenced by several factors. For instance, differences have been reported between correlates of female and male age, between speakers of good and poor physiological condition, between chronological age (i.e. the age of a speaker as measured in time from birth) and perceived age (i.e. the mean age of a speaker as estimated by a group of listeners), and also between different speech sample types (e.g. sustained vowels and read or spontaneous speech).

This chapter presents an overview of several known acoustic correlates of adult speaker age as well as brief reviews of some previous studies. The major part of the chapter describes a study where numerous acoustic features were extracted automatically from a large speech corpus. All features were analysed and compared in search of the most important correlates of adult speaker age.

4.2 Acoustic correlates of adult speaker age

There are several comprehensive overviews of acoustic correlates of age. For instance, Hollien (1987) has summarised research up till 1987, and Linville (2000, 2001) has provided excellent reviews of known acoustic aspects of the ageing voice. Based on these sources as well as on several other studies, this section gives an overview of the acoustic features usually related to speaker age. Furthermore, in an attempt to clarify which features have been found

to be important age correlates, some of the reported acoustic variation with increased age is summarised in Table 4.1. Variation with chronological age (CA) as well as with perceived age (PA) in women and men is described.

4.2.1 General variation

Old women and men alike demonstrate a general higher intra-subject as well as inter-subject variation of acoustic features when compared with young speakers. For example, increased variation has been found in some F_0 measures, as well as in speech rate (e.g. phoneme duration and VOT), vocal intensity, jitter, shimmer and HNR (Morris and Brown, 1994b; Decoster, 1998; Linville, 2001).

More age-related differences have been found for male than female speakers (Higgins and Saxman, 1991), and higher correlations of acoustic features with PA than with CA have generally been observed (Brückl and Sendlmeier, 2003). Moreover, correlations seem to vary with speech sample type (Brückl and Sendlmeier, 2003).

4.2.2 Speech rate

Temporal – static as well as dynamic – aspects of speech are strongly affected by the age of the speaker. The speech rate is linked to segment (syllable, phoneme, sub-phoneme, etc.) duration, to the number of speech segments per time unit and also to pause duration and frequency. A large number of studies have found a 20–25% decrease with older CA in speaking and reading rates. Increases have been found in consonant, vowel and sub-phonemic (prevoicing, plosive closure and release, vowel transition) durations as well as in pause duration and frequency (Benjamin, 1982; Oyer and Deal, 1985; Hollien, 1987; Morris and Brown, 1987; Brown et al., 1989; Shipp et al., 1992; Amerman and Parnell, 1992; Slawinski, 1994; Linville, 2001; Brückl and Sendlmeier, 2003). Women often demonstrate a smaller decrease in speech rate with older CA than men, or none at all (Hoit et al., 1989). Slower speech rates, a larger number of breaths and longer pause durations have been related to old male and female PA (Brückl and Sendlmeier, 2003).

The results for one sub-phonemic segment, namely voice onset time (VOT), are rather confusing. Some studies have found elderly (CA) women and men to exhibit shorter overall VOTs than younger people (Benjamin, 1982; Morris and Brown, 1987; Stölten and Engstrand, 2002). However, increased VOT with older male CA has also been observed (Decoster, 1998). Other researchers have reported only subtle differences and increased variation with advancing age (Petrosino et al., 1993; Linville, 2001). It has also been suggested that age-related differences in VOT appear only in certain phonetic

Table 4.1: Some reported acoustic variation with increased chronological age (CA) and perceptual age (PA) in female and male adult speakers (*decr.*: decrease, *dur.*: duration, *flat.*: flatter, *freq.*: frequency, *incr.*: increase, *int.*: intensity, *no*: no change, *sp.*: spectral, *steep.*: steeper). Please refer to the text for details

Group	*Feature*	*Variation with increasing adult age*			
		Female		*Male*	
		CA	*PA*	*CA*	*PA*
general	variation	incr.		incr.	
	overall changes	few	more	many	more
speech rate	syllables/second	decr. or no		decr.	decr.
	utterance dur.	incr.	incr.	incr.	incr.
	phoneme dur.	incr.			
	VOT	decr. or no		incr. or decr.	
	pause freq. & dur.	incr.	incr.	incr.	incr.
intensity	mean int.	no		incr.	
	max. int. range	decr.		decr.	
	amplitude SD	incr. or no	incr. or no	incr.	incr.
F_0	mean F_0	first no, then decr., incr. or no	decr.	first decr., then incr.	first decr., then incr.
	F_0 range	first incr., then decr.	incr. or no	first incr., then decr. or no	
	F_0 SD	incr. or no	incr. or no	incr. or no	incr.
tremor	vocal tremor	incr. or no	incr.	no	
jitter & shimmer	jitter	incr. or no	incr. or no	incr. or no	
	shimmer	incr.	incr. or no	incr.	
sp. noise	HNR	decr. or no		no	
	NHR	incr. or no	incr. or no	incr.	
sp. energy distribution	sp. tilt	flat.		steep. or flat.	
	sp. emphasis	no		no	
	sp. tilt (LTAS)	steep.		flat.	
resonance	F_1 and F_2	decr.	decr.	decr.	decr.
	F_1–F_3(LTAS)	decr.	no	decr.	decr.

contexts (Neiman et al., 1983).

4.2.3 Intensity

Conversational speech intensity appears to remain stable with increased CA for females, but it is higher for men after age 70, even for speakers without hearing loss (Ryan, 1972; Hollien, 1987; Morris and Brown, 1994a; Linville, 2001). The habitual intensity range in vowels is likely to increase with advancing female and male age (Xue and Deliyski, 2001). However, the maximum vowel intensity range seems to decrease in both women and men, while

minimum intensity levels increase for women (to the author's knowledge, no studies have been made concerning men) with advancing CA (Morris and Brown, 1987; Linville, 2001).

4.2.4 Fundamental frequency (F_0)

F_0 patterns in speech related to CA are different for women and men, as shown in Figure 4.1. Female F_0 remains fairly constant until menopause, when a drop (of about 10–15 Hz) usually occurs. F_0 then remains stable into old age, but may also rise or lower slightly (Hollien and Shipp, 1972; Kitzing, 1979; Higgins and Saxman, 1991; Traunmüller and Eriksson, 1995; Linville, 2000; Xue and Deliyski, 2001). A lower F_0 is also associated with older female PA (Linville, 2001; Brückl and Sendlmeier, 2003). In males, F_0 lowers slightly (by about 10 Hz) from young adulthood to middle CA, but then rises considerably (about 35 Hz) with old CA (Mysak, 1959; Kitzing, 1979; Linville, 2001). Higher F_0 has been reported to be a cue to old male PA (Shipp et al., 1992; Linville, 2001). However, there are also studies which have failed to find correlations between mean F_0 and CA in men (Ramig and Ringel, 1983). Moreover, the way changes in F_0 relate to perceptual cues is not in line with the above findings. For instance, listeners have reported lower male F_0 to be a cue to older age (Linville, 2001).

F_0 range

Maximum phonational frequency range – i.e. the complete range of frequencies which a speaker can produce, from the lowest (without creak) to the highest tone (including falsetto) – expands in the lower end following menopause in females, but is restricted in both the upper and lower ends later in life (Linville, 1987, 2001). Contradictory findings suggest that men either undergo similar changes in F_0 range as women (Ptacek et al., 1966; Linville, 2001), or that old and young males do not differ in F_0 range unless physiological condition and state of health are taken into account (Ramig and Ringel, 1983; Ringel and Chodzko-Zajko, 1987). A larger habitual F_0 range has been observed for the vowel /a/ in both women and men of old CA (Xue and Deliyski, 2001).

4.2.5 Variation in F_0 and amplitude

Fundamental frequency and amplitude instability and variation are related to various voice qualities. Jitter and shimmer (see p. 85) are often connected with harshness, hoarseness or vocal roughness, while increases in the more gross F_0 variation, as measured in standard deviation (F_0 SD), may cause

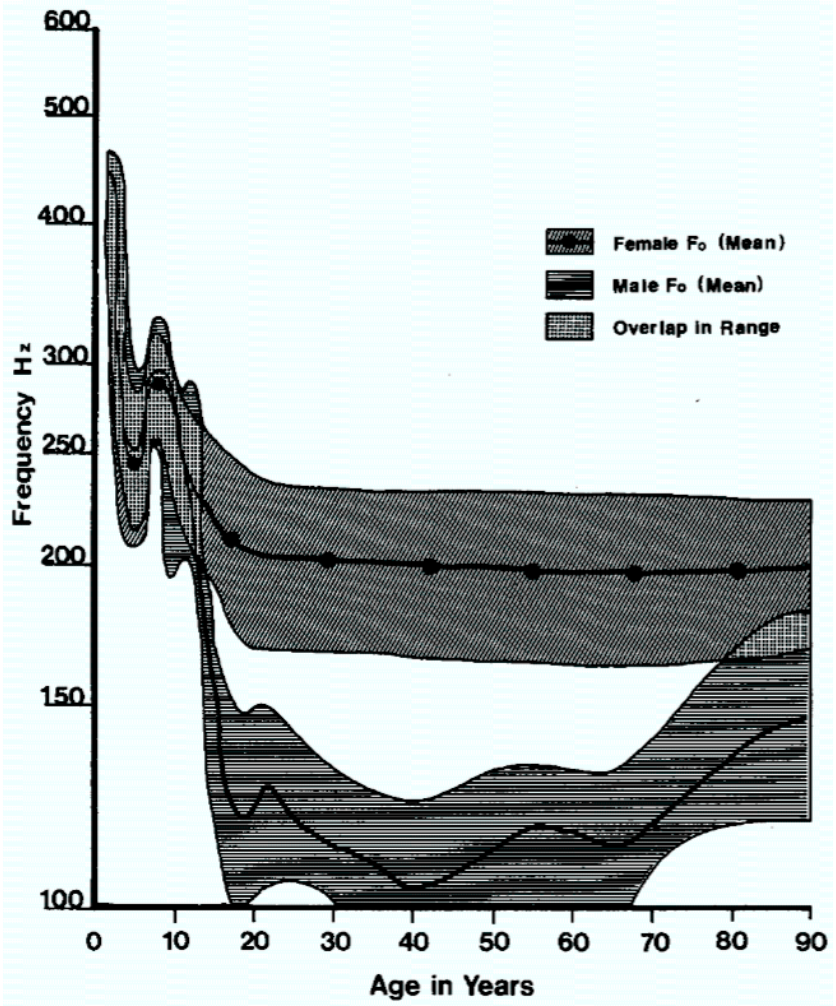


Figure 4.1: Speaking F_0 and its standard deviation as a function of speaker age (1–90 years) for female and male speakers (source: Brown et al., 1991)

vocal tremor or a "wobbling" voice quality (Hollien, 1987; Linville, 1996, 2001).

F_0 SD and amplitude SD

Higher F_0 SD (with greater variation for females) has been found in both men and women with advancing CA and PA (Hollien, 1987; Linville, 1987; Shipp et al., 1992; Linville, 2001; Xue and Deliyski, 2001), but sometimes only a minor correlation has been reported, or none at all (Ramig and Ringel, 1983; Brückl and Sendlmeier, 2003). Substantial increases in fundamental amplitude standard deviation (Amp SD) have been demonstrated in older men and women, and have been associated with both CA and PA (Orlikoff, 1990b; Xue and Deliyski, 2001). However, Brückl and Sendlmeier (2003) found a strong positive correlation with female CA and PA only in spontaneous speech but almost none in sustained vowels or read speech.

Jitter and shimmer

Jitter and shimmer are defined as period-to-period fluctuations in vocal fold frequency and amplitude, as shown in Figure 4.2, and they are considered to be correlates of rough or hoarse voice quality. These features have often been analysed in acoustic studies of age using a number of measures with varying results. Although sometimes no correlation with age has been found for jitter (Ringel and Chodzko-Zajko, 1987; Brown et al., 1989; Ferrand, 2002; Shuey et al., 2003), other researchers have reported increased jitter levels for older female and male CA (but not PA) (Linville, 1987; Orlikoff, 1990b; Decoster, 1998; Xue and Deliyski, 2001; Müller, 2005). However, higher and more variable jitter values seem to be more related to physiological health than to age (Ramig and Ringel, 1983; Orlikoff, 1990b; Linville, 2001; Brückl and Sendlmeier, 2003).

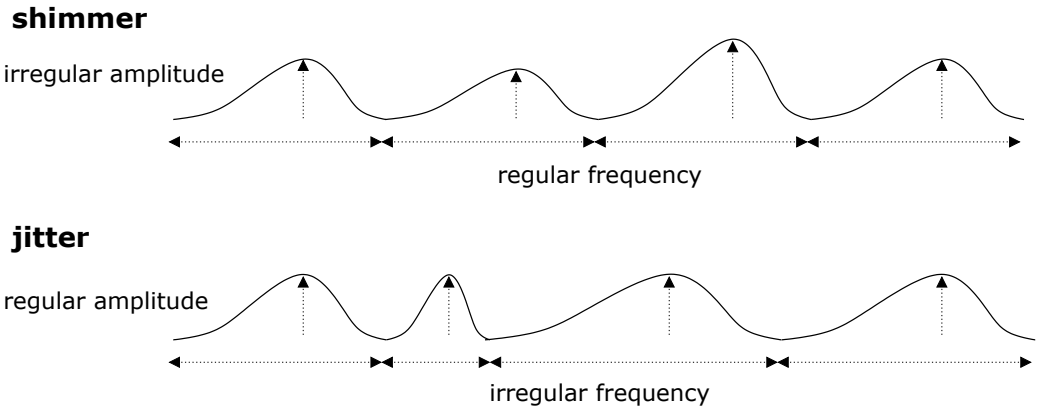


Figure 4.2: Irregularities (microvariations) in vocal fold movements can be measured as shimmer (variation in amplitude) and jitter (variation in frequency) (after Müller, 2005)

Higher shimmer levels have been found for older female CA and PA as well as for older male CA (independently of health) (Ramig and Ringel, 1983; Ringel and Chodzko-Zajko, 1987; Orlikoff, 1990b; Decoster, 1998; Xue and Deliyski, 2001; Müller, 2005). However, some studies have found shimmer to correlate strongly with CA and PA only in spontaneous speech samples (but not in read speech or in prolonged vowels) (Brückl and Sendlmeier, 2003). Other studies have observed correlations of shimmer and CA in sustained vowels, but only when 80-year-olds were compared with younger age groups (Shuey et al., 2003).

Linville (2001) concludes that it is impossible to draw any firm conclusions as to the effect of ageing on jitter and shimmer since several factors, including sound pressure level, mean F_0, analysis system differences and individual health and fitness variables, appear to have a strong effect on these measures, especially in women.

4.2.6 Other voice measures

Spectral energy distribution

Spectral tilt (ST), spectral emphasis (SE) and spectral balance (SB) are all measures of the relative energy levels in different frequency bands of the spectrum (Campbell, 1995; Heldner, 2001). ST usually represents the slope – i.e. the difference between the energy levels of two different frequency bands – of the source (inverse filtered) spectrum in dB per octave. SE is a measure of the relative energy levels in the higher frequency bands, while SB is often measured in four contiguous frequency bands. The three measures have sometimes been defined differently (Sluijter and van Heuven, 1996; Traunmüller, 1997; Heldner, 2001; O'Leidhin and Murphy, 2005).

ST has been observed to flatten (i.e. the energy in the frequency band 4–5 kHz increased with female and male CA) in some vowels, while a longitudinal study found a steeper spectral tilt in old men compared with the same men when young (Decoster and Debruyne, 1997). SE has been found not to correlate significantly with CA (Traunmüller, 1997), while no study of how SB correlates with speaker age is known to the author.

The age-related variation of the energy distribution in long-term average spectra (LTAS) has also been studied to some extent. An LTAS is an averaged spectrum of all voiced sounds across a relatively long speech sample. Elderly women have been observed to have higher spectral amplitudes at 320, 6080, 6240, 6400, 6560 and 6720 Hz but lower levels at 3040 and 3200 Hz than young women, and a tendency for older women to have higher levels at 160 Hz has been found as well (Linville, 2002). Differences in spectral amplitude have been found between old and young men, too. Old males had higher spectral amplitude values at 160 Hz and lower levels at 1600 Hz than young males (Linville, 2002). A strong spectral attenuation of high frequencies has also been observed in LTAS at older CA and PA in males, but not in females (Winkler et al., 2003).

Spectral noise

Spectral noise is defined as the unmodulated aperiodic energy in vowel spectra (Linville, 2001). It is considered an acoustic correlate of breathy and harsh or hoarse voice quality (McAllister et al., 1998; Kreiman and Gerratt, 2005), and has been analysed using various methods. Visual analysis of spectral noise in spectrograms has shown that this feature is much more strongly correlated with physiological condition than with CA (Ramig, 1983).

The harmonics-to-noise ratio (HNR) is a measure that quantifies the amount of additive noise in the voice signal, and it can be calculated in several ways (de Krom, 1993; Boersma, 1993). The ratio reflects the dominance of

the periodic level over the aperiodic one, as quantified in dB. HNR has sometimes been reported to decrease with older female CA (Ferrand, 2002), or to increase with younger male CA (Wang and Huang, 2004), while other researchers have failed to find strong correlations with CA in females (Wang and Huang, 2004) or both genders (Schötz, 2006b). No studies exist (to the author's knowledge) of HNR in relation to PA.

Other measures of spectral noise used in acoustic studies of speaker age include the parameters VTI, SPI and NHR of the commercial voice quality analysis software Multi-Dimensional Voice Program (MDVP, see e.g. Deliyski and Gress, 1998). Voice turbulence index (VTI) is a measure of the relative energy level in high-frequency noise. It is calculated as the average ratio of the inharmonic spectral energy in the 2.8–5.8 kHz range to the harmonic spectral energy in the 0.07–4.5 kHz range. Soft phonation index (SPI) measures the relative energy in low-frequency noise, calculated as the average ratio of the lower (0.07–1.6 kHz) to the higher (1.6–4.5 kHz) frequency harmonic energy. The noise-to-harmonics ratio (NHR) is the average ratio between noise in the frequency band 1.5–4.5 kHz and the harmonic energy in the frequency band 0.07–4.5 kHz; it is sometimes referred to as a low-frequency harmonics-to-noise ratio (Pereira Jotz et al., 2002). Increased values for all three features in women and men of older CA have been reported (Xue and Deliyski, 2001). Other researchers have failed to find correlations of these features with female and male CA (Shuey et al., 2004), though weak (NHR but not VTI) and strong (SPI) positive correlations with female PA have also been observed (Brückl and Sendlmeier, 2003).

Vocal tremor

Vocal tremor can be measured using the MDVP parameters FTRI (intensity of the strongest frequency modulation) and ATRI (intensity of the strongest amplitude modulation). FTRI (but not ATRI) has been found to increase with both female CA and PA in vowels, but not in read or spontaneous speech (Brückl and Sendlmeier, 2003).

4.2.7 Resonance measures

Research has revealed that age-related changes in the supralaryngeal structures provide acoustic cues to adult speaker age (Ptacek and Sander, 1966; Shipp and Hollien, 1969; Jacques and Rastatter, 1990). However, only relatively few studies have investigated age-related changes in the vocal tract resonance features.

Formant frequencies

Formant frequencies in vowels have been reported to lower with female and male CA and PA owing to increased vocal tract length (Endres et al., 1971; Linville and Fisher, 1985). There also seems to be a trend towards vowel centralisation (or reduction) for old CA (Rastatter and Jacques, 1990; Rastatter et al., 1997). However, it appears that some old speakers centralise more than others, suggesting an increase in formant frequency variation across speakers of old CA (Linville, 2001). For PA, formant information seems to lose its significance when F_0 information is present (Linville, 1996).

Energy peaks in long-term average spectra (LTAS), corresponding to the average formant frequencies across all vowels in a speech sample, have been studied by Linville and Rens (2001). They found a significant lowering of peaks 1, 2 and 3 (corresponding to F_1–F_3) with old female CA, and a significantly lower peak 1 (F_1) in old male CA. Moreover, the age-related lowering of peaks was greater in females than in males.

To sum up this chapter so far, previous research has found numerous acoustic correlates of chronological and perceptual speaker age. Some features, such as measures of F_0 and speech rate, have been found to be more important than others and have thus been investigated to a larger extent. However, the relative importance of the different age correlates has not yet been established. In addition, there are also a number of factors – partly the same ones claimed in Section 2.3 to affect human age perception – may also influence acoustic analysis of age correlates. Some of these factors are described in the following section.

4.2.8 Factors which may influence acoustic analysis of speaker age

In acoustic studies of speaker age, several factors (besides age) may affect the analysis outcome. These factors are related to the material and the methods used, and may contribute to the divergent and sometimes even contradictory results found in different studies. This section offers a brief overview of some of the factors which may influence analysis results. Some additional factors are similar to the ones affecting human perception of age; they are described in Section 2.3.

Speaker-related factors

Speaker-related factors include physical (anatomical and physiological) attributes such as gender, race, weight, health and physiological condition.

Women and men differ in several vocal characteristics. Some can be explained by anatomical differences while others, such as the paralinguistic use of breathy voice quality, appear to be learned behaviours (Klatt and Klatt, 1990). Differences in body physiology, vocal training and medical condition may also affect the age-related variation in speech (Ramig and Ringel, 1983; Orlikoff, 1990a; Sataloff et al., 1997; González, 2004), including effects of medication (Decoster and Debruyne, 1997) and cigarette smoking (Braun and Rietveld, 1995). For instance, smokers generally exhibit lower F_0 than non-smokers (González and Carpi, 2004), while professional sopranos and tenors have a higher F_0 than age-matched non-singers (Brown et al., 1991). Furthermore, age-related differences in habitual F_0 seem less prominent or even absent in singers and other voice professionals (Sataloff et al., 1997).

Cultural, social and psychological factors, including speaker language, dialect, sociolect, emotional state and attitude, may influence and even mask age-related acoustic variation. For instance, there are language-related, dialectal and attitudinal differences in habitual F_0, HNR and shimmer levels (Traunmüller and Eriksson, 1995; Wagner and Braun, 2003). Moreover, consideration must also be given to the fact that voice settings are more or less subject to swings in fashion (Lindblad, 1992), and that the pronunciation of a language is constantly changing. Young individuals often wish to speak differently from their parents (Roach, 2001). One example is the increased use of the more open allophones [æː] and [œː] of the /ɛ/ and /ø/ phonemes in Swedish (Andersson, 2006). Another example is the growing use of the glottal stop in British English (Roach, 2001).

Speech-material-related factors

Speech-material-related factors include the number and age distribution of the speakers and the duration and speech type (and number of speech types) of the speech samples analysed. Of course, fewer speakers will yield less reliable results, as will an unbalanced (for age) speech corpus. Valid measurements of some features are obviously obtained more easily from certain speech types. For instance, formant frequencies are more reliably measured in sustained vowels than in connected speech, and calculations of the average number of syllables per second are more reliable in longer speech samples. Moreover, studies which have used more than one speech type have sometimes found contradictory results for different speech types. One example is Brückl and Sendlmeier (2003), who found that vocal tremor correlated with age in sustained vowels, but not in read or spontaneous speech.

Methodological factors

Methodological factors, such as differences in recording and analysis equipment and techniques, may strongly influence the outcome of acoustic analyses. One example concerns the vocal effort made by speakers to adapt to the distance to a listener or a microphone, which may affect speech rate, voice quality, measures of F_0 and even some formant frequencies (Traunmüller and Eriksson, 2000). Different measurement techniques could also be one reason why, for instance, it has not yet been possible to draw any firm conclusions as to the effect of ageing on jitter and shimmer (Linville, 2001).

Another major methodological factor in acoustic studies of speaker age is whether the findings are related to chronological or perceived age. In automatic age recognition applications, the goal is in many cases to identify speakers' actual CA, and not the mean PA as estimated by a group of listeners. However, if only CA is considered in an acoustic study, no knowledge about the relative importance of the correlates to listeners will be gained (Ryan and Burk, 1974). On the other hand, when the acoustic correlates of PA are examined, the age judgements of a group of listeners – often quite small – will have to be trusted. Since PA is a subjective measure, results may not be reliable, as listener characteristics (gender, age, etc.) affect the age estimates (see Section 2.3.2). Thus, the purpose of each study or application will have to determine whether CA or PA is chosen as the frame of reference.

A connected question is whether we should use archival recordings in combination with recent ones of the same speakers (longitudinal studies) or speech samples from different speakers recorded close in time (cross-sectional studies). Although it may be tempting to use longitudinal data because of the invariant speaker-specific parameters, several aspects which may affect the results should be regarded. Differences in recording equipment and technical sound quality may yield unreliable results. Moreover, voice communication habits may change over time, one example being that Australian women aged 18–25 years recorded in 1993 had significantly lower F_0 levels than women of the same age recorded in 1945 (Pemberton et al., 1998). Another example concerns VOT and F_0 SD. Several cross-sectional studies have reported that VOT decreased and F_0 SD increased in males with advancing age. However, in a longitudinal study of male speakers recorded twice over a period of 30 years, Decoster and Debruyne (2000) found the opposite results.

In spite of the numerous factors which may affect acoustic analysis, different studies have agreed on several acoustic correlates of speaker age. However, many experiments have differed in the number and choice of speakers and acoustic features, as well as in speech material and method. Some studies have reduced the effect of certain factors by controlling variables or by using

a large material. The next section presents some of the previous acoustic research on speaker age.

4.3 Previous related studies

The first acoustic studies describing general aspects of speaker age emerged in the late 1950s and in the 1960s, e.g. Mysak (1959) and Ptacek et al. (1966). It soon became apparent that numerous features vary as a function of age, and that features which correlated with CA did not always correlate with PA or with perceptual cues (i.e. age-related perceptual features mentioned by listeners). To clarify the relationship between acoustic features and speaker age, numerous studies have been, and are still being, carried out.

This section offers a review of some previous acoustic studies in order to give the reader an idea of the various features, speakers and speech samples used in different experiments as well as whether the analysis was related to CA or PA. In some cases when the articles give no explicit information about the language analysed, it is assumed to be English (American or British). The studies reviewed are summarised in Table 4.2.

1. Endres et al. (1971) measured F_0 and F_1–F_4 in spectrograms of eight vowels and one nasal consonant in connected speech samples from two female and four male German speakers, recorded several times over a period of 13–15 years. They found that F_1–F_4 and mean F_0 decreased with increasing CA. F_0 variation decreased as well, indicating a loss in ability to vary F_0 with increased age.

2. Hollien and Shipp (1972) measured mean F_0 in read speech produced by 175 healthy male speakers aged 20–89 years (25 speakers in each decade). They found that mean F_0 decreased progressively from age 20 to 40, but increased after the age of 60.

3. Ramig and Ringel (1983) and Ramig (1983) studied the relationship between some acoustic voice characteristics and age-related physiological changes. Forty-eight men from three age groups (25–35, 45–55 and 65–75) and two levels of physiological condition (good, poor) produced three sustained vowels as well as read and spontaneous speech. Acoustic and statistical analyses revealed significant correlations with CA for shimmer, but not for jitter, maximum phonation range or mean F_0. Correlations with physiological condition were reported for jitter, shimmer and maximum F_0 range, but not for mean F_0. Ramig (1983) also examined vowel spectral noise using the same material. Four speech scientists visually classified spectrograms

Table 4.2: Data from 14 acoustic studies of speaker age (*CA*: chronological age, *PA*: perceived age, *sp.*: spectral, *jit.*: jitter, *shim.*: shimmer, *param.*: parameters, *(?)*: not stated explicitly in the source article)

No	*Authors*	*Year*	*Speech samples*		*Speakers*			*Analysis*	
			type	*duration*	*N*	*gender*	*age*	*features*	*related to*
1	Endress et al.	1971	vowels and /n/	(from connected speech)	6	♀ ♂	♀: 29–43 (?) ♂: 42–88 (?)	F_0, F_1–F_4	CA
2	Hollien & Shipp	1972	read	98 words (?)	175	♂	20–89	F_0	CA
3	Ramig & Ringel; Ramig	1983	vowels read spontaneous	(?) (?) ≈ 30 s	48	♂	25–35, 45–55, 65–75	F_0 (mean, max range), jit., shim. vowel sp. noise	CA
4	Oyer & Deal	1985	read	132 words	24	♀ ♂	44–82	speech rate, pauses	CA
5	Linville & Fisher; Linville	1985; 1987	phonated and whispered /æ/	≥ 1 s	75	♀	25–35, 45–55, 70–80	F_0 (mean, SD), jit., F_1, F_2	CA & PA
6	Morris & Brown	1987	syllables /ɑ/	(?) 5 s	50	♀	20–35, 75+	phoneme duration, VOT, intensity	CA
7	Shipp et al.	1992	read	(?) (22 words)	30	♂	27–35, 53–57, 75–85 (PA)	F_0, speech rate, breath pauses	PA
8	Decoster & Debruyne	1997	vowels syllables read	3 s (?) 180 s	265 20	♀ ♂ ♂	21–29, 60–99 33 & 63 (mean)	VOT, sp. measures	CA
9	Xue & Deliyski	2001	/a/	1 s	44	♀ ♂	young, middle-aged, 70–80	15 voice param. (F_0, jit., shim., NHR, tremor, etc.)	CA
10	Linville & Rens	2001	read	(?)	80	♀ ♂	19–24, 62–79	F_1–F_3 in LTAS source measures in LTAS	CA
11	Linville	2002		(98 words)					
12	Ferrand	2002	/ɑ/	2 s	42	♀	21–34, 40–63, 70–90	F_0, jit., HNR	CA
13	Brückl & Sendlmeier	2003	vowels read spontaneous	≈ 2.2 s ≈ 10 s ≈ 10–15 s	56	♀	20–87 (mostly middle-aged)	33 voice param. (F_0, speech rate, jit., shim., tremor, etc.)	CA & PA
14	Müller	2005	read spontaneous	(?) (utterances)	1,164	♀ ♂	10–97 (mostly aged 10–20 & 60–97)	F_0, speech rate, pauses, intensity, jit., shim., HNR	CA

of 48 sustained vowels, produced by equally many males in good and bad physiological condition from three age groups, with regard to the amount of spectral noise. After that, ten typical spectrograms were shown to 30 speech pathology students, who rank-ordered the spectrograms by the amount of visual spectral noise. Results showed that spectral noise was much more strongly correlated with physiological condition than with chronological age.

4. Oyer and Deal (1985) investigated the extent to which subjects of different CA could be divided into age groups based on reading rates and pauses. Readings of a 132-word paragraph by 24 male and female speakers (CA: 44–82 years) were analysed for total reading time and reading time within a passage as well as frequency, average duration and total time of orthographic and non-orthographic pauses. Results showed that 75% of the subjects older than 61 years had significantly longer reading times than the median value for all subjects. Noticeable increases in pause time and pause frequency from the youngest age group (42–60) to the oldest one (61–82) were also observed.

5. Linville and Fisher (1985) made acoustic analyses of phonated and whispered /æ/ vowels produced by 75 women from three age groups (25–35, 45–55 and 70–80 years) for jitter, mean F_0 and F_0 standard deviation (F_0 SD) as well as F_1 and F_2. These features were related to both PA (estimated by 15 young and 16 old female listeners) and CA. Results showed that (1) jitter was unrelated to both CA and PA, (2) mean F_0 lowered with advancing CA and PA, (3) F_0 SD was the most successful cue to the three age groups for both CA and PA (old speakers had higher and more variable values), (4) F_1 and F_2 in the phonated stimuli were not important for PA, but decreased (mainly F_1) with older CA, and (5) F_1 (but not F_2) in the whispered stimuli lowered with increased CA and PA. The study is also described in Linville (1987).

6. Morris and Brown (1987) studied phoneme duration and voice onset time (VOT) in CV, VCV and VC syllables, as well as vocal intensity in /a/ produced at three intensity levels by two CA groups of healthy females (25 aged 20–35, 25 aged ≥ 75). The young women showed higher maximum and lower minimum intensities than the old ones, but no significant differences were found at the conversational levels. The old women produced significantly longer consonant and vowel durations, and they exhibited shorter overall VOTs than the younger women.

7. Shipp et al. (1992) examined F_0, speech rate (total and segment[1] duration as well as number of syllables per second) and breath management (number of breaths and breath pause duration) as correlates of PA. The material consisted of a 22-word sentence read by 30 male speakers, who had in an earlier study been classified by listeners into three age groups (27–35, 53–57 and 75–85). Results showed that old speakers exhibited the highest mean F_0, followed by the young and the middle-aged ones. F_0 SD was highest for the old, lowest for the young and intermediate for the middle-aged speakers. No significant age-related differences were found in the percentage of utterance duration of voiced, unvoiced and silent segments. However, the number of intra-sentence breaths as well as pause duration increased with older PA. As for speech rate, the young group showed the shortest total duration and the largest number of syllables per second, while the middle-aged group had intermediate values and the old group had the longest total duration as well as the smallest number of syllables per second.

8. Decoster and Debruyne (1997) carried out an acoustic study of three spectral parameters in sustained [a] and [i], and also measured VOT in double syllables [kaˈka] and [paˈpa] from 205 elderly (60–95) and 60 young (21–29) Dutch-speaking females and males. For the vowels, two significant age-related differences were found: a stronger second harmonic relative to the first harmonic for elderly women in both vowels, and a relatively higher energy in the frequency band 4–5 kHz of [i] for all elderly speakers. For VOT, the only significant age-related difference was observed in [pa] for 60–69-year-old women compared with 90–99-year-old women (however, no indication of the nature of the difference is given).

Decoster and Debruyne also made a longitudinal study of 20 male radio speakers (mean CA: 63), who were asked to read the same texts as they had read 30 years previously. Comparison of the two recordings showed that the more recent recordings (with older speakers) had steeper spectral slopes between the frequency bands < 1 and 2–4 kHz as well as much longer VOTs.

9. Xue and Deliyski (2001) used the Multi-Dimensional Voice Program (MDVP, see p. 87) to extract 15 acoustic voice parameters, including mean F_0 and range, jitter, shimmer and noise-to-harmonics ratio (NHR), from sustained /a/ vowels produced by 21 male and 23 female elderly speakers (70–80 years old). The results were compared with those from a study of young and middle-aged (no information on the age of the speakers is given) adults (Deliyski and Gress, 1998). They found that, as a group, old female and male speakers had significantly lower F_0 than young and middle-aged

[1] Voiced and unvoiced segments, as well as the implosion phase in voiceless plosives.

speakers. The other 14 parameters showed significantly higher measurements for the old group than for the young and middle-aged groups.

10. Linville and Rens (2001) analysed three resonance peaks corresponding to F_1–F_3 in long-term average spectra (LTAS) of 40 male and 40 female healthy young and elderly (CA: 19–24, 62–79) speakers. Results showed that all three peaks lowered with advancing CA for both genders.

11. Linville (2002) studied spectral source measures in long-term average spectra (LTAS) of 40 male and 40 female healthy young and elderly (CA: 19–24, 62–79) speakers. She found that young women had higher spectral amplitudes around 3 kHz and tended to have flatter spectral tilts than elderly women, who instead had higher amplitudes around F_0 and above 6 kHz. Elderly men also had increased amplitudes around F_0, but unlike women they tended to have flatter spectral tilts than young men as well.

12. Ferrand (2002) examined harmonics-to-noise ratio (HNR), jitter and F_0 in sustained /a/ vowels, produced at comfortable F_0 and intensity levels by 42 women equally divided into three age groups (CA: 21–34, 40–63, 70–90). She found that while HNR and F_0 decreased significantly with advancing female CA, jitter demonstrated no significant difference between the three age groups.

13. Brückl (2002) and Brückl and Sendlmeier (2003) automatically extracted 33 voice parameters (using the voice program MDVP) from sustained vowels as well as from read and spontaneous speech, produced by 56 mostly middle-aged female speakers (CA: 20–87, PA estimated by 6 female and 9 male listeners). The parameters, which included several measures of F_0, jitter, shimmer, spectral energy distribution, tremor and speech rate, were analysed and compared with both CA and PA. Results revealed that correlations varied with speech sample type, and were also generally stronger with PA than with CA. F_0 lowered with both CA and PA, while F_0 variation increased slightly with CA and PA in some vowels. Shimmer displayed the strongest correlation of the study with both CA (0.551) and PA (0.602) in spontaneous speech, but almost no relation with age was found in read speech or vowels. Spectral energy distribution did not correlate with CA or PA. Vocal tremor increased significantly with both CA and PA in sustained vowels. Speech rate decreased with both CA and PA in vowels and read speech, but not in spontaneous speech.

14. Müller (2005, 2006) used several automatic methods and analysis tools (including Praat) to extract measures of F_0, jitter, shimmer, HNR,

articulation rate, pauses (e.g. number of pauses and pause duration) and intensity in read and spontaneous utterances. His material comprised three speech corpora (two German, one English) with 507 female and 657 male speakers, divided into four age classes according to their CA (children: 10–12, teenagers: 13–19, adults: 20–65, seniors: 60–97). The majority of the speakers were children and seniors. Results showed that F_0 decreased in women but increased in men with advancing adult age. Seniors displayed higher jitter and shimmer values than younger speakers of both genders. Children and teenagers of both genders had higher HNR values than adults and seniors. Articulation rate increased in young adults, but decreased in seniors. Adults demonstrated the smallest number of pauses, seniors the largest number, and children and teenagers were in between. Pause duration was equal for all age classes except for seniors of both genders, who demonstrated longer pause durations.

As has been shown, the numerous studies carried out so far have often differed in the number of speakers, the type of speech material, the features analysed and the method used, as well as in whether the results were related to CA or PA. Several studies agree on which acoustic features are important, but sometimes contradictory results have been found as well. Moreover, no study (to the author's knowledge) has yet attempted to make direct comparisons of a large number of acoustic features in order to identify the most important acoustic correlates of speaker age. The next section describes an acoustic study of speaker age where such an attempt is made.

4.4 Study 5: Acoustic correlates of speaker age

Previous studies have yielded much knowledge about acoustic cues to CA and PA. It has been demonstrated that variation in numerous acoustic features is related to both the chronological and the perceived age of an individual speaker. Although some features obviously seem more important than others, it is still unclear which (group of) features constitute the most significant acoustic age correlates. Some features, such as resonance measures, have not been studied as extensively as others. Moreover, less age-related research has been carried out using phonetic segments such as isolated words or consonants and vowels in connected speech than using sustained vowels or read and spontaneous utterances.

Lately, automatic methods have become more common in acoustic studies of age. For instance, computer programs have been used to extract features automatically, an approach which has several advantages. It is less time- and

labour-consuming than manual approaches, allowing researchers to analyse more features in a larger speech material. Moreover, automatic methods do not make "human factor" errors, e.g. measurement mistakes due to fatigue or boredom. The drawbacks of automatic methods include errors which are easier to avoid in manual analysis, for instance when identifying F_0 and formant frequencies in high-pitched or creaky speech segments (Klatt and Klatt, 1990). So far, automatic analysis has been used mainly for voice parameters and speech rate (Brückl and Sendlmeier, 2003; Müller, 2005; Winkler and Sendlmeier, 2006), while resonance measures need to be studied further.

In an attempt to answer some of the remaining questions regarding acoustic correlates of speaker age, an experiment was carried out where numerous acoustic features were automatically extracted from a large speech corpus with a wide age distribution. The features were analysed in order to identify and compare important acoustic correlates of female and male speaker age. This study is described in the following sections.

4.4.1 Purpose and aim

The purpose of the study was to obtain data on a large number of acoustic features in order to learn if and how they are related to speaker age, and to compare the age-related variation in the different features. Specifically, the study was designed to investigate features in isolated words, in stressed vowels and in voiceless fricatives and plosives. A second purpose was to use mainly automatic methods when extracting and analysing the features, in order to be able to cover more features and a larger material.

The experiments presented here aimed at identifying the relatively most important acoustic correlates of speaker age among a large number of features. The study was further motivated by an ambition to gain more knowledge about how automatic approaches can be used in acoustic-phonetic studies, thereby contributing somewhat to narrowing the gap between (traditional) acoustic phonetics and speech technology.

4.4.2 Questions and hypotheses

The research questions concerned acoustic feature variation with advancing speaker age and also automatic approaches to acoustic analysis:

1. What age-related differences in acoustic features can be identified in female and male speakers?

2. Which are the most important acoustic correlates of speaker age?

3. Would it be possible to verify the results of previous studies in this area of research using mainly automatic techniques in the acoustic analysis?

Based on the observations reported in the studies reviewed, which are summarised in Table 4.1, the following hypotheses were made for how the various features examined in this study were expected to vary with increasing speaker age:

1. Speech rate will generally decrease with advancing age. The number of syllables and phonemes per second will either decrease (both genders) or remain the same (females). Segment (word and phoneme) duration will increase for both genders.

2. Habitual intensity range will increase for both genders. Intensity standard deviation in vowels (Amp SD) will increase or remain constant in women but increase in men.

3. F_0 will display different patterns for female and male speakers. In females, F_0 will remain stable until around the age of 50 (menopause), when a drop will occur, followed by either an increase, a decrease or no change. In men, F_0 will decrease until around middle age (somewhere between 40 and 60 years), when an increase will follow until old age. F_0 range will first increase for both women and men, and then decrease (women) or decrease or remain stable (men). F_0 standard deviation (F_0 SD) will either increase or display no change in both genders.

4. Jitter and shimmer will either increase or remain stable in both women and men.

5. Spectral energy distribution (spectral tilt, see p. 86) will generally either increase or decrease. However, at the higher frequencies (spectral emphasis, see p. 86) there will be no change.

6. Spectral noise will increase in women, and either increase or remain stable in men.

7. Resonance measures in terms of formant frequencies will decrease in both female and male speakers.

Owing to the contradictory findings regarding VOT, it was hypothesised merely that VOT would vary with age, but no predictions were made as to whether it would increase or decrease. No hypothesis was made for average intensity, since the speech samples to be used in this study would be normalised for intensity. Furthermore, owing to the small amount of directly comparable normative data, no hypotheses were made for spectral energy distribution in LTAS.

4.4.3 Speech material

The most suitable material for this study would be an age-balanced speech corpus where every age from 20 to 90 years is represented by equally many female and male speakers. No such corpus exists today (to the author's knowledge), but the material finally chosen still represented speakers of both genders and most adult ages. It comprised speakers from the SweDia 2000 corpus (Bruce et al., 1999; see also p. 30) as well as several new recordings.

Speakers

In all, 259 female and 268 male speakers were chosen for the study. The goal was to analyse five female and five male speakers, preferably from the south of Sweden (to reduce dialectal variation), of each chronological age from 20 to 90 years, i.e. five female and five male speakers aged 20 years, the same number of speakers aged 21 years, and so on. At first, the SweDia 2000 corpus seemed like a good choice. However, since the aim of the SweDia project was to document dialectal variation of an old and a young generation, the corpus comprises mainly speakers from the two age groups 20–30 and 55–75 years. Additional recordings of speakers of the ages missing in the SweDia material were thus made by the author.

The material was compiled as follows: As many SweDia speakers with South Swedish dialects (*sydsvenska*) as possible were selected to fill the quotas of five speakers per age for each gender. Whenever fewer than five female or male speakers of a certain age were found, SweDia speakers from the other Swedish dialectal areas (Göta, Svea, Dala, Gotland and Finland Swedish) were chosen. However, many ages still had no or nearly no speakers. Therefore, 109 additional speakers from the south and south-west of Sweden (mainly in and around the localities of Barsebäck, Göteborg, Landskrona, Lund, Malmö and Växjö) were recorded by the author. The goal was to record speakers and speech samples as similar as possible to those in the SweDia corpus. Mostly, speakers from smaller localities and of similar social backgrounds were selected. Moreover, the exact same recording equipment and similar procedures and environments (home, quiet office room) were used (the SweDia speakers were recorded in their homes with a Sony portable DAT recorder TCD-D8 and a Sony tie-pin type condenser microphone ECM-T140 at 48 kHz/16-bit sampling frequency). Although the goal of having a material containing five speakers per gender of each age was not reached (owing to limitations in time), a corpus was created which had at least one speaker from each age (except for the ages 85, 87 and 90 years). Table 4.3 shows the age and gender distribution of the material.

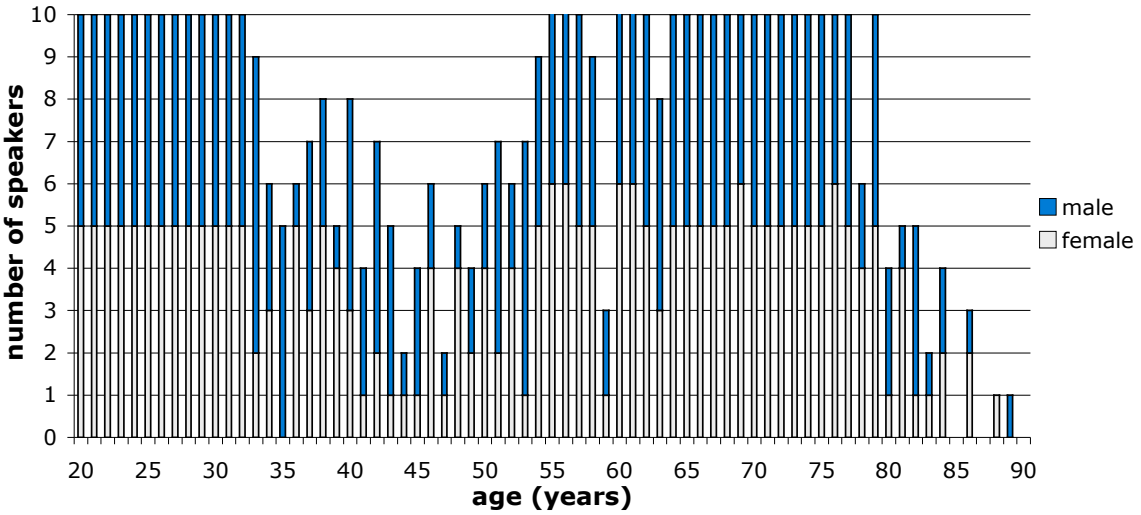


Figure 4.3: Gender and age distribution of the speakers used in Study 5

Speech type

One purpose of this study was to use automatic analysis of specific phoneme segments, such as voiceless fricatives and plosives. This approach led to restrictions regarding the choice of speech samples. Automatic recognition of spontaneous speech is still not reliable enough to be used in these kinds of acoustic studies (Ikeno et al., 2003; Furui et al., 2006; Li et al., 2006). For Swedish, even automatic aligners used to determine borders between known phonetic segments (e.g. phonemes in words whose orthography is known) have not reached a high enough performance to be used in studies such as this one (Sjölander, 2001).

After some consideration, the following six words from the word list of the SweDia 2000 material were chosen for analysis: *käke* [ˈɕɛ̀ːkə] ('jaw'), *saker* [ˈsɑ̀ːkəʁ][2] ('things'), *själen* [ˈɧɛːlən] ('the soul'), *sot* [suːt] ('soot'), *typ* [tyːp] ('type' [noun]) and *tack* [tak] ('thanks'). These words were selected because they had previously been used by the author in a perceptual study (see Section 2.7). Moreover, the words contained phonemes which in a previous study had shown tendencies to contain age-related information, namely the voiceless stops /p/, /t/ and /k/ and the voiceless fricatives /s/, /ɕ/ and /ɧ/ (Schötz, 2001b). All words – both the SweDia words and the ones recorded by the author – were elicited to yield semi-spontaneous speech. For instance, the word *tack* (thanks) was elicited by asking the speakers what single word to say when receiving a gift. Every speaker produced at least three versions of each word (sometimes four or five), yielding a total of 810 female and 836 male versions of each word. All versions were used as material in order to

[2]The majority of the speakers spoke a South Swedish dialect variety where the velar or uvular [ʁ] is used. However, there were also some speakers from other dialectal areas where the alveolar [ɹ] is used instead.

include some intra-speaker variation. As the recording sessions carried out by the author did not include as many words as the ones in the SweDia corpus, measures were taken to reduce effects related to the unusual situation which a recording session can be. The actual recording sessions were preceded by practice sessions.

Problems with pre-annotated material

There are several dangers associated with using a material which has been previously annotated by different transcribers. Intra- and inter-transcriber agreement is difficult to maintain when numerous transcribers annotate a large speech corpus during a period of several years. Problems in the SweDia corpus included differences in the criteria for word beginnings and ends. For instance, some transcribers did not place the initial word border for words with initial plosives at the beginning of the plosive release (burst), but rather several milliseconds before the release. Another discrepancy concerned the segmentation of words ending with a final aspirated plosive. Such differences may influence the results of e.g. the analysis of segment duration.

4.4.4 Method and procedure

The methodological approach of the present study can be described as mainly automatic acoustic analysis of a large number of features. Automatic methods offer several advantages over manual ones, including that the exact same task is consistently carried out and that decisions are not affected by opinion or arbitrariness. However, occasional manual interventions were necessary at some points. The procedure involved automatic extraction of 161 acoustic features from selected phonetic segments (a different number of features was extracted for different segments) using a Praat script. The feature values were stored in text files. The subsequent analysis was performed with m3iCat, a toolkit specially developed for corpus analysis (Müller, 2005). M3iCat was used to calculate statistical measures and to generate tables and diagrams displaying the variation of a certain feature as a function of age.

Tools

Two major analysis tools were used in the experiment. Praat (see p. 31) was used in the pre-processing and feature extraction stages of the procedure.

In the pre-processing, a Praat script was used to normalise all words for intensity, and the aligner used to segment the words into phones and sub-phone segments also consisted mainly of Praat scripts (see p. 59).

Feature extraction was also carried out automatically using a Praat script. The script was originally kindly developed for the author by Johan Frid[3], and was further adapted and extended for this study by the author. Most features were extracted using Praat's built-in functions and algorithms, namely values for duration, intensity, F_0, voice report (including HNR and NHR, see p. 105), formant frequencies, jitter, shimmer and LTAS. Some of the arguments to the functions had to be adjusted for female speakers. For instance, the min and max F_0 values were set differently for female and male speakers in Praat's pitch analysis. Additional functions were developed to extract other features, e.g. spectral balance and spectral emphasis.

After the feature extraction, the corpus analysis tool m3iCat (mobile multi-modal interaction corpus analysing toolkit), developed by Christian Müller[4], was used to calculate and display statistical measures, including normalised mean values which enabled comparison of the different features and their variation with speaker age. M3iCat was developed within the project m3i, which in turn is part of the project COLLATE (computational linguistics and language technology for real-life applications), sponsored by the German Federal Ministry for Research and Education, and carried out by DFKI at Saarland University. It is an extensive tool for performing and controlling corpus analysis and for selecting and displaying results (i.e. by generating several graphs and tables of statistical values). For more comprehensive descriptions of m3iCat, see Müller (2005) and Germesin (2006). Christian Müller adapted the analysis and layout (of the resulting tables and graphs) parts of the tool for the present study, and was also very helpful in explaining the functions of the tool to the author.

Pre-processing

The pre-processing of the data involved intensity (SPL) normalisation, downsampling to 16kHz of all SweDia speakers, alignment and concatenation of the words. Intensity normalisation was done to reduce intensity variation related to microphone distance. As a consequence, only *relative intensity* (see p. 10) could be measured. The alignment and concatenation procedures are described in the following sections.

Alignment

The six words were aligned separately and automatically, i.e. each single word was transcribed on the phoneme level, and the plosives were further divided into closure and voice onset time (VOT) or final aspiration segments.

[3]Department of Linguistics and Phonetics, Centre for Languages and Literature, Lund University.

[4]German Research Center for Artificial Intelligence (DFKI), Saarland University.

Figure 4.4 displays an alignment example of the word *tack*. An attempt was made to use the same aligner which had been used previously to align the word *rasa* with acceptable accuracy (see p. 59). However, it turned out that it did not reach the same accuracy when transcribing these six words. Therefore, a number of Praat scripts were developed to align the six words and to perform additional adjustments of segment borders for each of them, using a knowledge-based approach. When these measures did not yield an acceptable accuracy, another script was used to align some of the words using fixed phoneme and sub-phoneme segment boundaries based on average segment duration percentages. After that, each border was further adjusted using additional Praat scripts.

The alignments were further checked several times using a number of additional Praat scripts in order to detect errors including wrong label names, wrong number of segment boundaries, etc. Any errors were corrected manually. Sample checks revealed that most errors came from a limited number of speakers, and they were subsequently checked extra carefully.

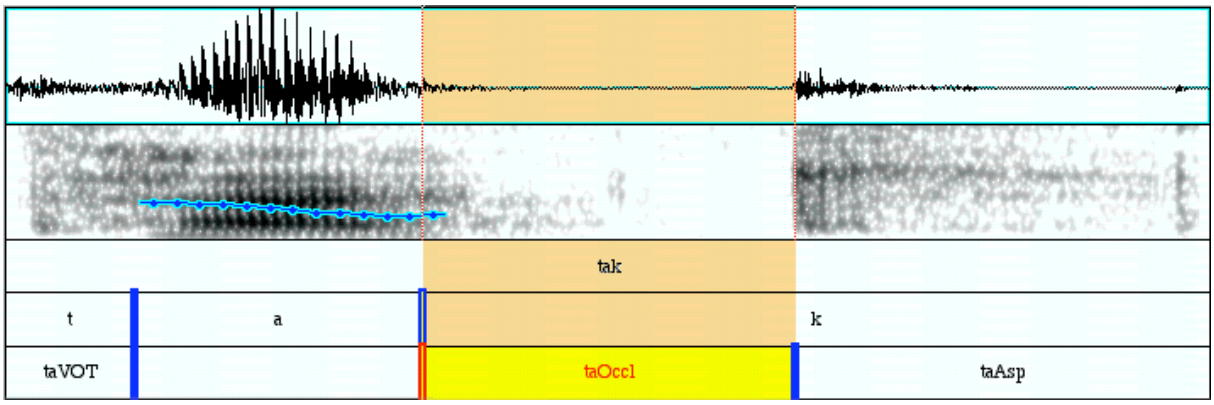


Figure 4.4: Example of the word *tack* aligned into word, phoneme, VOT and final plosive closure and aspiration segments (female speaker, 38 years)

Concatenation

The aligned words were concatenated into larger files. All first elicitations of one speaker were combined into a six-word sound file, all second elicitations combined into a second concatenated segmented file, and so on, until all elicitations by each speaker had been concatenated into files containing all six words. The words were concatenated in the same order as in a previous perception test (see Section 2.7), i.e. *käke, saker, själen, sot, typ, tack*. An example of a concatenated file with all six words is shown in Figure 4.5.

Once all files had been concatenated, additional Praat scripts as well as several manual sample checks were run in an attempt to find any remaining alignment errors, which were corrected manually. It is difficult to tell how

many errors still remain in the material. However, after several sample checks the alignment was judged to be of sufficient accuracy to be used for acoustic analyses of potential age correlates.

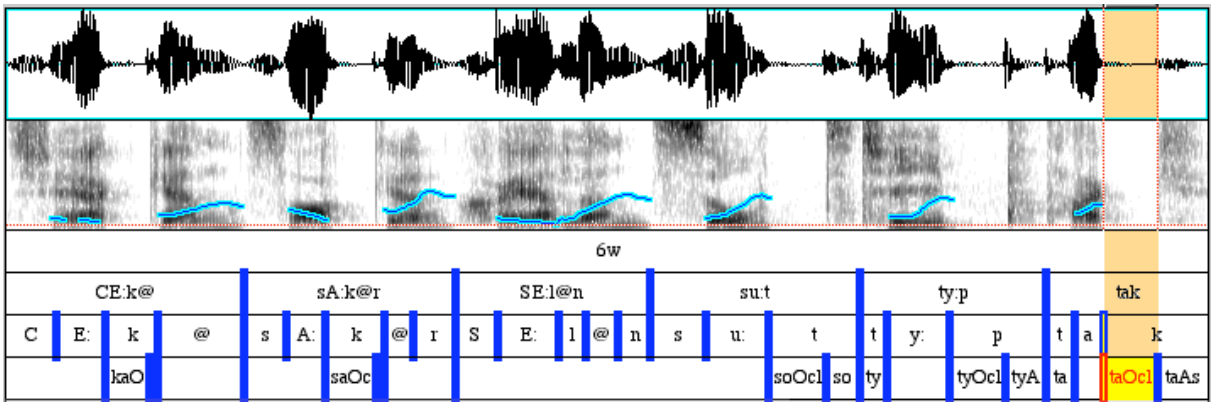


Figure 4.5: Example of a concatenated file aligned into word, phoneme, plosive closure, VOT and aspiration segments (male speaker, 42 years)

Feature extraction

The aligned concatenated words were subjected to a Praat script that extracted a large number of features from the various segments. Different data sets were created for each feature group as well as for female and male speakers. The total number of features was 161, divided into seven feature groups. Some features (e.g. syllables and phonemes per second, jitter and shimmer) were extracted only for the whole file containing all six concatenated words, while others (e.g. F_0, formant frequencies and segment duration) were extracted for several segments, including the six words and the stressed vowels. Table 4.3 offers an overview of the segments analysed in each feature group.

Table 4.3: Segments analysed in each feature group (*sp.*: spectral)

No	*Feature group*	*Segments analysed*
1	syllables & phonemes per second	all six words
	segment duration (ms)	all six words, words, stressed vowels, fricatives, plosives (closure, VOT and aspiration)
2	intensity (dB)	
3	F_0 (Hz, semitones)	all six words, words, stressed vowels
4	jitter, shimmer	all six words
5	sp. tilt, sp. emphasis, inverse filtered SB, LTAS	
6	HNR, NHR, other voice measures	all six words, stressed vowels
7	formant frequencies (F_1–F_5) sp. balance (SB)	stressed vowels fricatives and plosives

Feature group 1 contained the number of syllables and phonemes per second as well as segment durations for all six words and each individual word,

vowel, fricative and plosive (including closure, VOT and aspiration intervals). However, since there were no natural pauses in the six concatenated words, no analyses of pause frequency and duration were possible.

Groups 2 and 3 contained the following built-in Praat measures of intensity and F_0: mean and median (only relative measures for intensity), range, range without upper and lower 10% (to exclude outlier values) and standard deviation.

The five measures of jitter and the six measures of shimmer provided by Praat as built-in functions were used to extract these features in group 4.

Group 5 comprised several voice measures. Spectral tilt was calculated as the energy difference (in dB per octave) between 1 and 4 kHz in the inverse filtered signal, while spectral emphasis was calculated as the difference (in dB) between the overall intensity and the intensity in a low-pass filtered signal set to 1.5 times F_0 (Traunmüller and Eriksson, 2000; Heldner, 2001). Five contiguous frequency bands were used for the measurements of inverse filtered spectral balance: 0–0.5 kHz, 0.5–1 kHz, 1–2 kHz, 2–4 kHz and 4–6 kHz, the first four being in accordance with Sluijter and van Heuven (1996). Inverse filtering was carried out using Praat's built-in inverse filtering function to obtain measures of the spectral distribution of the source signal.

For LTAS, mean amplitudes were measured from 0 to 8 kHz at contiguous intervals of 160 Hz, following Mendoza et al. (1996) and Linville (2002). In addition to the "normal" LTAS function in Praat, a pitch-corrected LTAS function was used, which corrected away the influence of F_0 without sacrificing frequency selectivity. "The resulting Ltas is meant to reflect only the resonances (formants) in the vocal tract and the envelope of the glottal source spectrum" (Boersma and Weenink, 2005).

Group 6 consisted of two spectral noise features (HNR and NHR) as well as three additional voice measures (fraction of locally unvoiced pitch frames, number and degree of voice breaks) which were all extracted using the built-in voice report function in Praat. HNR was calculated as $10 \times log10(e_h/e_n)$, where e_h is the harmonic energy and e_n the noise energy. NHR was calculated using the same algorithm as in the voice program MDVP (see p. 87). The fraction of locally unvoiced pitch frames was calculated as the number of pitch frames analysed by Praat as unvoiced divided by the total number of pitch frames in the speech sample. The number of voice breaks was calculated as the number of distances longer than 1.25 between consecutive glottal pulses divided by a pre-defined pitch floor. The degree of voice breaks was measured as the total duration of the breaks between the voiced parts of the speech signal divided by the total duration of the speech signal. The reader is kindly referred to the Praat manual (Boersma and Weenink, 2005) for a more detailed description of these features.

The last feature group, group 7, included measures of formant frequencies

and spectral balance. For F_1–F_5, the following measures were extracted: mean, median, range, range minus upper and lower 10% (to exclude outlier values) and standard deviation. The spectral balance was measured as the peak amplitude of the spectrum in the following frequency bands: 0–0.5, 0.5–1, 1–2, 2–4 and 4–6 kHz, i.e. the same bands as for inverse filtered spectral balance.

Appendix A (p. 170) lists the individual features of each group.

Feature analysis

The data files were analysed with the m3iCat tool. As it would be impractical to analyse the data (and present the results) for each year of age, a subdivision of the speakers into eight age classes was made, based on the human listening test with the same six words which is described in Section 2.7. The best estimation accuracy obtained in that test was ± 8 years for female speakers. Based on this accuracy, the age classes were tentatively defined as eight overlapping "decade-based" classes with 14 ages in each class (except for the youngest and oldest classes):

- 20: aged 20 to 27
- 30: aged 23 to 36
- 40: aged 33 to 46
- 50: aged 43 to 56
- 60: aged 53 to 66
- 70: aged 63 to 76
- 80: aged 73 to 86
- 90: aged 83 to 89

Overlapping classes were preferred over discrete ones owing to the problems associated with defining discrete age classes. For instance, if one class is defined as 30–39 years and another as 40–49 years, 39-year-old and 40-year-old speakers will fall into two separate classes, despite their closeness in age as well as their assumed similarity in features. Allowing overlapping classes enables this problem to be avoided. A consequence of overlapping classes is, however, that some speakers will belong to two different age classes. For instance, a 25-year-old speaker will belong to both the age classes 20 and 30. Nevertheless, this was judged to have little or no influence on the results, because of the relative closeness in age to both classes for these speakers.

Several other age class definitions (three, five and seven discrete classes, as well as five overlapping classes) were also tested in the analysis. As it turned out, the results were quite similar for all age class definitions. Therefore, only the eight overlapping classes will be described in the results.

For each feature, m3iCat calculated actual means (μ), standard deviations (σ) and normalised means ($\bar{\mu}$) for each age class. Normalisation involved mapping the domain of the values in the following way:

$$a_i = \frac{(v_i - mean)}{stdev} \tag{4.1}$$

where v_i represents the actual value, *mean* represents the mean value of the data and *stdev* represents the corresponding standard deviation. Occasionally, normalisations were also carried out separately for each gender. This was done in order to see the age-related variation more distinctly when there were large differences in the real mean values between female and male speakers, e.g. in F_0 and formant frequencies. Because of the normalisation process, almost all values (except a few outliers) fall within the range between -1 and $+1$, which allows direct comparison of all features regardless of their original scaling and measurement units.

The values calculated for the eight age classes were displayed in tables, separately for female and male speakers. In addition, line graphs were generated for the age-class-related profiles or tendencies, with the age classes on the x-axis and the normalised mean values on the y-axis. The differences between the normalised mean values of all pairs of adjacent age classes are displayed as labels at the top of the diagrams (female labels above male ones). Statistical t-tests were carried out to calculate the significance of the differences; all differences except the ones within parentheses are statistically significant ($p \leq 0.01$). Figure 4.6 shows an example of a tendency diagram where the normalisations were carried out using all speakers, while Figure 4.7 shows the same tendencies but normalised separately for each gender.

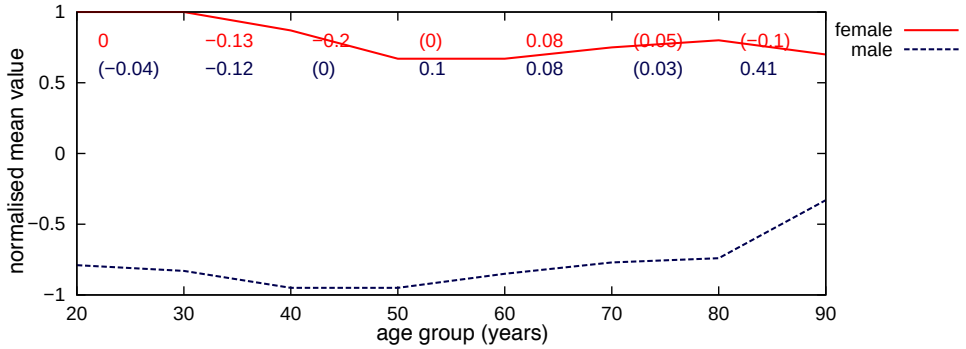


Figure 4.6: Normalised tendencies for *mean F_0 (Hz) (all six words)* (8 overlapping age classes, normalised for all speakers)

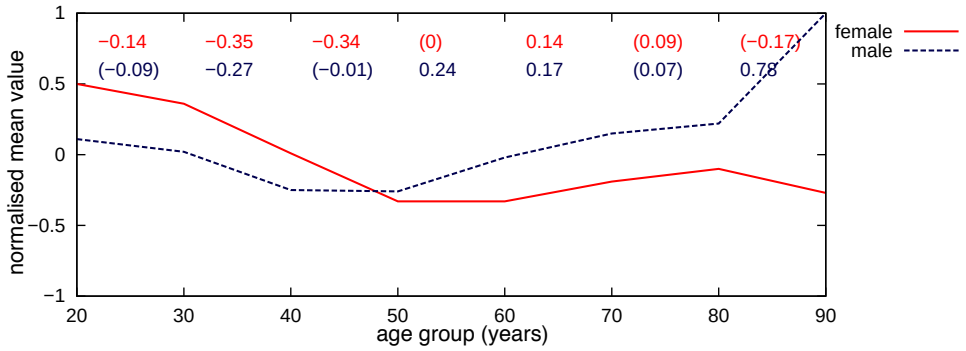


Figure 4.7: Normalised tendencies for *mean F_0 (Hz) (all six words)* (8 overlapping age classes, normalised separately for female and male speakers)

The advantage of using normalised means is that variation can be studied across features regardless of differences in the original scaling and units of the features. For instance, it allows direct comparison of the age-related variance between duration and F_0 by comparing the tendency for segment duration (in seconds) with the tendency for mean F_0 (in Hz).

4.4.5 Results and discussion

Owing to the large number of features investigated, the results are presented by feature group (see Table 4.3) and each set of results is immediately followed by a discussion in order to spare the reader unnecessary browsing and scrolling. Moreover, only a selection of interesting results for each feature group will be described here. It would be impossible to present the results for every single feature within the scope of this thesis. Instead, something is said about each feature group as a whole, and then one or several particularly interesting results within the group are presented. Normalised means have been calculated using all speakers, unless otherwise explicitly stated (cf. Section 4.4.4).

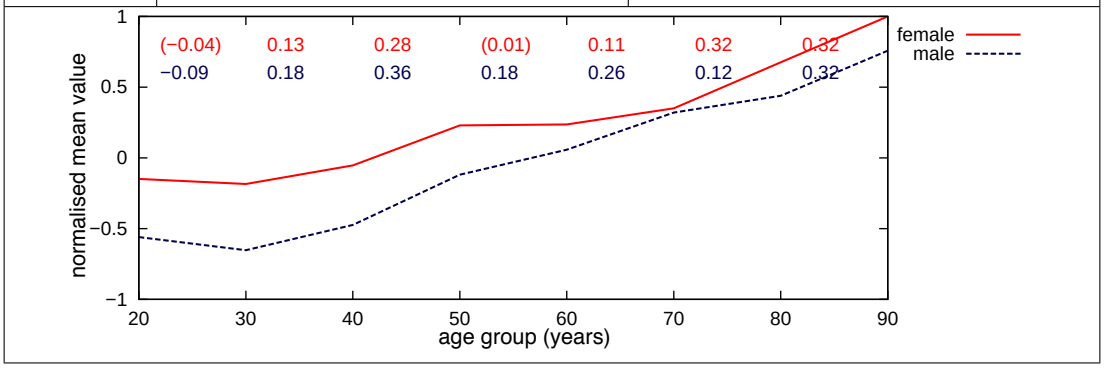
Speech rate

The number of syllables and phonemes per second generally decreased with advancing age group for both genders, while segment duration increased for most segments. The tendencies were less clear for the female than the male speakers, especially for the age classes 30 and 60, where sometimes reversed tendencies were observed. Statistically significant results were found for all six concatenated words and for each single word. Stressed vowel duration also increased significantly with age. However, the duration of the two vowels [a] in *tack* and [yː] in *typ* tended to decrease from age class 30 to 40, before a significant increase followed. Plosive duration increased with age in [k],

while [p] and [t] demonstrated irregular duration patterns, except for initial [t], which decreased somewhat with male age. Fricative duration displayed only minor increases with advancing female age. Male fricative duration remained relatively stable until age class 80, where significant increases were observed, at least in [ʃ] and [ɕ]. In the sub-phonemic segments, plosive closures increased with advancing female and male age in the two disyllabic words *käke* and *saker*, but not in the three monosyllabic ones. VOT in the medial plosives showed no significant change with age, except for women in the word *tack*, which rose considerably from age class 20 to 50, then fell substantially until age class 70, when another significant rise followed into age class 90. Final plosive aspiration duration generally increased with age. The strongest tendencies were found in the word *sot*. Table 4.4 shows the results for all six words. Interestingly, the standard deviation (σ) almost doubled from age class 20 to 90 for the female speakers, while it shows much less variation for the male speakers. A normalised tendency diagram for the number of syllables per second is displayed in Figure 4.8.

Table 4.4: Mean (μ), standard deviation (σ) and normalised mean ($\bar{\mu}$ and also in the diagram) values for *duration (all six words)*

dur_6w	female speakers			male speakers		
age class	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
20	3.6539	0.4764	-0.1491	3.4156	0.4809	-0.5599
30	3.6335	0.4618	-0.1842	3.3619	0.4583	-0.6525
40	3.7095	0.5284	-0.0532	3.4656	0.5697	-0.4737
50	3.8734	0.4887	0.2294	3.6718	0.5275	-0.1183
60	3.8773	0.4959	0.2362	3.7742	0.4915	0.0583
70	3.9437	0.557	0.3506	3.9259	0.5629	0.3201
80	4.1319	0.6954	0.6752	3.995	0.5689	0.4391
90	4.4181	0.8926	1	4.18	0.4307	0.7582



This study generally confirms the hypothesis that speech rate decreases with age and that segment duration consequently increases. Explanations

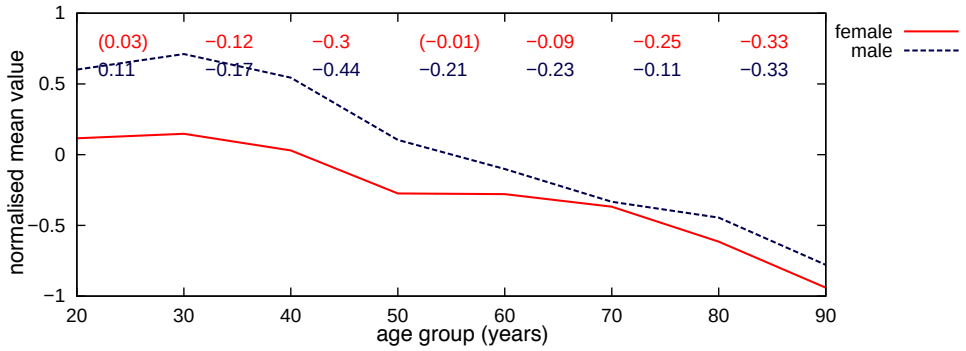


Figure 4.8: Normalised means and tendencies for *syllables per seconds (all six words)*

suggested for this phenomenon include neuromuscular slowing, altered nerve supply, increased cautiousness and adjustment by older speakers of their tempo to maintain speech fluency (Linville, 2001). Segment duration did increase in most cases, even in sub-phonemic segments such as aspiration. Deviating tendencies were observed mostly in very short segments, such as [a] in *tack*, initial [t] in *typ* and *tack*, and in VOT segments. The fact that VOT did not increase significantly in this study is in line with some previous research. It is not unlikely that age-related variation in VOT is connected to phonetic context and perhaps even language. The minor or non-significant variation in fricative duration may be explained by the fact that the segments examined here were utterance initial, while several previous reports of age-related variation seem to be based on fricatives in carrier phrases (e.g. Morris and Brown, 1987, 1994b). The standard deviation was in some cases relatively high for women (19% for female age class 90 in duration for all six words; see Table 4.4). These results confirm the reports of increased variation in older speakers, at least for female speakers. An interesting result was that speech rate decreased almost continuously throughout the ageing process, interrupted only by occasional plateaus, mainly for women. This indicates that speech rate is a powerful cue not only for distinguishing old speakers from young ones, but also for identifying speakers who are relatively close in age to each other.

Intensity

Since the words had been normalised for intensity, the mean and median measures reveal results only for relative segment intensity (see p. 10). Average relative intensity generally either decreased slightly or remained relatively constant with advancing female and male age in most analysed segments. Occasionally, intensity increased for one age group (often the oldest) or gender (often females). However, some exceptions were found. For instance,

the vowel [u:] displayed a sharp drop in intensity from age class 30 to 40 (females) or 50 (males), but then increased significantly again. An intensity peak was also observed in [s] in age group 40 for both genders. Figure 4.9 shows the tendency diagram for mean relative intensity for all six words.

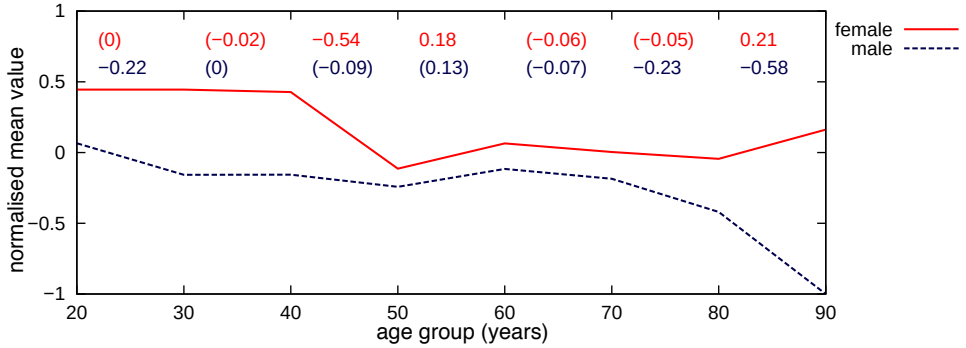
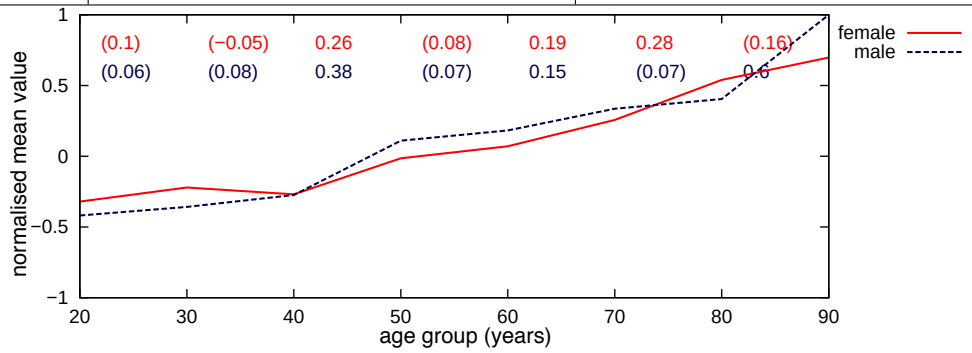


Figure 4.9: Normalised tendencies for *mean relative intensity (all six words)*

Table 4.5: Mean (μ), standard deviation (σ) and normalised mean ($\bar{\mu}$ and also in the diagram) values for *intensity range (the word käke)*

in_r_ceeke	female speakers			male speakers		
age class	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
20	34.913	4.569	-0.32	34.295	5.26	-0.419
30	35.538	4.821	-0.221	34.679	4.843	-0.358
40	35.235	6.653	-0.269	35.203	6.083	-0.274
50	36.836	6.521	-0.014	37.616	6.795	0.11
60	37.366	6.133	0.07	38.07	6.721	0.182
70	38.54	6.121	0.257	39.034	6.527	0.336
80	40.316	6.36	0.54	39.462	6.563	0.404
90	41.313	7.569	0.698	45.171	4.207	1



The intensity range either increased (six concatenated words, single words, consonants, plosive closure and aspiration) or remained relatively stable (vowels, VOT) with advancing age, as did the intensity standard deviation. Again, the vowel [u:] was an exception, with a large intensity range and SD

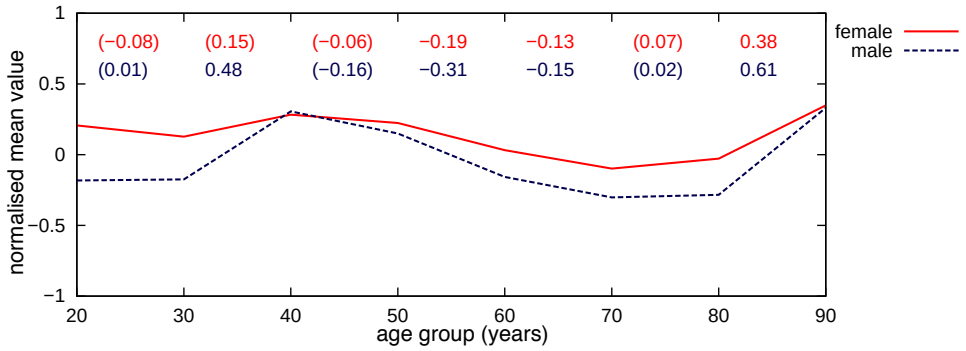


Figure 4.10: Normalised tendencies for *intensity standard deviation ([ɛː] in the word själen ['ʃɛːlən])*

increases from age class 30 to 40, followed by a considerable decrease in the older age classes. Table 4.5 shows the results for intensity range in the word *käke*.

The only feature measured in this study which may compare with the amplitude stability (or Amp SD) investigated in some previous studies is intensity SD in the stressed vowels. No consistent increase or decrease with older age was found for any of these vowels in either gender. Figure 4.10 displays the results for the vowel with the longest duration: [ɛː] in the word *själen* (mean duration: 236.7 ms).

These results concerning average relative intensity of normalised data are difficult to compare with the findings of previous studies. Therefore, hypotheses were made only about intensity range. Apart from the normalisation of the data, there were other differences compared with other studies as well. The speech material used in this study consisted of elicited isolated words. The general trend observed here of decreasing or stable mean intensity levels for the segments investigated was occasionally broken for certain age groups or one of the genders. Further analysis of the results is needed in order to understand the reason for this. One explanation may be that intensity variation with age is phoneme dependent (cf. the results for the vowel [uː]).

The relatively consistent increase in intensity range with increasing age for both genders suggests a larger variation in intensity with advancing age, thereby confirming the hypothesis of increased habitual intensity range. This feature may thus be an important correlate of age.

Since intensity standard deviation (or Amp SD) remained relatively stable with the strongest increase after age group 80, the hypothesis of an increase in this feature with age was only partially confirmed. One explanation may be that previous studies (on which the hypothesis was based) have only compared young and elderly speakers. Another explanation may be the

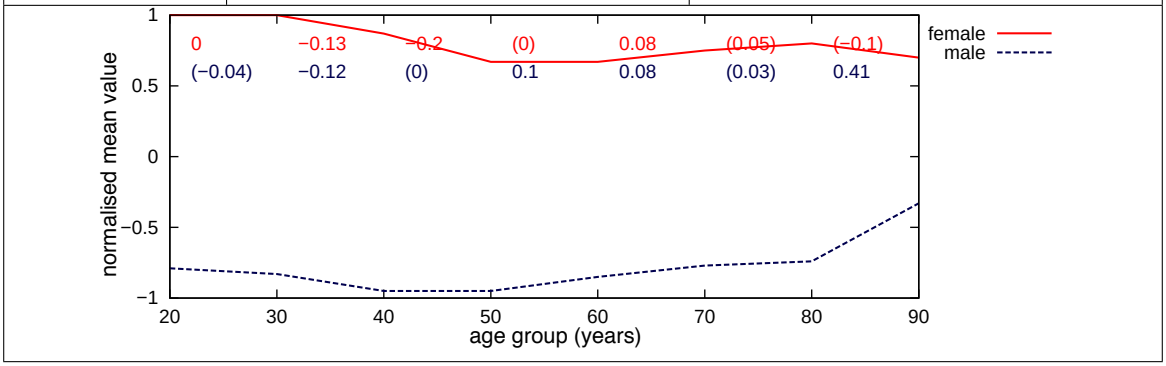
relatively short duration of the vowels in this study. Mean duration values ranged from 109.8 ms (male [a] in *tack*, age class 40) to 330.2 ms (female [ɛː] in *själen*, age class 90).

F_0

For mean and median F_0, significant age-related variation was found in all analysed segments (all six words, single words, stressed vowels), both in Hz and in semitones. Female F_0 decreased until age group 50. Then followed a slight increase until age group 70, when another decrease was observed. Male F_0 lowered slightly until age group 50, but then rose into old age. Owing to the gender-related differences in F_0, the results for mean F_0 (Hz, all six words) are presented both in Table 4.6 (normalised for all speakers) and in Figure 4.11 (normalised separately for each gender); the latter shows more distinct tendencies.

Table 4.6: Mean (μ), standard deviation (σ) and normalised mean ($\bar{\mu}$ and also in the diagram) values for *mean F_0 (Hz, all six words)*

f0_meanhz_6w	female speakers			male speakers		
age class	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
20	216.95	24.89	1	120.42	17.49	-0.79
30	212.8	26.24	1	118.51	18.87	-0.83
40	202.42	31.26	0.87	112.76	24.93	-0.95
50	192.21	30.2	0.67	112.56	21.43	-0.95
60	192.22	28.87	0.67	117.59	21.18	-0.85
70	196.38	28.51	0.75	121.29	21.19	-0.77
80	199.03	29.74	0.8	122.79	23.39	-0.74
90	194.08	34.51	0.7	143.17	12.44	-0.33



The results for F_0 range (in semitones) are shown in Figure 4.12. Both female and male values remained stable until about age group 60. After that, women displayed a significant increase followed by a slight decrease after age group 80, while a significant decrease after age group 80 was found for men.

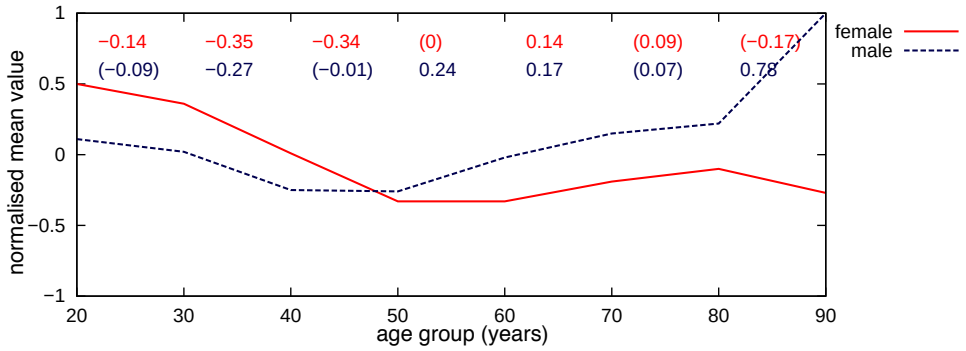


Figure 4.11: Normalised (separately for each gender) tendencies for *mean F_0 (Hz, all six words)*

F_0 standard deviation (F_0 SD) remained relatively stable, except for a rise in females after age class 60, as shown in Table 4.7. No statistically significant age-related change in standard deviation (σ) of this feature was found for either gender.

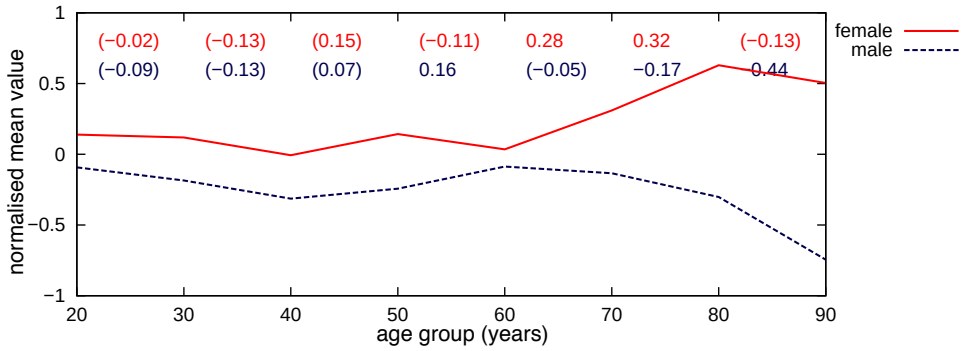


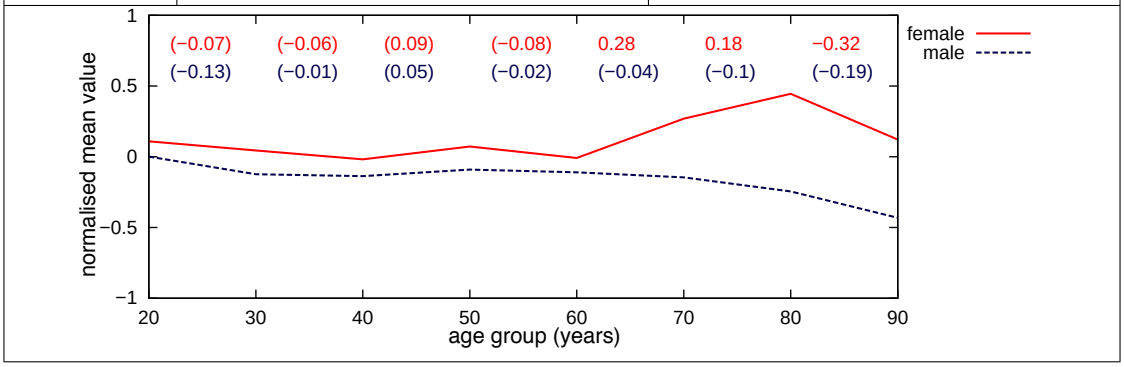
Figure 4.12: Normalised tendencies for *F_0 range (semitones, all six words)*

Variation with age in F_0 measures can be explained by age-related changes in the respiratory and/or phonatory systems (Linville, 2001). The different results found for female mean and median F_0 only partially verified the hypothesis about female F_0 (i.e. stable until menopause, then a significant decrease, followed by either an increase, a decrease or no change). This is because this study observed a significant decrease also in the young age classes, perhaps due to the more detailed subdivision into a larger number of age classes used in this study. Still, women of age classes 50 and 60 demonstrated the lowest F_0 values. Also, the varying results for women over age class 60 supported the hypothesis of different female F_0 patterns in old age. For men, the results clearly confirmed the hypothesis of a slight decrease until around age 50, followed by a considerable increase into old age.

Conversational F_0 range remained relatively stable for both genders until old age, where either an increase (women) or a slight decrease (men) was

Table 4.7: Mean (μ), standard deviation (σ) and normalised mean ($\bar{\mu}$ and also in the diagram) values for *F_0SD (semitones, all six words)*

f0_sdst_6w	female speakers			male speakers		
age class	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
20	4.096	1.6639	0.1095	3.922	1.756	0.001
30	3.9913	1.595	0.0442	3.722	1.653	-0.124
40	3.8909	1.4859	-0.0185	3.701	1.773	-0.137
50	4.0365	1.5078	0.0724	3.775	1.685	-0.091
60	3.9063	1.4777	-0.0089	3.744	1.411	-0.11
70	4.3517	1.5172	0.2692	3.687	1.504	-0.146
80	4.6327	1.4348	0.4447	3.528	1.629	-0.245
90	4.1149	1.4507	0.1214	3.23	1.327	-0.431



observed. The hypothesis, which suggested an increase followed by a decrease or a stable curve, was thus not verified in this study. The differences in speech material may be one reason for the different results.

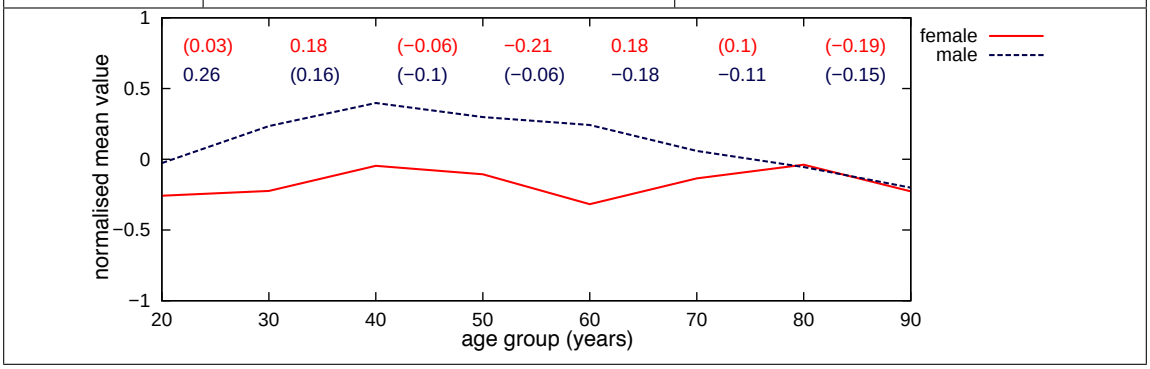
The results for F_0 SD confirmed the hypothesis of stable values regardless of speaker age, but not that of increased F_0 SD – except for age class 80 in women. Increased F_0 SD has previously been suggested to be a powerful cue to speaker age, at least to PA (Linville, 2001). However, statistically significant but only weak correlations of F_0 SD with age (for the vowel /i/ and for spontaneous speech) or no statistically significant correlation (in the vowels /a/ and /u/ and in read speech) have been reported as well (Brückl and Sendlmeier, 2003). In the present study, F_0 SD in women remained stable at around 4 semitones until age class 70, when it rose to about 4.6 semitones. It then decreased to about the same level again after age 80. Male F_0 SD did not change much with age at all. The results of this study indicate that F_0 SD may not, after all, be a strong acoustic correlate of chronological age, at least not when the speech samples consist of concatenated isolated elicited words.

Jitter and shimmer

The five measures for jitter and the six for shimmer all displayed quite similar results. Although generally higher for male than female speakers, no continuous increase with age was found in either gender for these two features. Female jitter and shimmer values remained relatively stable from young to old age. Male values generally increased slightly until age group 40, and then decreased slowly until old age, except for a considerable decrease in shimmer after age class 80. Table 4.8 displays the results for local jitter, while Figure 4.13 shows local shimmer for all six words.

Table 4.8: Mean (μ), standard deviation (σ) and normalised mean ($\bar{\mu}$ and also in the diagram) values for *local jitter (all six words)*

jitt_local_6w	female speakers			male speakers		
age class	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
20	0.0233	0.0061	-0.2575	0.0256	0.0067	-0.0274
30	0.0236	0.0063	-0.2243	0.0281	0.016	0.2341
40	0.0254	0.0076	-0.0459	0.0297	0.019	0.3973
50	0.0248	0.0071	-0.1058	0.0288	0.0084	0.2985
60	0.0227	0.006	-0.3174	0.0282	0.008	0.2426
70	0.0245	0.0072	-0.135	0.0264	0.0067	0.0586
80	0.0255	0.0079	-0.0383	0.0253	0.0062	-0.0555
90	0.0236	0.0074	-0.2271	0.0239	0.0041	-0.2007



The rather broad hypotheses for jitter and shimmer, which predicted either increased or stable values with advancing age, were partly confirmed in this study. Relatively stable values were found for both female and male speakers, except for male shimmer, which decreased significantly after age class 40. However, no increases in values with advancing age for either feature were observed. While this does support the findings of some previous studies, it is hard to draw any firm conclusions from these results because little reliable normative data exist (Linville, 2001). The large number of measures for these features and the differences in speech material used in various studies also contribute to the problem with comparison of results. More studies

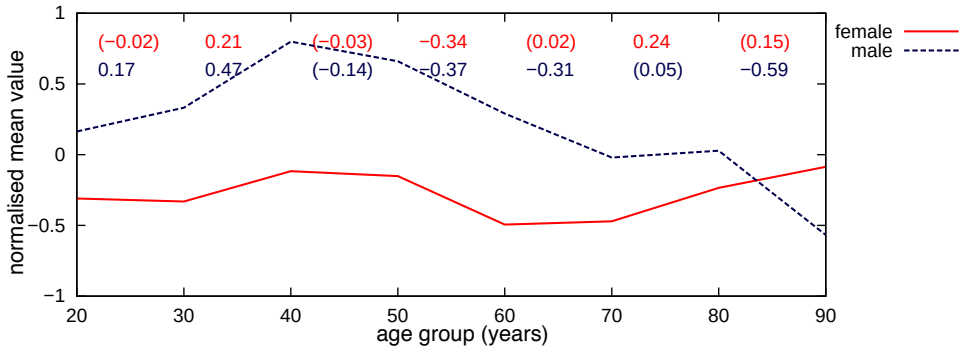


Figure 4.13: Normalised tendencies for *local shimmer (all six words)*

and a larger material are needed. The use of automatic methods may facilitate this. Furthermore, perturbation measures are probably more reliable when derived from sustained vowels than from connected speech (McAllister et al., 1998). Because of this, Shipp et al. (1992) did not investigate jitter and shimmer, as their speech material contained no prolonged vowels.

Spectral energy distribution

The distribution of spectral energy was analysed in several features. To facilitate for the reader, the LTAS results will be presented separately in the next section. Table 4.9 shows the results for spectral tilt (ST), and Figure 4.14 shows the tendency diagram for spectral emphasis (SE) for all six words. ST was rather constant in females, except for a slight increase after age class 60. For males, ST remained relatively stable until a considerable rise after age class 80. The SE pattern was similar for both genders, though the values were much higher for males. It can be described as a general rise to age class 40 or 50, followed by a decrease to age class 70 or 80. After that, SE rose into the final age classes. The standard deviation (σ in Table 4.9) of female ST as well as SE increased with advancing age, while the highest values for men were observed around age class 50.

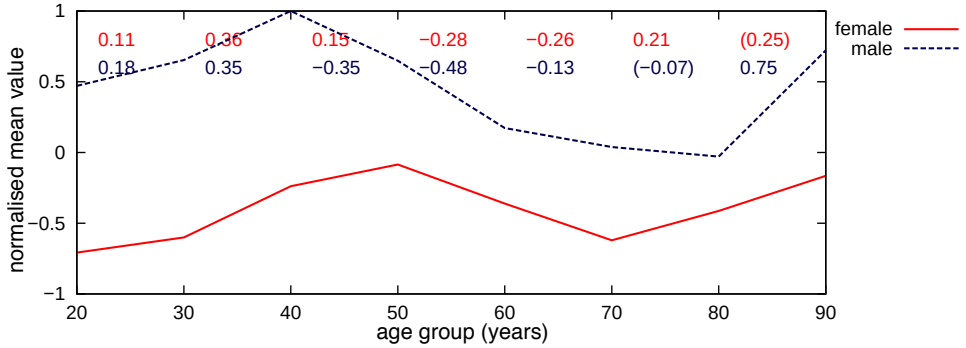
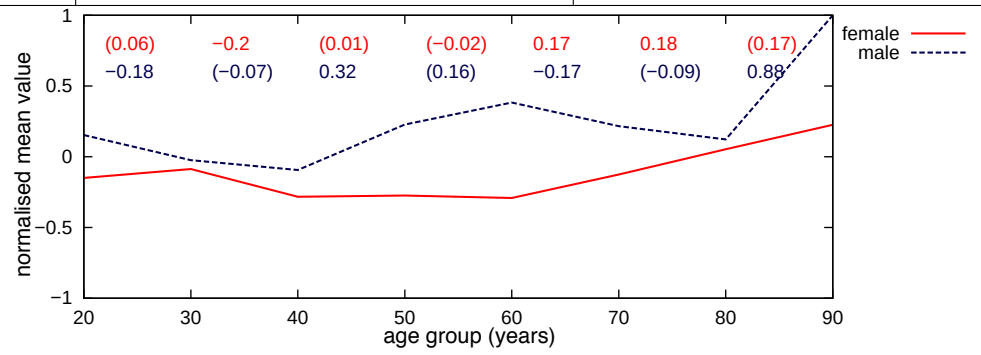


Figure 4.14: Normalised tendencies for *spectral emphasis (all six words)*

The results for inverse filtered spectral balance (ISB), i.e. the inverse

Table 4.9: Mean (μ), standard deviation (σ) and normalised mean ($\bar{\mu}$ and also in the diagram) values for *spectral tilt (all six words)*

sptlt_6w	female speakers			male speakers		
age class	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
20	-10.9862	1.9728	-0.15	-10.2147	1.822	0.1537
30	-10.8249	2.1984	-0.0866	-10.6661	1.9376	-0.024
40	-11.3224	2.6626	-0.2824	-10.8424	3.9642	-0.0934
50	-11.3024	2.5087	-0.2745	-10.0266	3.9514	0.2277
60	-11.3445	2.0524	-0.2911	-9.6311	2.1406	0.3834
70	-10.9244	2.1479	-0.1257	-10.0569	1.9863	0.2158
80	-10.4697	2.5786	0.0533	-10.2935	2.1377	0.1226
90	-10.0323	3.164	0.2255	-7.92	2.7864	1



filtered spectral energy peaks in the five frequency bands 0–0.5, 0.5–1, 1–2, 2–4 and 4–6 kHz, did not demonstrate any changes with age in females, except for a peak around age class 40 in the lowest band. For the male speakers, small (mostly statistically insignificant) peaks in the age classes 40 and 50 were observed in all bands. Figure 4.15 shows tendency diagrams for ISB in the two frequency bands 0–0.5 and 2–4 kHz.

There were no unidirectional tendencies for either of these measures. Therefore the hypothesis for spectral emphasis (no change with age) was confirmed, while that for spectral tilt (a change in either direction) was not. The results for ST tend to support a previous study (Decoster and Debruyne, 1997), which found a flatter spectral tilt in the vowel [i] for older female and male speakers. However, owing to the different speech samples used, no definite conclusions can be drawn as to the effect of ST on speaker age until this feature has been investigated more. SE displayed large differences between female and male speakers, but the most significant tendency was a strong rise after age class 80 in males. From the rather complicated tendency contours, one might suggest that SE is no reliable correlate of speaker age, which is in line with Traunmüller (1997).

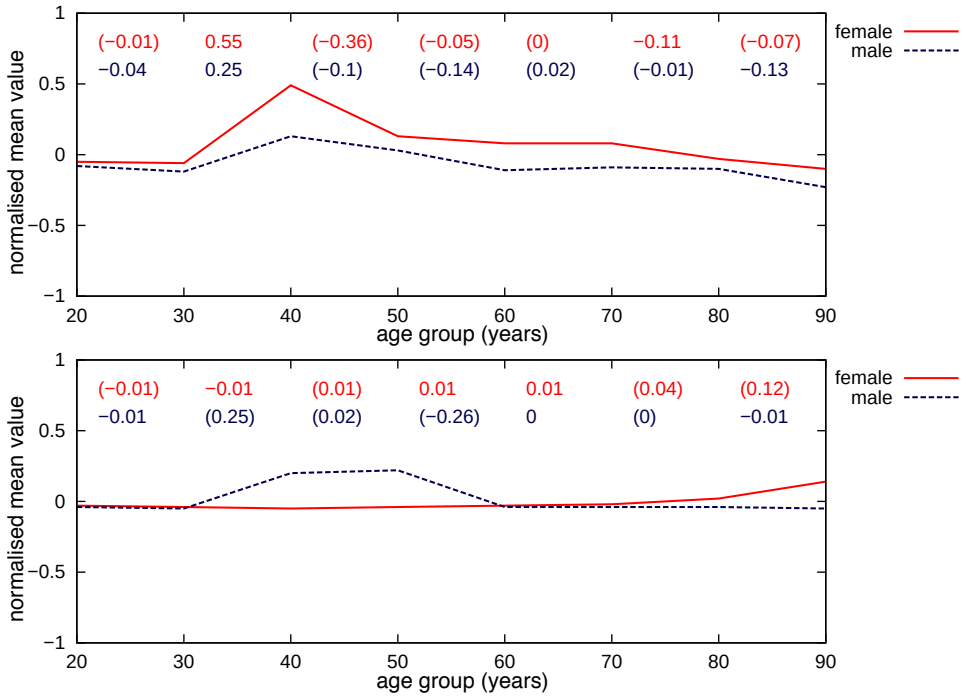


Figure 4.15: Normalised tendencies for *inverse filtered spectral balance in 0–0.5 kHz* (top) *and 2–4 kHz* (bottom) *(all six words)*

Inverse filtered spectral balance did not turn out to vary consistently with age. The peak around age class 40 in the frequency band 0–0.5 kHz found in both females and males could be explained by the small number of speakers in this age class. More studies with more speakers and different measures, including narrower frequency bands, mean energy values as well as maximum energy peaks, may help to clarify this result.

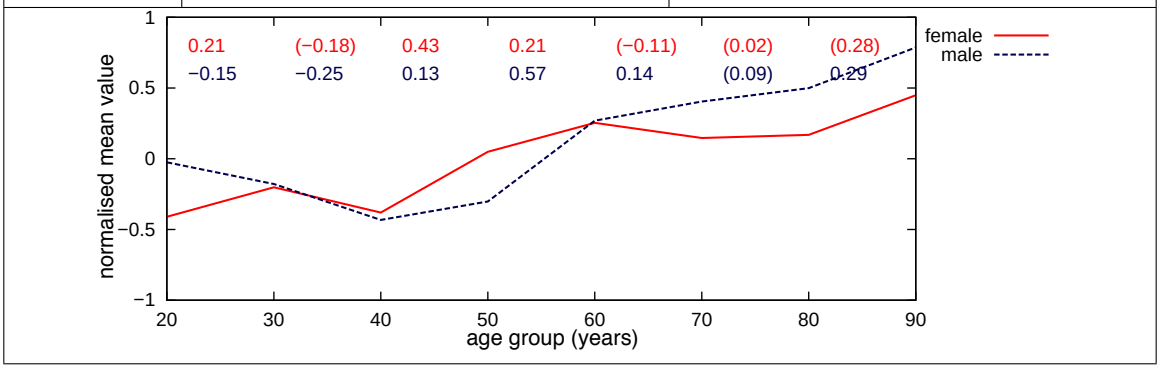
LTAS

Although part of the *spectral energy distribution* feature group, the results for LTAS are presented separately for convenience. The LTAS varied considerably in amplitude in many of the frequency intervals. Females showed somewhat higher amplitudes with advancing age at 160 (but only from age group 40 to 70), 320 and 2240–2560 Hz in both normal (i.e. where the measure was calculated without correcting for pitch) and pitch-corrected LTAS. Slightly lower female LTAS levels with increased age were found at 5920–7200 Hz in the normal measure, while the pitch-corrected measure mostly increased significantly from age group 50 to 70 at 4960–7040 Hz. Table 4.10 shows the normal LTAS results for 320 Hz, while Figures 4.16 and 4.17 display tendency diagrams for the spectral amplitude at the normal LTAS intervals of 1920 and 2560 Hz as well as 6880 and 7840 Hz. All diagrams show the results for the normal (non-pitch-corrected) LTAS.

For male speakers, moderately higher LTAS amplitudes with advancing

Table 4.10: Mean (μ), standard deviation (σ) and normalised mean ($\bar{\mu}$ and also in the diagram) values for *LTAS amplitudes at 320 Hz (all six words)*

ltas320_6w	female speakers			male speakers		
age class	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
20	48.86	2.748	-0.41	49.716	1.601	-0.025
30	49.322	2.567	-0.202	49.373	1.869	-0.179
40	48.926	2.495	-0.38	48.81	1.854	-0.432
50	49.88	1.996	0.049	49.1	2.031	-0.302
60	50.337	2.228	0.254	50.371	1.795	0.269
70	50.099	2.374	0.147	50.67	1.799	0.404
80	50.148	2.381	0.169	50.881	1.571	0.499
90	50.77	2.295	0.449	51.523	1.104	0.787



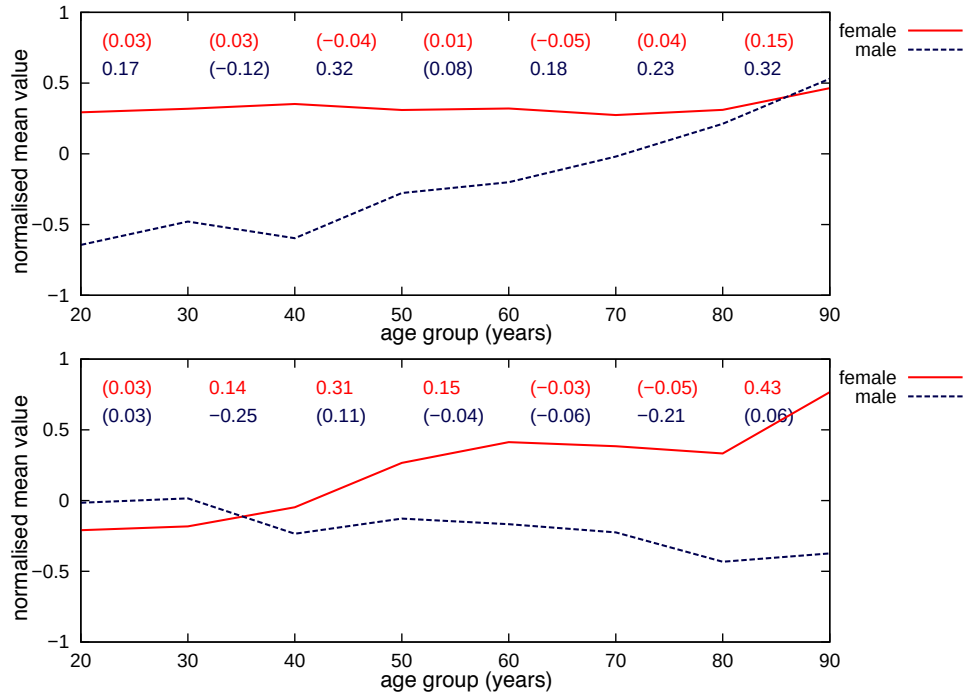


Figure 4.16: Normalised tendencies for *LTAS for 1920 Hz* (top) *and 2560 Hz* (bottom) *(all six words)*

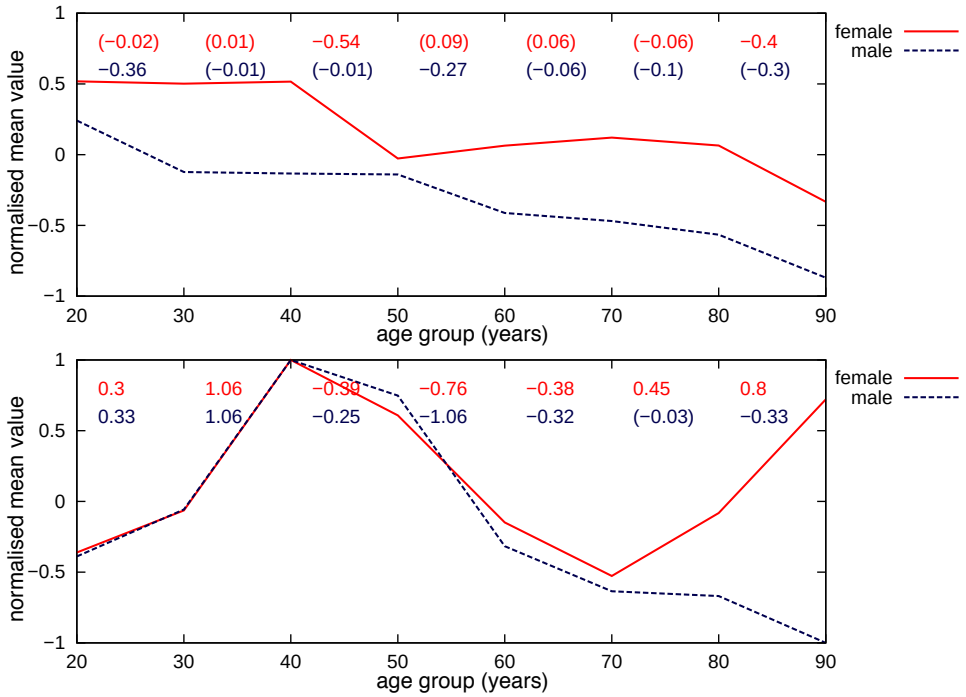


Figure 4.17: Normalised tendencies for *LTAS for 6880 Hz* (top) *and 7840 Hz* (bottom) *(all six words)*

age were for the most part observed at 160 (but only for age group 40 to 70), 320 and 1760–2080 Hz in both the normal and pitch-corrected measure. LTAS amplitudes lowered slightly with increased male age at 6560–7200 Hz in the normal measure. However, the pitch-corrected LTAS decreased till age group 40, and then increased – often statistically significantly – till age group 80 or 90 at 4160–7040 Hz.

The most dramatic variation was observed in the normal LTAS at 7680–8000 Hz, where a considerable increase occurred until age group 40. Then followed a significant decrease into old age (men) or age group 70 (women, for whom another substantial increase followed into the oldest age group).

No hypotheses had been made about LTAS, owing to the small amount of existing normative data. A few of the LTAS results in this study were consistent with those of previous studies, such as increased LTAS amplitudes with age at 320 Hz, while most were not. A possible explanation is the speech material used here, which may not be optimal for LTAS analyses.

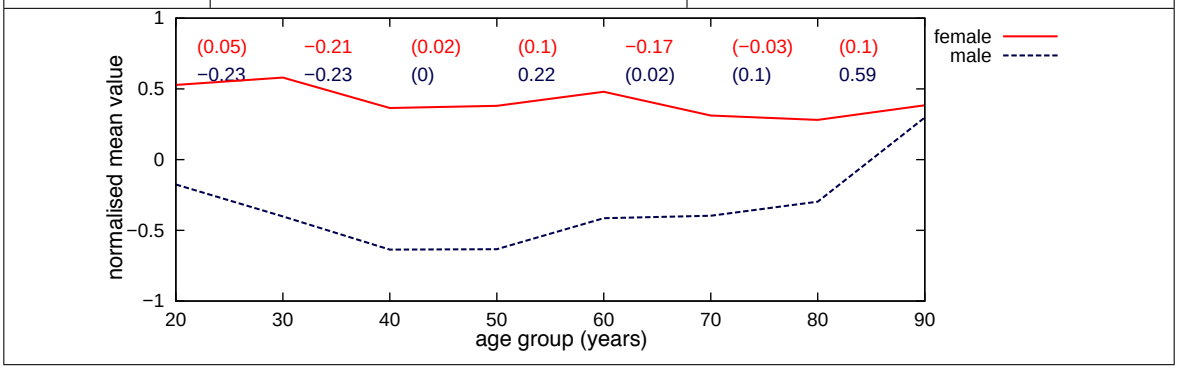
Neither spectral energy distribution nor LTAS measures seem to be strong acoustic correlates of speaker age by themselves. Still, the findings of this study did reveal at least some age-related variation, which could be used in combination with other features.

Spectral noise and other voice measures

Spectral noise measures were obtained for the stressed vowels as well as for all six words. The two measures HNR and NHR are closely related (see p. 87 and 105). HNR was found to be generally higher – and NHR lower – in females than in males. Almost no age-related variation was found for these two measures in females. In male HNR, a slight decrease followed by an increase after age class 50 was observed, while the opposite relationship was found for NHR. Table 4.11 shows the results for HNR, while Figure 4.18 displays a tendency diagram for NHR, both for the vowel [ɑː]. The other vowels displayed similar tendencies. No significant changes in standard deviation (σ) were observed for either gender.

Table 4.11: Mean (μ), standard deviation (σ) and normalised mean ($\bar{\mu}$ and also in the diagram) values for *HNR ([ɑː] in the word saker ['sàːkəʁ])*

mean_hnr_aa	female speakers			male speakers		
age class	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
20	11.647	4.153	0.528	8.604	3.81	-0.176
30	11.869	4.071	0.58	7.629	3.754	-0.402
40	10.943	4.392	0.365	6.619	3.389	-0.636
50	11.005	3.926	0.38	6.631	3.671	-0.633
60	11.437	3.638	0.48	7.578	3.788	-0.414
70	10.712	3.902	0.312	7.651	3.749	-0.397
80	10.58	4.442	0.281	8.085	3.499	-0.297
90	11.022	4.141	0.384	10.649	2.99	0.297



None of the voice measures (fraction of unvoiced pitch frames, number of voice breaks, degree of voice breaks) showed any correlation with speaker age.

The fact that HNR was higher (and NHR lower) in females than in males seems to contradict reports of breathier voice quality in women than in men, but it is actually consistent with some earlier studies (Biemans, 2000). More

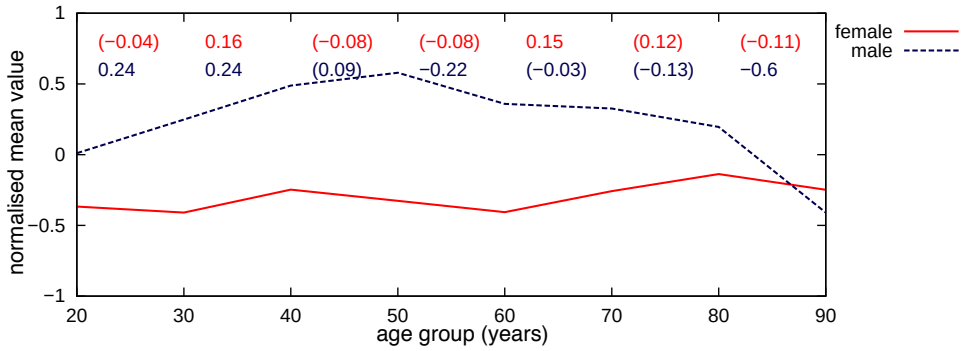


Figure 4.18: Normalised tendencies for *NHR ([aː] in the word saker ['sàːkəʁ])*

research is needed to explain this phenomenon. It was hypothesised that an increase in spectral noise would accompany advancing age in females and that spectral noise would either increase or remain constant in males. The hypothesis was thus confirmed only for the male speakers, since there was no substantial change in any of the spectral noise measures, except for the considerable increase of noise amplitude after age group 80. The stable noise levels in females may be explained by the fact that young females may apply glottal gaps to obtain a breathier voice quality (Linville, 2001). However, the slight increase in noise followed by an equally large decrease in males is difficult to explain.

Resonance measures

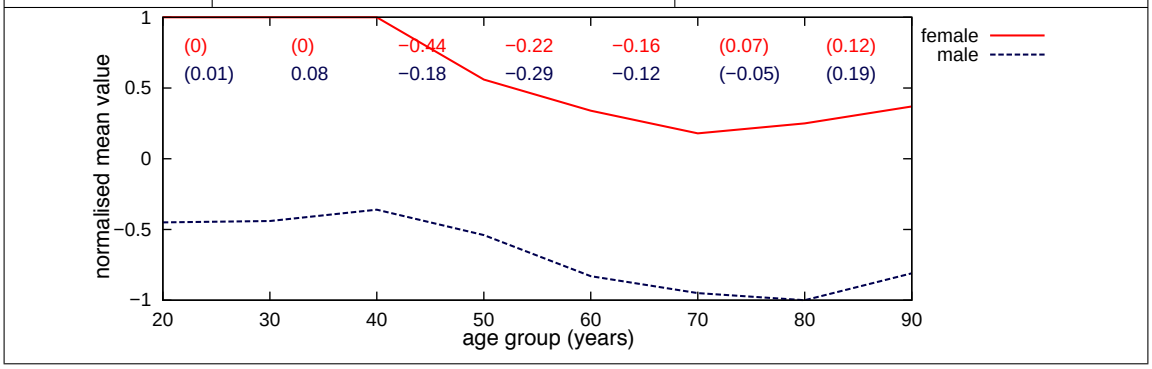
The resonance measures analysed were F_1–F_5 in stressed vowels and spectral balance (spectral energy maxima in five consecutive frequency bands: 0–0.5, 0.5–1, 1–2, 2–4 and 4–6 kHz) in fricatives and plosives. As expected, female formant frequencies were always higher than male ones.

F_1 in [a], [ɑː] and [uː] did not change much with age in either gender. In [ɛː], F_1 decreased with age for both genders from age class 40, but increased slightly again after age group 80, as shown in Table 4.12 (normalised for all speakers) and Figure 4.19 (normalised separately for female and male speakers with more conspicuous tendencies). Female [yː] displayed a general F_1 decrease. In male [yː], a similar decrease was interrupted by an increase around age class 40, and there was a substantial increase after age class 80.

F_2 increased slightly with advancing age in [ɑː] and [ɛː] for both genders. In [a] and [uː], F_2 tended to decrease slightly, interrupted by increases and peaks at age group 40 in both genders. A fairly stable F_2 was observed in [yː]. Table 4.13 displays the results for F_2 in [ɛː] as normalised for all speakers, while Figure 4.20 shows a tendency diagram for F_2 in the same vowel, but normalised separately for female and male speakers. Considerable increases in F_2 after age class 80 were found for females in [ɛː] and [ɑː], and for males in

Table 4.12: Mean (μ), standard deviation (σ) and normalised mean ($\bar{\mu}$ and also in the diagram) values for *mean F_1 ([ɛː] in the word själen ['ɧɛːlən])*

f1_mean_ee_2	female speakers			male speakers		
age class	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
20	725.29	61.01	1	584.5	44.74	-0.45
30	724.52	74.67	1	585.26	48.5	-0.44
40	715.18	77.46	1	592.39	51.47	-0.36
50	672.13	68.43	0.56	576.76	56.52	-0.54
60	653.63	67.63	0.34	551.08	47.65	-0.83
70	639.34	55.86	0.18	540.7	51.1	-0.95
80	645.04	67.48	0.25	535.94	48.09	-1
90	655.92	51.35	0.37	552.96	53.68	-0.81



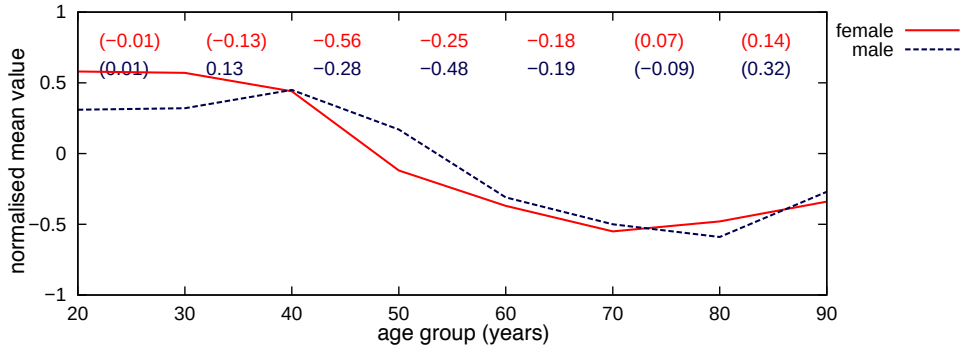


Figure 4.19: Normalised (separately for each gender) tendencies for *mean F_1 ([ɛː] in the word själen ['ɧɛːlən])*

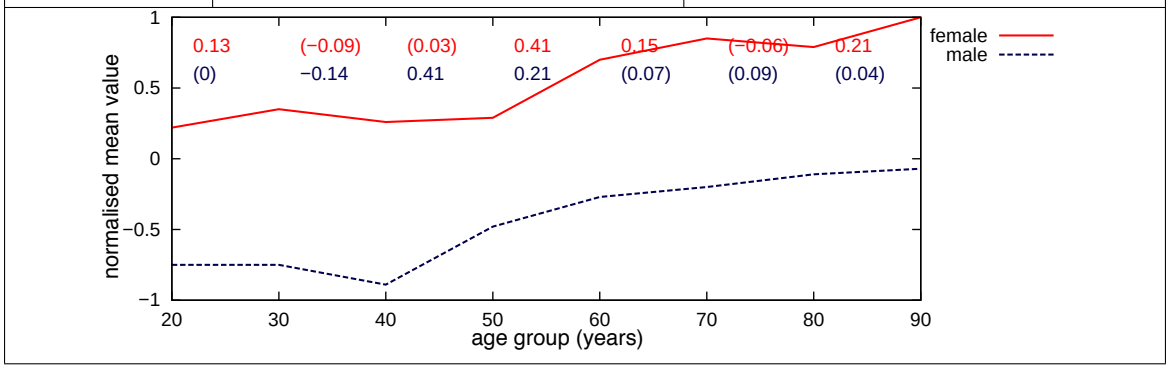
all vowels except [ɛː] and [yː]. Only minor fluctuations in F_1 and F_2 standard deviation (σ) were found in the vowels examined.

F_3, F_4 and F_5 displayed somewhat different patterns for the two genders and the six vowels. Often (but far from always) a decrease was observed from age class 20 to 30, followed by little change or a very slight increase, with an occasional rise or fall after age 80 for one or both genders. Figure 4.21 shows the results for F_3 and F_4 in [ɑː].

The results for spectral balance did not indicate any consistent changes

Table 4.13: Mean (μ), standard deviation (σ) and normalised mean ($\bar{\mu}$ and also in the diagram) values for *mean F_2 ([ɛː] in the word själen ['ɧɛːlən])*

f2_mean_ee_2	female speakers			male speakers		
age class	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
20	1690.49	127.14	0.22	1488.87	137.5	-0.75
30	1718.43	144.1	0.35	1488.64	135.51	-0.75
40	1699.52	153.67	0.26	1459.87	150.55	-0.89
50	1705.31	215.19	0.29	1545.09	169.71	-0.48
60	1790.93	205.42	0.7	1588.78	175.06	-0.27
70	1822.71	214.27	0.85	1604.18	174.52	-0.2
80	1810.24	211.32	0.79	1621.68	110.05	-0.11
90	1876.87	130.08	1	1630.07	133.85	-0.07



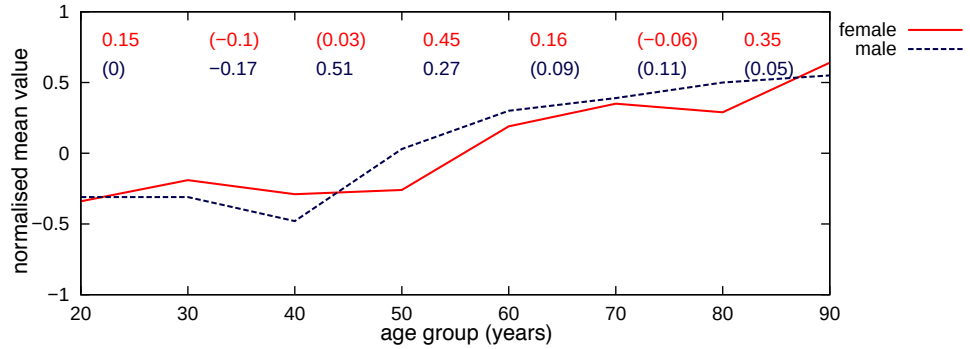


Figure 4.20: Normalised (separately for each gender) tendencies for *mean F_2 ([ɛː] in the word själen ['ɧɛːlən])*

with female and male age for either fricatives or plosives in any of the frequency bands examined. Occasional strong tendencies were found for a few bands in some segments. However, in most cases neither increases nor decreases with advancing age were observed. Figure 4.22 displays one of the largest age-related variations found in spectral balance, namely for 4–6 kHz in [k] in the word *käke* ['ɕɛ̀ːkə].

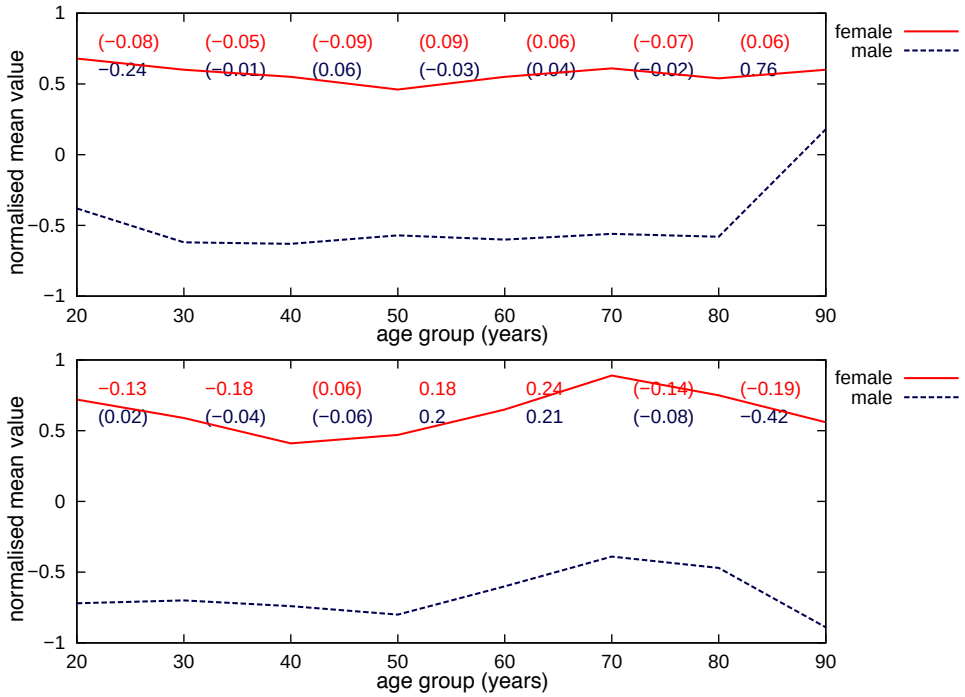


Figure 4.21: Normalised tendencies for *mean F_3 and F_4 ([aː] in the word saker ['sàːkəʁ])*

In this study, the only formant frequency that decreased with advancing age was F_1 in [ɛː] and [yː]. Therefore, the hypothesis of a general decrease of formant frequencies is supported only for F_1 in some vowels. F_2 was found to increase with age in some vowels, which contradicts previous studies. The same can be said for the mainly non-significant changes in F_3–F_5 with increasing age. A possible explanation for these results may be the short vowels of the material used here. However, before any firm conclusions can be drawn, further studies using a larger material and more reliable formant extraction algorithms are needed. Automatic formant extraction algorithms can be unreliable, especially in high-pitched voices and in the higher formants.

As for the spectral balance in the fricatives and plosives, the poor results found in this study could be explained by the relatively broad frequency bands examined. In the feature extraction script, an attempt was made to implement a function which detected the frequency above which the typical energy plateau in fricatives can be found. An earlier study (Schötz, 2003a) had shown that this frequency tended to be lower in old speakers than in young ones. However, the function in its current form did not yield any reliable results compared with manual identification of the frequency, and it was thus abandoned. More studies using narrower frequency bands are needed to examine age-related resonance characteristics in voiceless obstruents more carefully. Another possible interpretation of the results found here is that fricatives and plosives do not contain any significant age-related information.

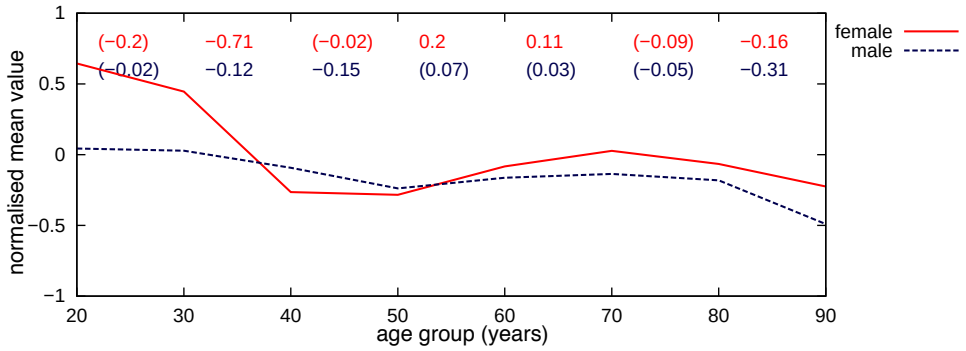


Figure 4.22: Normalised tendencies for *spectral balance 4–6 kHz ([k] in the word käke [ˈɕɛ̀ːkə])*

4.4.6 Further discussion

Since the results for the features analysed have already been discussed in the previous sections, this section discusses more general aspects of the study, such as the overall results, the effects of the speech material as well as the automatic feature extraction and analysis approach.

Speech data

Numerous previous studies of speaker age have controlled the speaker-related variables to some extent. Information on the speakers' drinking or smoking habits was frequently gathered, and the subjects often underwent hearing or oral mechanism screenings. When pre-recorded speech corpora are used, which is mostly the case in automatic studies (cf. Minematsu et al., 2002a; Müller, 2005), such variables are often uncontrollable. For instance, the SweDia 2000 corpus does not contain speaker-related information about smoking habits or regular medication. The only information available on speaker physiology is that the speakers are healthy and active members of their community. It is possible that some values, e.g. mean F_0 values for some age groups, may have been influenced by speakers who are smokers, which may have affected the overall results of this study. Therefore, in the recording of new speech corpora, speaker-related information should be collected for the benefit of future research on the speech material. However, if a large enough number of speakers is used, such speaker-related effects may be reduced. Moreover, the age classes with relatively few speakers (40, 50 and 80) have probably yielded less reliable results than the other classes. The relatively large number of speakers used in this study, in combination with the detailed definitions of age classes, may still provide interesting data on how speech is affected by the chronological age of the speaker.

Automatic segmentation and feature extraction

This study was an attempt to use automatic methods to annotate (i.e. align) a relatively large speech corpus, and to extract features from numerous phonetic segments. The approach proved to have both advantages and drawbacks. One benefit was that it was less time-consuming to analyse the material. However, owing to the many errors in the initial alignment of the phonetic segments (especially on the phonemic and sub-phonemic levels), the speech material had to be subjected to a large number of automatic and manual alignment checks, which were rather time-consuming. In this case a small number of speakers seemed to cause the most problems. Manual reviews and correction of these files saved time and also, it is hoped, reduced the number of remaining errors considerably. Furthermore, the scripts developed to detect errors automatically were also successful, for instance in detecting creaky segments, which were hard to align.

Owing to the numerous problems with the aligner, the features extracted here may still contain values from other phonemes. There is a strong need for improved aligners, and until this need is met, studies like the present one must be prepared to deal with problems relating to sub-optimally aligned segments. Moreover, not until automatic speech recognition can be reliably used for reduced and dialectal speech will automatic acoustic analysis of phoneme segments using spontaneous speech samples be possible.

Important and less important correlates of speaker age

All features analysed in this study could be directly compared using the normalised tendency values in order to identify the most important correlates of adult chronological age (based on six concatenated Swedish words). The following general results were observed regarding how well the features from the seven feature groups correlated with speaker age (unless stated specifically, the results apply to both female and male speakers):

1. Habitual speech rate, here represented by syllables and phonemes per second and by segment duration, appears to be an excellent correlate of speaker age for both female and male speakers. The tendencies were clear and comprehensible.

2. The most significant tendency for intensity was the continuous increase in intensity range for the six concatenated words, which was as clear as the tendency for the duration of all six words.

3. Mean and median habitual F_0 were also found to be fairly good indicators of speaker age – at least better than F_0 range and F_0 SD.

Although displaying different patterns for female and male speakers, the tendencies are easy to distinguish. However, since the general decreasing F_0 tendency curves from age class 20 to 50 are followed by significant increases, F_0 cannot be used as a sole correlate of age. A 20-year-old male may very well have a mean F_0 level similar to that of a 65-year-old.

4. Jitter and shimmer did not correlate very well with speaker age.

5. Spectral energy distribution measures, including spectral tilt and spectral emphasis, did not display any general or continuous variation with speaker age, although LTAS measures varied with age in some frequency intervals.

6. Spectral noise was not found to correlate clearly with speaker age.

7. Resonance measures such as formant frequencies varied in their relation to speaker age. F_1 decreased only in some vowels, and F_2 and the higher formants demonstrated quite divergent frequency patterns. The occasional consistent variation with advancing age of these measures (decrease in F_1 and increase in F_2 in some vowel segments) did not show as clear tendencies as for speech rate and intensity range.

Segment duration and intensity range showed stronger tendencies to vary with different age classes than the other features and may, according to this study, be considered the most important correlates of speaker age.

Acoustic correlates vs. perceptual cues

Section 2.2 describes some perceptual cues suggested by human listeners as important for the task of judging an individual's age. As for the lower pitch and speech rate, these suggestions are in line with the findings of this study. The lower mean relative intensity levels confirm the reduced loudness mentioned by humans. As for voice quality, increased hoarseness has been correlated with spectral noise measures, including HNR and NHR. However, no general increases in these values were found. Duller or darker voices could be related to lower formant frequencies, which is confirmed only for F_1 in some vowels examined here. Less precise articulation is a cue to speaker age which could be related to the centring of formant frequencies observed in some elderly speakers as well as to the lower frequency of the sharp borders of the energy plateaus in voiceless fricatives. However, neither phenomenon could be confirmed in this study.

4.4.7 Conclusions

The following conclusions represent the general interpretations of the data analysed in the present study:

1. Automatic feature extraction and acoustic analysis can be used to analyse large speech data sets in relation to speaker age, and may lead to similar results as previous manual approaches in acoustic studies of speaker age.

2. The relatively most important correlates of adult speaker age are likely to be speech rate (measured as syllables or phonemes per second and as segment duration) and intensity range. Moreover, mean and median F_0 also provide consistent variation with speaker age, as do F_1 and F_2 in some vowel segments and LTAS in some frequency intervals. It is possible that these features can be used in combination with other features as correlates of speaker age. The other features analysed, i.e. jitter, shimmer and spectral noise, etc., do not seem to be as strongly related to adult speaker age.

3. The type of speech material used in acoustic analysis of speaker age is very likely to influence the results.

4.5 Summary

Numerous acoustic aspects of speech undergo significant change with ageing. Earlier studies have found age-related variation in duration, fundamental frequency, intensity, acoustic correlates of voice quality and spectral energy distribution (both phonatory and resonance). Moreover, a general increase of variation and instability, for instance in F_0 and amplitude, has been observed with increasing age. Numerous factors influence these acoustic features. Speaker-related factors include gender and physiological condition as well as cultural and social characteristics. Differences in the speech samples examined (sustained vowels, read or spontaneous speech) may also yield contradictory results. Methodological factors include differences in recording and analysis equipment and techniques as well as whether acoustic features are correlated with chronological or perceived age and whether studies are cross-sectional or longitudinal.

Despite the numerous studies conducted in this area of research, no consensus on which correlates are the most important in differentiating between speakers of different ages has been reached. Reasons for this include that there have been few previous attempts to compare features in search of the most relevant age correlates. As automatic phonetic studies are becoming

more common, identifying and rank-ordering the most important acoustic features may become a much easier task.

In order to identify the most important age correlates, an acoustic study was carried out. The speech samples consisted of 810 female and 836 male versions of six Swedish isolated words produced by 259 female and 268 male speakers. Pre-processing of the material included concatenation into files containing one version of all six words as well as automatic alignment into word, phoneme and several sub-phoneme segments.

A total of 161 acoustic features were extracted from various segments of the concatenated words. The features were divided into seven feature groups: speech rate, intensity, F_0, jitter and shimmer, spectral energy distribution, spectral noise and resonance. The features were analysed using normalised mean values and tendency diagrams, which enabled direct comparison of different features regardless of the original scaling and units. Results indicated that a continuous decrease in speech rate and a continuous increase in intensity range were the most powerful correlates of speaker age. Decreases in F_0, a substantial decrease in F_1 between the age classes 40 and 70 (and an occasional increase in F_2) in some vowels and changes in several frequency intervals in LTAS may be used in combination with other features as correlates of speaker age. Other features, including jitter, shimmer, spectral energy distribution and spectral noise, did not provide substantial age-related variation. Although several interesting results were found, more research using a larger and more varied speech material is needed in order to further clarify the relationship between acoustic features and speaker age.

Chapter 5

Data-driven formant synthesis of speaker age

5.1 Introduction

Synthetic speech is being increasingly used in human–machine communication and other applications, such as speaking and reading aids for people with communication handicaps. Although modern talking machines have obtained a high degree of intelligibility, they still cannot be considered very natural-sounding. One reason is their limited number of voice variations in terms of gender, age, emotions and other speaker-specific qualities. The lack of adequate tools for analysis and synthesis of such qualities may be one reason why they have not been implemented in speech synthesisers to a larger extent. In applications such as spoken dialogue systems and voice prostheses there is an apparent need for voice (and speaker-specific) variation. Gender and emotion have already been quite successfully simulated using synthesis. However, there have been few attempts to synthesise speaker age, although several applications would benefit from it. If voice prostheses could reflect the approximate age of the user, this would provide the vocally handicapped with more distinctive and personal voices. Also, the performance of some spoken dialogue systems might improve if the synthetic speech were adapted to the user by means of a voice which was similar in age, while other systems might benefit from having a voice of a specific age.

This chapter offers a short background on speech synthesis approaches and a brief survey of earlier research related to this theme. Furthermore, the development of a tool for the analysis and simulation of speaker age by data-driven formant synthesis is described. Results from two evaluations of the tool, carried out at different stages of the development process, are presented. The chapter ends with a discussion of the possibilities and limitations of the research tool as well as suggestions for further studies.

5.2 Speech synthesis approaches

People have attempted to generate intelligible and natural-sounding artificial speech since the late 18th century. Some approaches have been more successful than others, and the present-day speech synthesisers can be divided into three groups based on the synthesis method used: articulatory, concatenative and formant synthesis (Narayanan and Alwan, 2004). Comprehensive reviews of older and more modern speech synthesis approaches are given in Klatt (1987) as well as in Carlson and Granström (1997) and Narayanan and Alwan (2004), respectively.

This section offers a short survey of some different speech synthesis approaches, with focus on formant synthesis, GLOVE and the LF-model.

5.2.1 Articulatory synthesis

Articulatory synthesis uses a model of the human speech organs to generate synthetic speech. It is thought of as the ultimate speech synthesiser since it explicitly mimics human speech production, and may include aspects such as emotions and voice quality. With its intuitive and flexible controls, this approach is also potentially able to generate both acoustic and visual speech using the same set of articulatory parameters (e.g. jaw, lip and tongue movements). The first mechanical speech synthesisers, e.g. those built by Kratzenstein (1779) and von Kempelen (1791), were articulatory (Flanagan, 1972; Klatt, 1987; Schroeder, 1993). Although other methods have been predominating since the introduction of electronic synthesis in the 1930s, in recent times articulatory synthesis has again attracted the attention of researchers. Improvements in computer power and technology have allowed scientists to build three-dimensional models of the vocal tract using safe and non-invasive techniques, e.g. magnetic resonance imaging and electropalatography (Carlson and Granström, 1997; Engwall, 2002). However, because the modelling task is so complex and relatively little research has been devoted to this approach so far, the quality of articulatory synthesis is still much lower than that of both concatenative and formant synthesis.

5.2.2 Concatenative synthesis

Concatenative synthesis is a data-driven method which concatenates prerecorded natural speech samples (units). This method is currently the most common in both research and commercial synthesis applications (Carlson and Granström, 2005). Apart from the modelling of speech rhythm and intonation, it requires a minimal amount of phonetic knowledge and often produces highly natural-sounding synthetic speech (Narayanan and Alwan,

2004). Unit-selection or corpus-based synthesis approaches have increased naturalness further by concatenating larger speech chunks (units) of variable length (Möbius, 2000). Concatenation synthesis "inherits" the gender, age, emotion, speaking style, etc., of the speaker of the pre-recorded units. It is highly restricted in speaker characteristics and voice variation, since most of these "inherited" features are difficult to manipulate (Klatt, 1987). Although recent attempts have been made with prosodic modelling (Möbius and van Santen, 1996; Schweitzer et al., 2002; Batliner and Möbius, 2005; Schneider et al., 2006) and with databases of emotional speech (e.g. Campbell, 2000; Burkhardt, 2002; Burkhardt et al., 2005), most concatenative synthesisers employ a neutral speaking style (Engwall, 2002).

5.2.3 Formant synthesis

Formant synthesis models the acoustic speech signal produced by the vocal tract through serial or parallel connections of analogue or digital resonators, representing the first three to five formant frequencies and bandwidths, and generates speech from a set of rules and acoustic parameters. Rule-based synthesis has been frequently used by phoneticians because of its cognitive and generative approach to the speech mechanism (Pols, 1990). It implements the source–filter model, in which the acoustic speech signal is seen as a source signal (either produced at the glottis or at a constriction in the vocal tract) which is filtered by resonance from the oral and nasal cavities downstream from the source (Fant, 1960). The widespread use of the Klatt formant synthesiser (Klatt, 1980) is one example (Dutoit, 1997). Being both robust and highly flexible, this approach has successfully been used to model different voice qualities, emotions, attitudes and other voice variations (Klatt and Klatt, 1990; Carlson et al., 1991; Granström, 1992; Karlsson, 1992; Montero et al., 1998). Such experiments with "copy synthesis", which aim at generating close synthetic copies of natural speech samples, have often produced very natural formant-synthesised speech. Still, the general unnaturalness often found in rule-based formant synthesis is considered a basic disadvantage (Narayanan and Alwan, 2004). The main cause of the unnatural quality is considered to be our inability fully to control the parameters. Lately, data-driven and hybrid synthesisers have been developed in order to increase the naturalness of formant synthesised speech (Högberg, 1997; Carlson et al., 2002; Öhlin and Carlson, 2004; Carlson and Granström, 2005). This improvement, combined with its flexibility and "compressibility" into small applications, is beginning to make formant synthesis an attractive alternative to other synthesis techniques (Carlson and Granström, 2005).

In the following study, formant synthesis was chosen over other synthesis techniques because it is (1) simple and robust in operation, (2) undemanding

of computational resources, (3) flexible (as it permits easy manipulation of the voicing source characteristics), (4) closer to physical modelling than concatenation synthesis and (5) higher in synthesis quality than articulatory synthesis (Klatt, 1987; Burkhardt and Sendlmeier, 2000).

5.2.4 GLOVE

One formant synthesis system which has been used for experiments with voice variation since the late 1980s is GLOVE, an extension of the cascade formant synthesiser OVE III (Liljencrants, 1968), with an expanded Liljencrants–Fant (LF) glottal pulse model (Fant et al., 1985). Figure 5.1 shows the structure of GLOVE as used in the experiments described in this chapter.

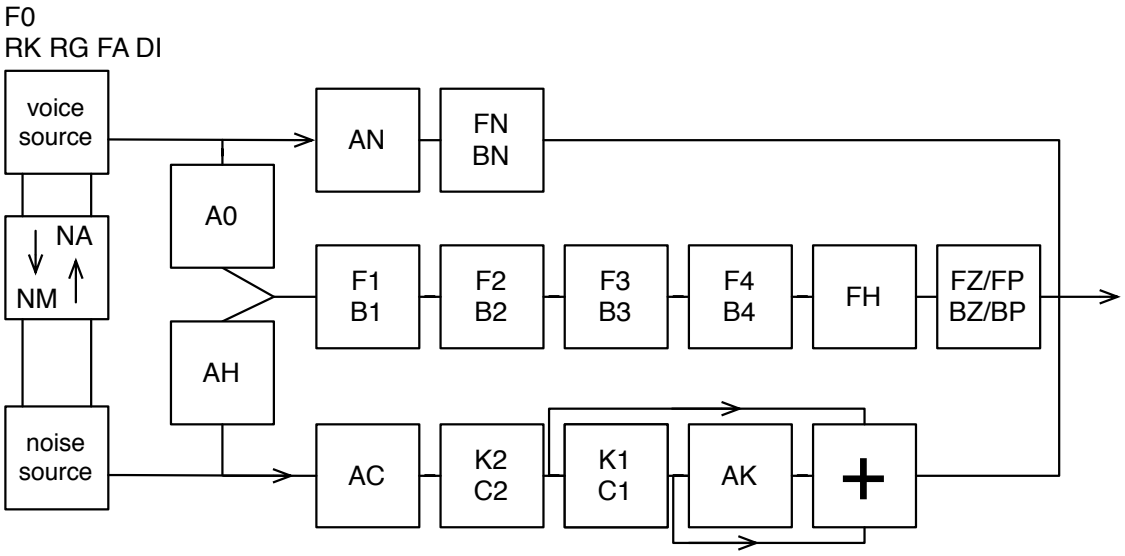


Figure 5.1: Block diagram of the GLOVE synthesiser (after Carlson et al., 1991). The parameters are described in the text

In Figure 5.1, the GLOVE parameter names are mainly constituted by two-letter symbols. The two source parameters *voice source* and *noise source* can be seen to the left in the figure. *NM* modulates the flow of the noise source, e.g. in voiced fricatives, while *NA* adds noise to the glottal source, e.g. to simulate breathy voice quality. *F0* controls the fundamental frequency, while *DI* simulates creak, laryngalisation and diplophonia. (The other three source parameters *RK*, *RG* and *FA* are defined in the next section, which describes the LF-model). Next, the signal is fed into three parallel branches containing amplitudes as well as formant frequencies and antiresonances (poles and zeros) and formant bandwidths. The signal amplitudes for the branches are controlled by the four parameters *AN*, *A0*, *AH* and *AC*. Poles and zeros for the three branches are nasal (*FN*, *BN*), glottal (*F1*–

F4, *B1–B4*, *FH*[1], *FZ/FP*[2] and *BZ/BP*[3]) and fricative noise (for fricatives and plosives, *K1–K2, C1–C2, AK*[4]). The GLOVE parameters are also summarised in Table 5.1 (p. 146).

5.2.5 The LF-model

Fant et al. (1985) introduced a mathematical model of the voicing source, which is used in GLOVE. Without going into too much technical detail, the model can be described as follows:

Figure 5.2 provides examples of glottal airflow (top) and LF-model (bottom) pulses. $U_g(t)$ represents the true glottal airflow (U_g) waveform as a function of time and ${U_g}'(t)$ is the derivative waveform of this function (i.e. the differentiated airflow), which can be divided into a sinusoidal and an exponential part and is described using four waveform parameters. The sinusoidal function models the waveform from the time of glottal opening, t_o, to the time of maximum excitation (i.e. the time of maximum negative amplitude), t_e. Then follows the exponential part from t_e to t_c, where t_c is the time of the opening of the next glottal pulse. In this part, t_p is the time of the first zero crossing of ${U_g}'(t)$, i.e. the time of the maximum flow in $U_g(t)$. $-EE$ represents the maximum negative amplitude of the main excitation, and T_a is the duration of the segment between T_e and a tangent fitted at the starting point of the return phase.

From these waveform parameters, a number of other parameters can be calculated, of which *RK*, *RG* and *FA* are used in GLOVE. *RK* represents the slope of the glottal pulse, i.e. the quotient between the time from maximum flow to maximum excitation and the time from zero to maximum flow. It can be calculated as $RK = T_n/T_p$. *RG* is the frequency which determines the opening phase (T_p) expressed as a percentage of F_0. It is calculated as the time of the glottal cycle divided by twice the time from zero to maximum flow, i.e. $RG = T_0/(2T_p)$. *FA* is the cut-off frequency above which a first-order low-pass filter adds an extra −6 dB per octave to the glottal source signal, influencing the higher frequencies of the spectrum. It can be calculated as $FA = 1/(2\pi T_a)$ or as $FA = F_0/(2\pi R_a)$, where $R_a = T_a/T_0$ corresponds to the glottal leakage. More detailed descriptions of the LF-model and its parameters are given in Fant et al. (1985), Lin (1990), Fant (1995, 1997) and Gobl (2003), and an excellent short summary can be found in Granström (1992).

[1]Parameter for higher pole correction; see Section 5.4.4, Table 5.1.

[2]Parameters for extra poles.

[3]Parameters for extra zeros.

[4]Parameter for a zero in the frication branch.

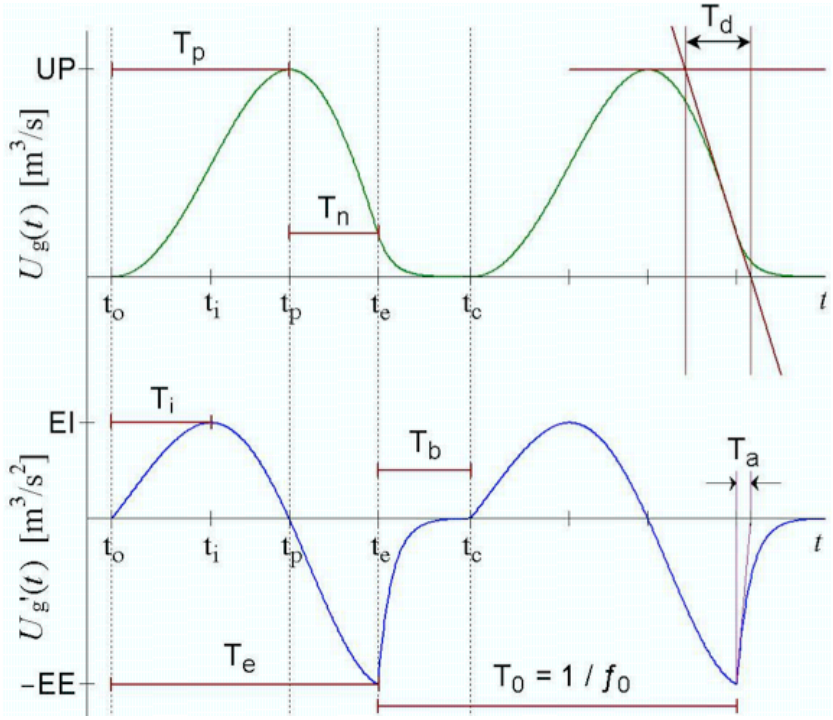


Figure 5.2: Examples of LF-model pulses (bottom) and corresponding true glottal airflow pulses (top) (source: Gobl, 2003)

5.3 Previous related studies

In this section, a brief review is given of previous research on formant synthesis (GLOVE) of speaker-specific qualities and simulation of speaker age.

5.3.1 Previous studies with GLOVE

Carlson et al. (1991) used a superimposed spectrogram to modify the rule-generated GLOVE parameters in order to adjust the target values of a number of different voices. The voice source parameters were obtained from inverse filtering. This approach was applied to produce acceptable stylised female synthesised versions of the Swedish sentence *Pia odlar blå violer* ('Pia grows blue violets').

Carlson et al. (1992) analysed and synthesised the Swedish sentence *Sommarlovet börjar sent i år* ('The summer holidays begin late this year') produced by one actor expressing four acted emotions (neutral, happy, angry, sad) using the same technique. The synthesised versions were manipulated ones where the original version of each emotion was mixed with the duration and pitch contour of the others. A listening experiment revealed that emotions use both segmental and prosodic cues, including phoneme realisation, F_0 and speech rate as well as non-phonetic supportive sounds, e.g. sighs, voice breaks and laughter.

Karlsson (1992) carried out a series of experiments on reported acoustic differences between female and male voices, with the aim of enhancing the female character of a synthesised voice. She systematically varied a number of parameters in single vowels and whole sentences using GLOVE, and carried out listening tests to evaluate the results. Although none of the acoustic differences which were tested generated a convincing female synthesised voice, the tests gave indications about areas to investigate further in order to generate more realistic synthesis of female voices.

Montero et al. (1998) developed an emotional text-to-speech system for Spanish where they implemented emotional voices in GLOVE as a small set of emotional rules. The synthesised emotions of happiness, sadness and anger obtained quite good recognition rates in the evaluation of the system, although some confusion between happiness and anger did occur.

5.3.2 Previous studies on synthesis of speaker age

Until now there have been, to the author's knowledge, very few attempts to simulate speaker age.

On his web page, Traunmüller (1996) presents resynthesised examples from experiments on the simulation of speaker age using linear predictive coding (LPC) after having manipulated speech rates and recalculated F_0 and F_1– F_3 according to values reported in earlier studies (Traunmüller, 1988; Traunmüller et al., 1989). The examples contain transformed samples of the same male speaker at the ages of 5, 12, 14 and 21 years as well as transformations of an adult female speaker into a kindergarten girl, into an adolescent girl and even into a man.

Stölten and Engstrand (2003) manipulated F_0 and duration in speech samples to simulate age differences. Original versions of the samples as well as F_0- and duration-manipulated samples (longer duration and raised F_0 for all men as well as for old women, increased duration and F_0 for young women) were used as stimuli in an age estimation test. Results showed that the age judgements of the manipulated samples were about 10 years higher than those of the original samples.

5.4 Study 6: Formant synthesis of speaker age

Analysis by synthesis is a method frequently used to study speaker characteristics and voice variation in speech (see e.g. Section 5.3.1). As already mentioned in Chapter 4, speaker age has numerous acoustic correlates and

the relationships among them appear to be rather complex. In the investigation of speaker age, this approach would be a possible method for identifying important phonetic age cues as well as for studying how important parameters interact. However, analysis by synthesis of speaker age has (to the author's knowledge) not yet been investigated.

This section describes the development of a research tool for the analysis of speaker age by data-driven formant synthesis, optimised for the generation of speaker age. Two evaluations, carried out at different stages of the development process, are presented, as well as a brief discussion of performance.

5.4.1 Purpose and aim

The main purpose of this experiment was to develop a research tool for the analysis of speaker age by data-driven formant synthesis. If parameter values could be obtained automatically from natural speech, a number of speakers with similar voices but of different ages could be used to simulate speaker age by age-weighted parameter interpolation.

The first aim of this experiment was to simulate speaker age using the research tool for analysis by synthesis. The second aim was to build the analysis tool so that it would rely mainly on automatic methods. As it turned out, these two goals sometimes led to problems concerning which goal should be regarded as more important. Compromises were sometimes inevitable. For instance, although automatically extracted formants in some cases yielded imprecise descriptions of age-related changes, they were still preferred over manually extracted values in order to meet the second aim of this study. The problems related to these matters will be discussed.

The study was further motivated by an aim to contribute to the bridging of the gap between phonetics and speech technology – more specifically by a desire to see what a phonetician could accomplish in speech synthesis with tools and methods which were relatively easy to use.

5.4.2 Questions and hypotheses

Two related research questions were asked in this experiment:

1. Would it be possible to successfully simulate speaker age using formant synthesis?

2. Could this be done using mainly automatic methods?

It was hypothesised that it would be possible to simulate speaker age using formant synthesis and linear interpolation of parameters.

5.4.3 Speech material

It might be assumed that the best speech material for the research tool would be a speech database containing recordings of the same speakers from childhood to very old age. However, such longitudinal recordings are not only rare, but also likely to vary considerably in method, such as the equipment used, the room acoustics, the storage media, etc. Moreover, in a longitudinal study, purely organic ageing effects are hard to differentiate from other variables, including emotional state, level of comfort in the recording situation, etc. (Decoster and Debruyne, 2000).

The second-best material, then, might be recordings of members of the same family with equivalent recording conditions. Therefore, four very similar and closely related non-smoking female native speakers of the same Swedish dialect were chosen to represent four different ages in the tool:

- Speaker 1: girl (aged 6)
- Speaker 2: mother (aged 36)
- Speaker 3: grandmother (aged 66)
- Speaker 4: great-grandmother (aged 91)

The speakers were recorded in their homes with a Sony portable DAT recorder TCD-D8 and a Sony tie-pin type condenser microphone ECM-T140 at 48kHz/16 bit sampling frequency. After some of the words and utterances from the recordings had been listened to, and attempts had been made to synthesise them, the elicited single Swedish isolated word *själen* ['ɧɛːlən] ('the soul') was chosen as test material, mainly because the synthesis of this word sounded acceptable (to the author) for all four speakers right from the start. Many of the other recorded words and utterances contained plosives, which were difficult to synthesise using the methods chosen for this study; extraction of parameters every 10 ms is not an optimal technique for capturing a plosive burst. Other words contained dialectal allophones (e.g. [ʁ]), which might create problems if the tool were later to be extended for the analysis of speakers of a different dialect.

An acoustic pre-analysis of the four speakers showed several known acoustic correlates of age, e.g. lowering with age of F_0 and F_1–F_2 (Linville, 2001). Figure 5.3 displays F_0 contours of the word as well as plots of F_1 and F_2 in the steady-state part of the stressed vowel [ɛː]. In the figure, the four versions of the word are lined up from the beginning of the word, i.e. from the initial [ɧ]. As expected, the youngest speaker (Speaker 1) had the highest F_0 and formant values, while the oldest speaker (Speaker 4) had the lowest values, except for the final rise of F_0. F_1 values were lower for Speaker 3 than

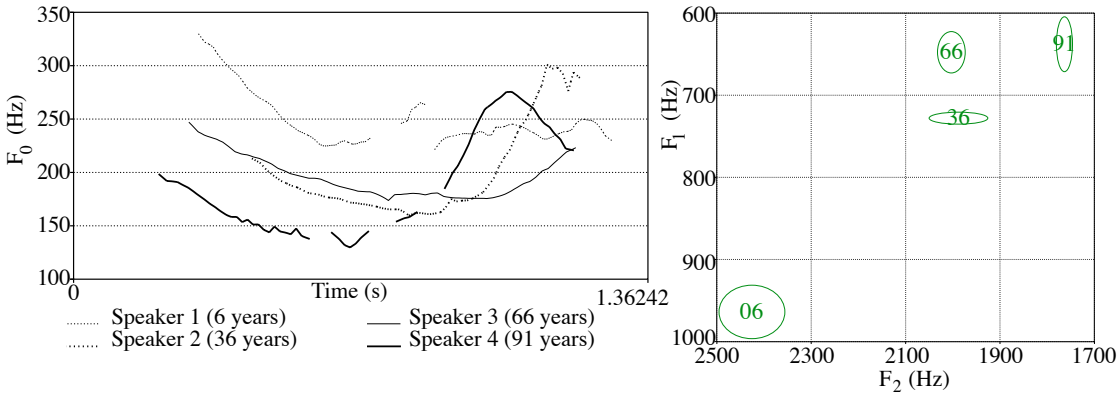


Figure 5.3: F_0 contours for ['(ɧ)ɛːlən] (left) and mean F_1–F_2 plot for the steady-state vowel [ɛ] (right) for the 6-, 36-, 66- and 91-year-old speakers

for Speaker 2, which is in line with previous studies, while F_2 values were quite similar for these two speakers. Speaker 3 displayed a somewhat higher F_0 than Speaker 2, perhaps owing to a slight difference in speaking style. Still, although additional parameters based on a larger number of speakers have to be considered in order to make more reliable attempts to simulate speaker age, the material chosen for this study seemed to include potential age-related parameters which could be examined using the research tool.

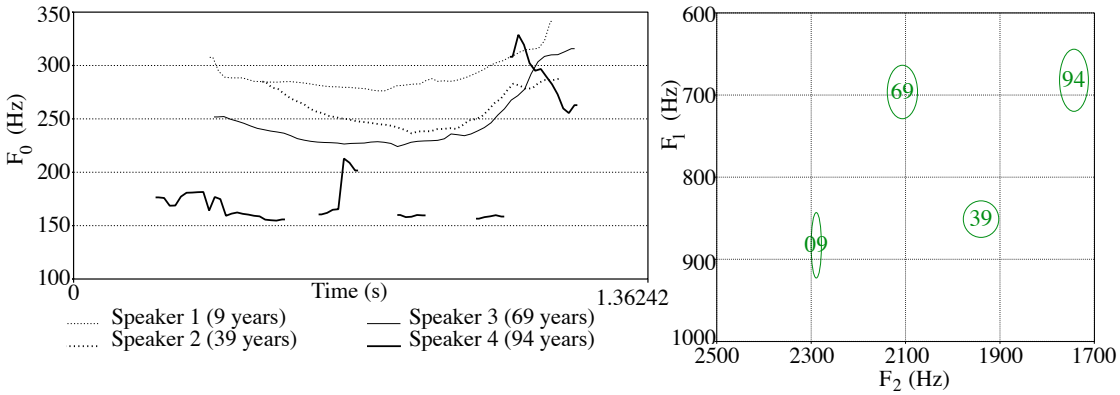


Figure 5.4: F_0 contours for ['(ɧ)ɛːlən] (left) and mean F_1–F_2 plot for the steady-state vowel [ɛ] (right) for the 9-, 39-, 69- and 94-year-old speakers

At a later stage of the study, the material was extended through the addition of new recordings of the same four speakers, made three years later under identical conditions and integrated after the first evaluation of the research tool. These new versions of ['ɧɛːlən] were subjected to the same pre-analysis as the first recordings, and they are shown in Figure 5.4. Again, the words were lined up from the initial [ɧ]. The F_0 contours for these speakers were more consistent with the reports from other studies, as they lowered with increasing age. In the F_1–F_2 plot, the same pattern can be observed as in Figure 5.3, i.e. the older speakers tended to display lower formant frequencies than the younger ones.

5.4.4 Method and procedure

The approach for developing the research tool consisted of two parts. In the first one, the focus was on generating closely resembling synthetic versions of words produced by natural speakers using automatic extraction of parameters and "data-driven analysis by synthesis", in several ways similar to the method employed by Carlson et al. (1991). However, instead of the parameter values being obtained manually at only a few points in every segment, they were extracted automatically for each time frame (every 10 ms). Several parameters were adjusted by a small set of simple rules in order to generate more natural-sounding synthesis. The second part focused on the simulation of age. Here, the first part of the method was applied specifically to speaker age in an attempt to simulate the age of natural reference speakers as well as new ages in between by age-weighted linear interpolation.

Tools

Two main tools were used to develop the research tool. For the acoustic analyses and the feature extraction, the speech analysis software Praat (Boersma and Weenink, 2005) was used. Because of its built-in functions for displaying acoustic diagrams (waveforms, spectrograms and spectra) on the computer screen to enable the user to compare the synthesised speech with the original speech, Praat also served as the main graphical user interface, and it was the program from which additional programs were called. The synthesis was performed with an internal and non-public software version of the GLOVE formant synthesis system along with the small script *Dat-convert*, which converts parameter files into readable files for GLOVE. Both GLOVE and Dat-convert were used with the kind permission of the Centre for Speech Technology (CTT) at the Royal Institute of Technology (KTH) in Stockholm. Additional software was developed using the Java and Perl programming languages.

Procedure

Pre-processing of the words included normalising for intensity (see p. 10) as well as resampling to 16 kHz. Development then proceeded in several steps, where the various parts were built individually and then integrated into the research tool. One part concerned automatic extraction of acoustic parameters from the natural speakers, which were then used to generate synthesised versions of the words. The next part comprised a set of parameter adjustment rules to produce more natural-sounding synthesis. A schematic overview of the research tool can be seen in Figure 5.5. When acceptable (but still far from perfect) synthesised versions had been obtained for all

the first recordings of the four natural speakers, the tool was used in an initial experiment to synthesise speaker age by age-weighted interpolation of parameters. In the following paragraphs, each step is explained in more detail.

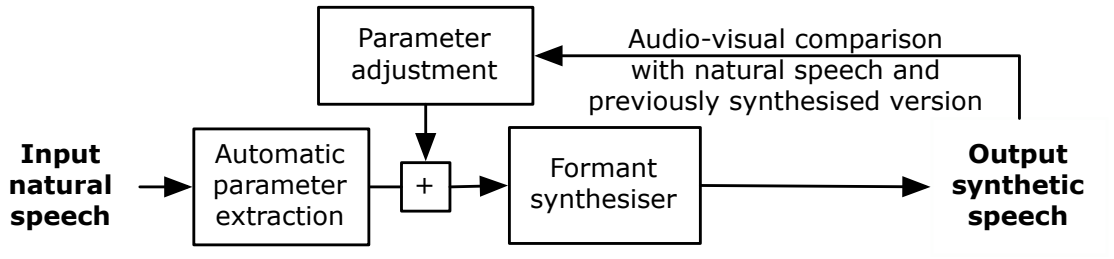


Figure 5.5: Schematic overview of the research tool

Parameter extraction

A Praat script performed several acoustic analyses on each natural word and extracted 23 parameters for each time frame of 10 ms, storing the values in a parameter file which was used as input to the GLOVE synthesiser. The parameters are briefly described in Table 5.1. Early in the development process, the script extracted only formant and F_0 parameters. In the next step, amplitude and the rest of the voice source parameters were extracted. The parameter *NA* was added at a later stage to introduce a small amount of pitch-synchronous noise so as to generate a breathy voice quality; and the *DI* parameter, which simulates creak, was the last one to be integrated.

To enable comparison of the natural speech with the synthesised versions derived from the parameter extractions, another Praat script was developed, which took the sound file of a natural word as input. First, the script called the parameter extraction script, and then it displayed waveforms and spectrograms of the original word, the resulting synthesised word and the previously synthesised version. Figure 5.6 shows such a screenshot for Speaker 2, where the user has also applied the *spectral slice* function in Praat to compare the realisations of [ʃ]. By means of auditive and visual comparison of the three displayed files, the user could easily judge the degree of similarity between the synthesised and the natural speech as well as determine whether the latest added parameter had improved the synthesis.

Parameter adjustment

The automatic parameterisation was followed by several further adjustments to improve the synthesis. A small set of adjustment rules were integrated into the parameter extraction script and varied, sometimes systematically and

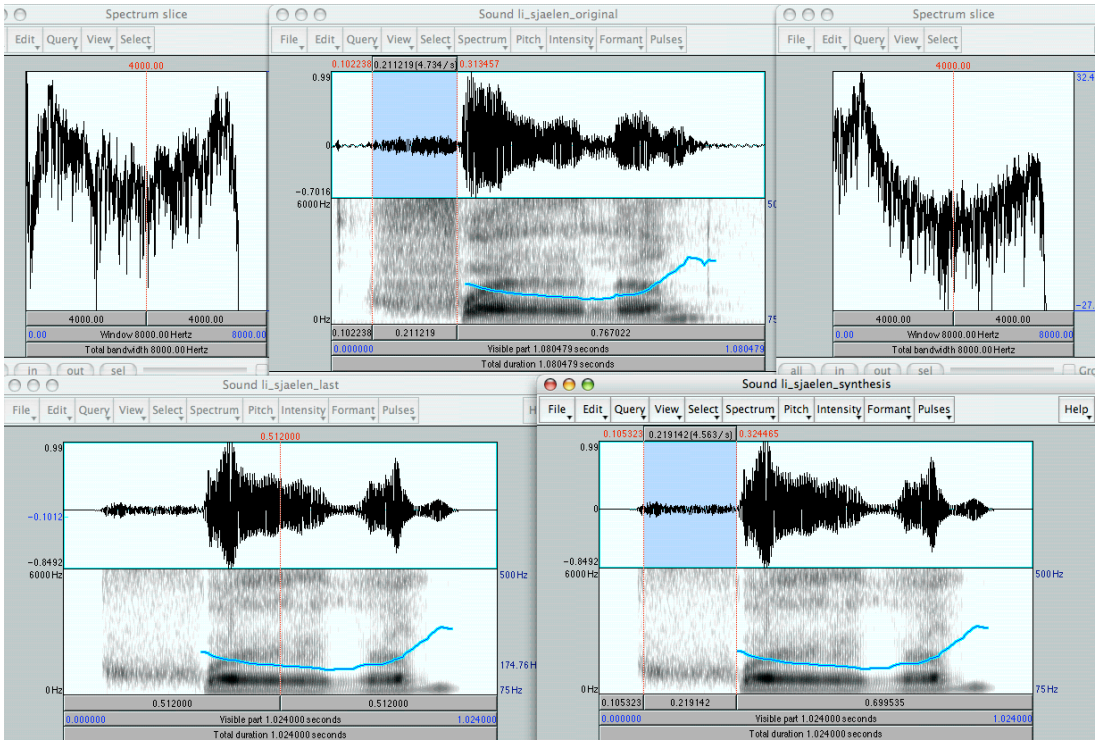


Figure 5.6: Waveforms and spectrograms for the natural word (top middle) and the previous (bottom left) and latest (bottom right) synthesised words, as well as natural (top left) and synthesised (top right) spectra for [ɧ]

sometimes using ad hoc methods, to improve similarity to the natural voices. After each adjustment, the resulting synthesis was compared with the natural word using the spectrograms and waveforms displayed in Praat. Whenever an adjustment seemed to improve the resulting synthesis, it was added to the adjustment rules. The adjustments carried out so far are described in this section.

Formants Praat's formant analysis occasionally extracted incorrect formant values. One typical error was that the program sometimes failed to identify the two lowest formants of the natural words. Instead, incorrect formant parameters were extracted, leading to unnatural phoneme qualities in the synthesised words. Early efforts to improve the formant analysis involved experiments with Praat's formant tracking function, which tries to find formants close to pre-specified reference frequencies. When this did not improve the formant analysis, the Praat script was rewritten to replace extracted outlier values with typical female formant frequencies and bandwidths for each phoneme. After some initial problems with distortion of the phoneme qualities as well as in the transitions between phonemes, relatively good results were obtained, both for vowels and for consonants. Another problem concerned the *FH* parameter, which was set to F_5. Higher formants

Table 5.1: Brief description of the GLOVE parameters

Parameter	*Description*
F1–F4, B1–B4	formant frequencies and their bandwidths
	(nasal formant and bandwidth (FN and BN) not used)
FH	higher pole correction with 3 double poles
	(FH, FH × 1.2 and FH × 1.4) with fixed bandwidths (set to F5)
	(extra formant zeros and poles (FZ/FP and BZ/BP) not used)
K1–K2, C1–C2	fricative formant frequencies and their bandwidths
	(K1 set to F2, K2 to F3, C1 to B2 and C2 to B3)
AK	zero for fricatives
F0	fundamental frequency
AC, AH	noise amplitudes for frication (AC) and aspiration (AH)
	(nasal amplitude (AN) not used)
A0	voice amplitude
RG	frequency which determines the opening phase of the glottal pulse
RK	glottal pulse skewness factor
FA	frequency above which an extra –6 dB per octave is added to the spectral tilt
NA	noise added, mixing of noise into the voice source
	(noise modulation (NM) not used)
DI	simulation of diplophonia, creak or laryngalisation

were very hard to extract because of their low amplitudes, and distortion may occur if two formants are given the same value. No attempt was made by the script to keep the frequencies of two adjacent formants apart.

F_0 Praat sometimes failed to generate reliable F_0 values, especially in the creaky segments, and it occasionally identified a non-existing F_0 in parts of the voiceless fricatives. This was solved by adjusting the arguments to Praat's pitch analysis and by introducing a second analysis specially adapted for F_0 contours below 150 Hz; this second analysis was used whenever Praat failed to find F_0 in voiced segments. Synthesis quality was improved by these adjustments, especially for Speakers 1 and 4.

Amplitude The automatically extracted amplitude values were generally too high. To avoid distorting the synthesised versions, the values were scaled down using ad hoc factors. Different factors had to be used for each segment, since the relationship among the three amplitude parameters *A0*, *AH* and *AC* depended on the specific phoneme quality. Furthermore, linear interpolation of amplitudes in the transitions between unvoiced and voiced segments helped to generate less abrupt transitions between the initial voiceless fricative and the stressed vowel.

Voice source When the voice source parameters were first added to the tool, the values were not extracted automatically. Instead, different fixed val-

ues for each segment were copied from Karlsson (1992), who had provided typical values for some female segments. A first crude attempt was made to implement the LF-model using Praat's inverse filtering function, and by calculating values for the *RG* and *RK* parameters. Owing to the highly variable trajectories, a rough linear (low-pass filtering) smoothing was obtained by letting the last value provide 90% of each value while the newly calculated value provided only 10%. The result was judged by auditive analysis to be a slight improvement on the fixed values. To be able to adjust the voice source parameters more systematically, a small Perl program was developed to enable the user to make quick adjustments of one voice source parameter for one segment at a time by a selected factor. This program was used to fine-tune some of the parameters which could not be intuitively extracted from the natural speech or which had not yet been implemented in the Praat script. Values for *FA* and *NA* were adjusted in this way. Whenever the resulting synthesised voice improved, the new parameter settings were integrated into the adjustment rules.

Parameter interpolation and synthesis of age

The basic idea for the simulation of speaker age was to use the synthesised versions of the four words to synthesise new words of other ages by age-weighted linear interpolation between two reference parameter files. To accomplish this, a small Java program was developed to calculate the weights and to perform the interpolations. An overview of the procedure is shown in Figure 5.7. For each target age provided as input by the user, the program

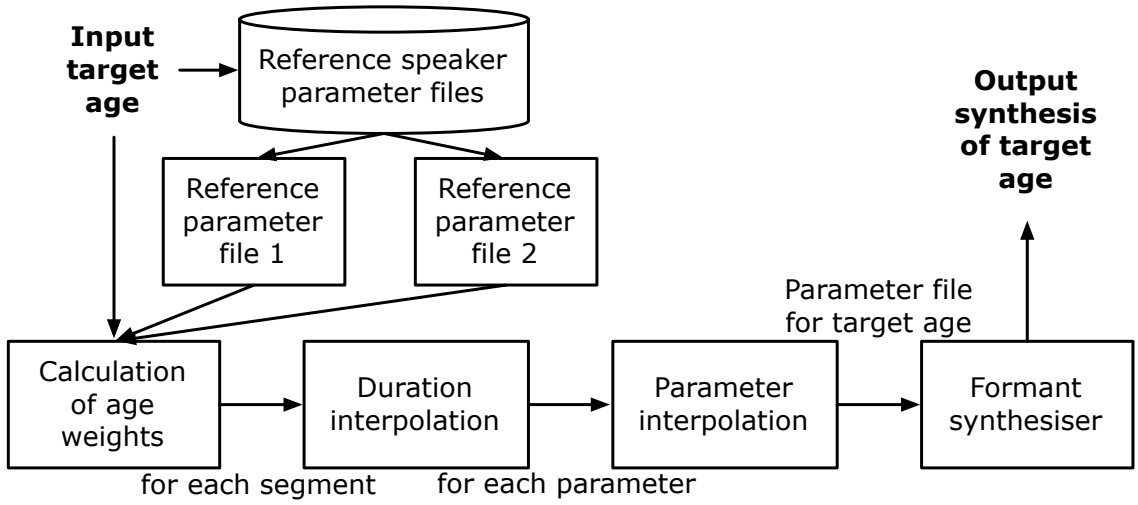


Figure 5.7: Overview of the interpolation method for simulation of age

selects the parameter files of two reference speakers (the speakers closest in age on either side of the target age) and generates a new parameter file from the interpolations between the two reference parameter files. For instance,

for the target age of 51, i.e. exactly half-way between the ages of Speaker 2 (aged 36) and Speaker 3 (aged 66), the program selects these two speakers as reference speakers, and then calculates the age weights as (in this case) 0.5 for each of them. Next, the program calculates the duration for each segment using the age weights and the durations of the reference words. If the duration of a particular segment is 100 ms for Reference Speaker 1 and 200 ms for Reference Speaker 2, the target duration for the interpolation is $200 \times 0.5 + 100 \times 0.5 = 150$ ms. All parameter values are then interpolated in the same way. Finally, the target parameter file is synthesised using GLOVE and displayed (waveform and spectrogram) in Praat along with the two input synthesised words for comparison.

Obviously, linear interpolation is a crude simplification of the age-related changes in speech, which are certainly not linear (Linville, 2001). Moreover, some age parameters may change during a certain period in life (e.g. puberty or menopause) while others remain constant. Therefore, the method described here should be considered only a tentative first attempt to simulate speaker age. Another problem concerns whether the interpolation actually simulates age or rather other speaker characteristics. For this question to be answered, further listening tests concerning speaker age must be carried out.

5.4.5 Results and first evaluation

Data-driven formant synthesis

A first look at the results of the research tool showed that although there were similarities between the natural and synthesised versions, there were differences as well. The synthesised words generally contained less energy at higher frequencies and sounded more muffled than the natural ones. Problems with formants and amplitudes also influenced results. However, the considerable resemblance between the natural and synthesised F_0 contours for all four speakers, which can be seen in Figure 5.8, shows that the automatic extraction of F_0 and duration was unproblematic. Only a few minor differences could be observed, and they can mainly be explained by the fact that the F_0 values were extracted only once every 10 ms. Also note the halved F_0 in the successfully simulated creaky parts of the synthesised versions for Speakers 1 and 4.

The durations of the natural and synthesised words are not identical in Figure 5.8, as the natural words had been extracted from the recordings with about 30–60 ms extra at the beginning and end of the words. Moreover, up to 20 ms at the beginning and end were lost in the parameter extraction, which extracted features only once every 10 ms.

Waveforms and spectrograms for the natural words as well as the resulting

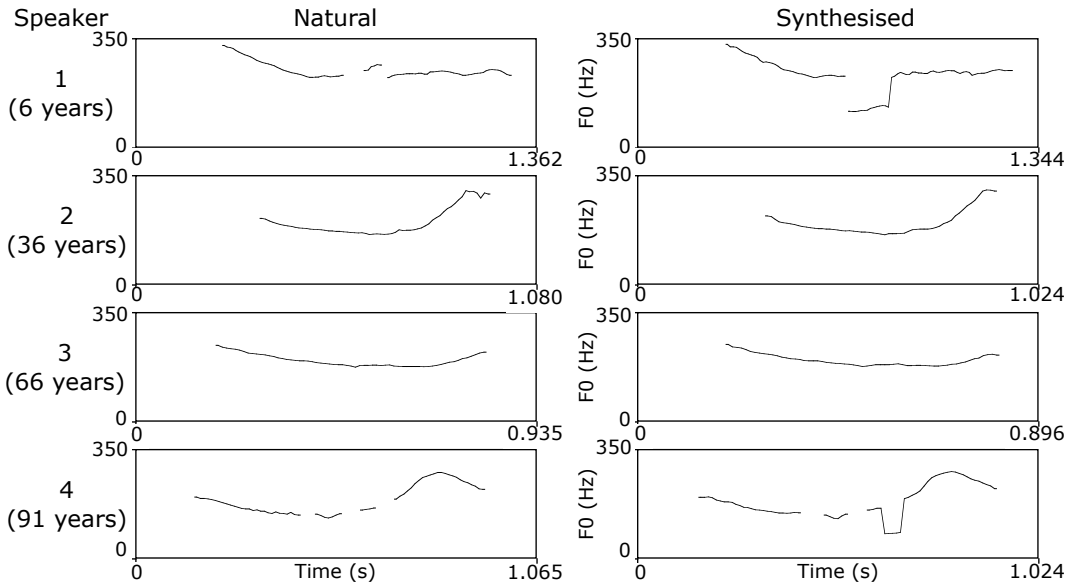


Figure 5.8: Natural and synthesised F_0 contours for the 6-, 36-, 66- and 91-year-old speakers

synthesised versions are shown in Figure 5.9. There seems to be a fair amount of resemblance between the natural and synthesised words for each speaker, a similarity corresponding to what could be heard. Formant frequencies often displayed a close resemblance to the natural words. As for voice quality, use of the *NA* and *DI* parameters contributed to a certain resemblance in the segments with breathy and creaky voices in the synthesised versions.

Shortcomings of the tool are also easy to identify. Abrupt phoneme transitions due to a lack of amplitude interpolation in the transitions (except perhaps for the initial fricative and the following vowel) can be seen in the synthesised words. For Speaker 1, differences in the formants can be observed, especially in the second syllable, where the formant analysis in Praat seems to have failed. Failure to synthesise exact copies of [ɧ] is also easily observable in the spectrograms. The waveforms also show disagreement in the amplitude envelopes, probably partly because of the ad hoc downscaling of amplitudes. Moreover, the spectrograms for the natural words generally display more energy in the higher frequencies than do the synthesised words, and auditive analysis confirmed a more muffled sound quality for the synthesised words.

Linear interpolation and synthesis of age

The resulting waveforms and spectrograms of interpolations between two reference speakers to synthesise a target age are shown in Figure 5.10 for the target ages of 20 and 80. As can be seen in the waveforms and spectrograms, the interpolation bears some resemblance to both reference speakers. For

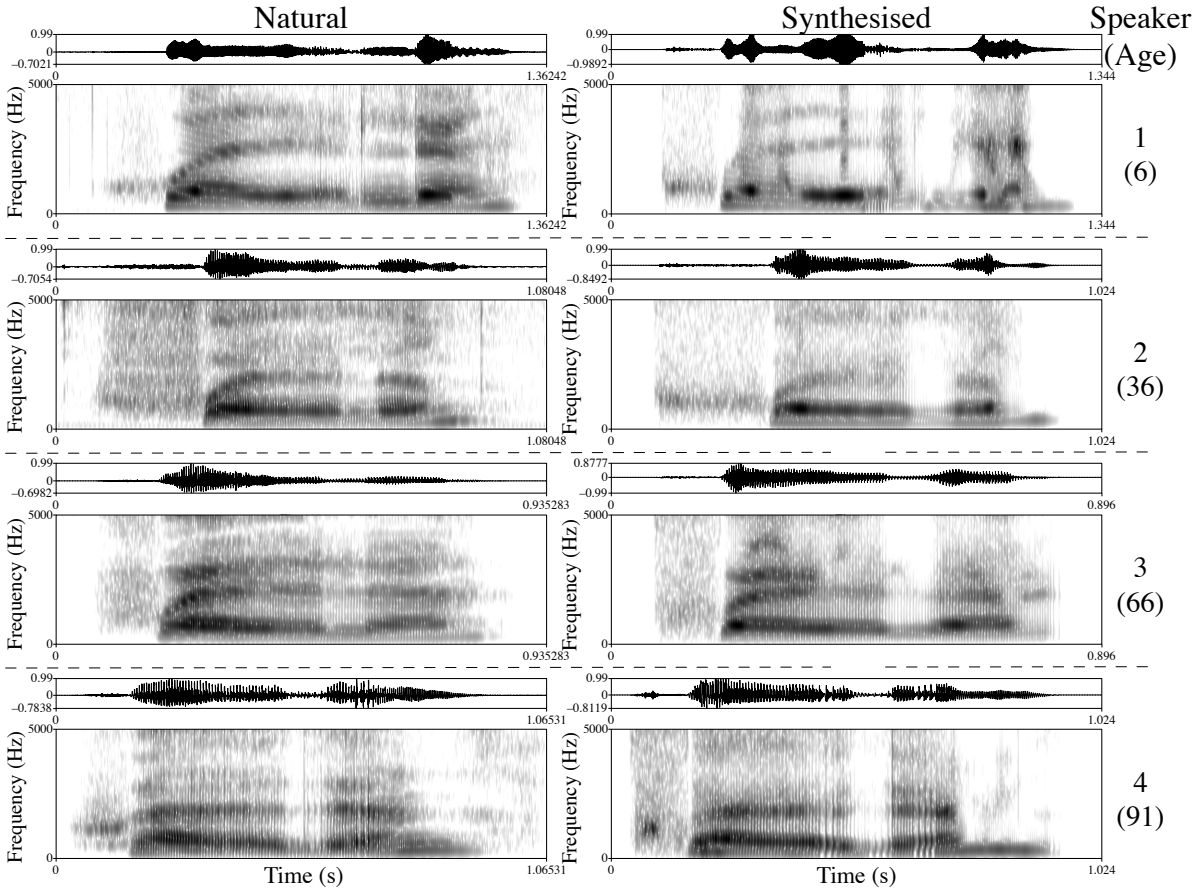


Figure 5.9: Waveforms and spectrograms for the four natural (left) and synthesised (right) words for the 6-, 36-, 66- and 91-year-old speakers

instance, the interpolation with the target age of 80 displays a small amount of creak in both vowels and in [l], where Reference Speaker 3 has no creak at all, and Reference Speaker 4 has quite a large amount. The duration of the interpolated word cannot be directly compared with the durations of the reference words, as silent segments not belonging to the word were ignored by the interpolation algorithm.

First evaluation

To evaluate the performance of the research tool at this stage, a listening test was carried out. Thirty-one students of phonetics (25 female, 6 male; mean age: 24 years) listened to the stimuli and judged direct age (in years) and naturalness (on a seven-point scale where 1 is very unnatural and 7 is very natural). The stimuli for the age estimation task consisted of the four natural and the four corresponding synthesised versions, along with age interpolations for eight decades from 10 to 80 years. The natural version of Speaker 3 was used twice to test judging consistency. In the naturalness test, only the four natural and the four synthesised words served as stimuli.

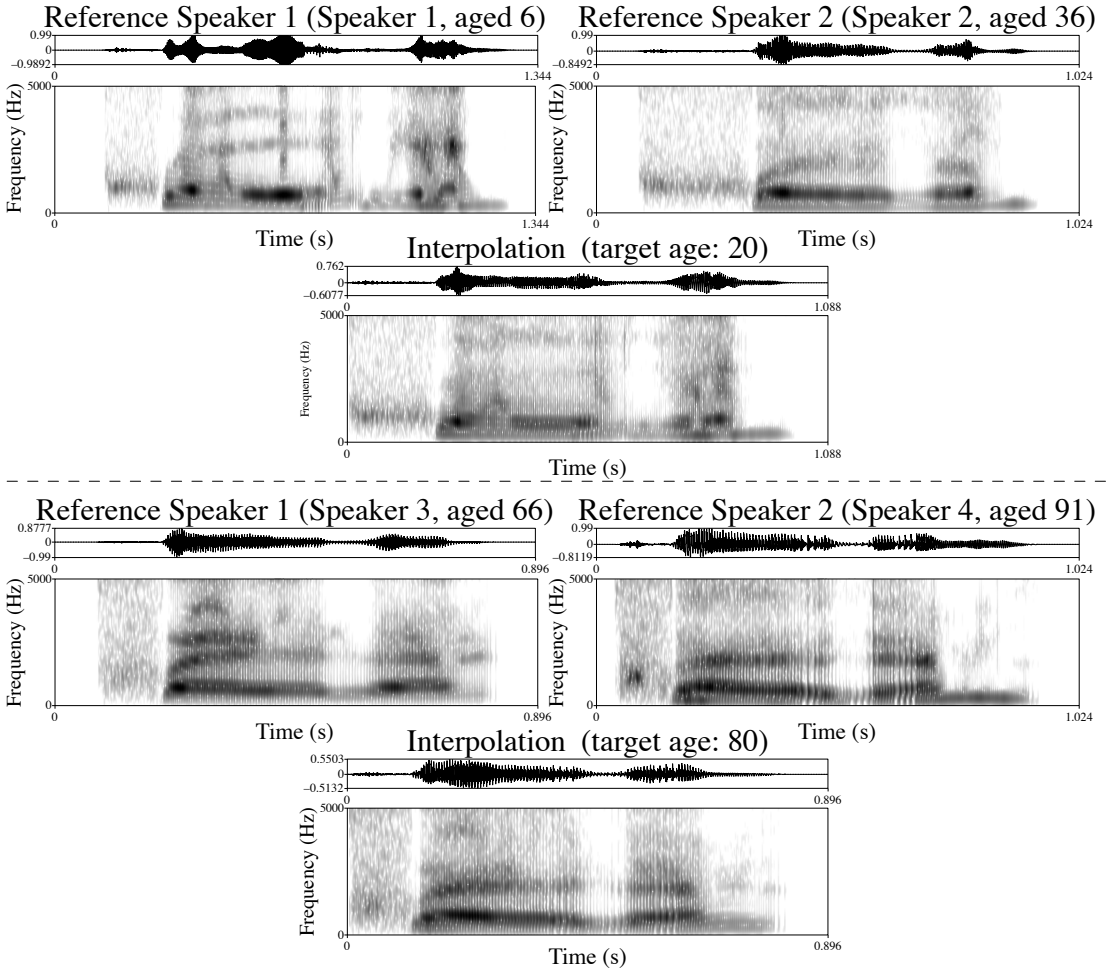


Figure 5.10: Waveforms and spectrograms for the two synthesised reference speakers and age interpolations for the target ages of 20 (top) and 80 (bottom)

Figure 5.11 shows correlations between chronological (or simulated chronological for the synthesised and interpolated versions) age (CA) and mean perceived age (PA) for the natural, synthesised and interpolated stimuli. The curves for the natural and synthesised words are not too different, although the synthesised versions were mostly judged to sound older – except for the 91-year-old voice, which was judged to be younger than its corresponding natural version. The natural and synthesised versions of Speaker 3 (66 years) were both judged to be much younger than her CA, and there was a fair amount of listener consistency for the two natural versions of this speaker: the first one had a mean PA of 35 and the second a mean PA of 36 years. The interpolation curve looks quite different from the others. Listeners often judged the interpolated stimuli to be much older than both the natural and the synthesised words. An especially striking observation was that the interpolations for 20 and 30 years were both judged to be older than 50 years. Only the age estimates for the interpolations for 10, 40 and

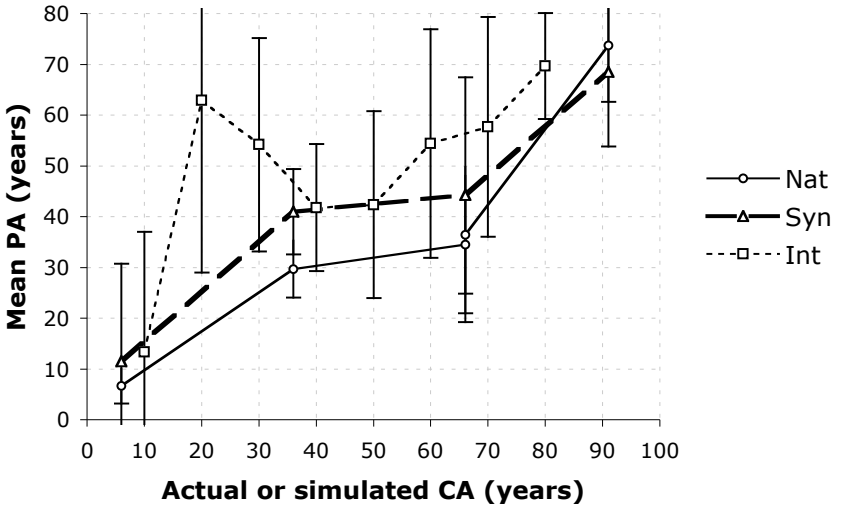


Figure 5.11: Scatter plot (points connected with lines) of chronological age (CA) and mean perceived age (PA) for the natural, synthesised and interpolated stimuli in the first evaluation (error bars = standard deviation)

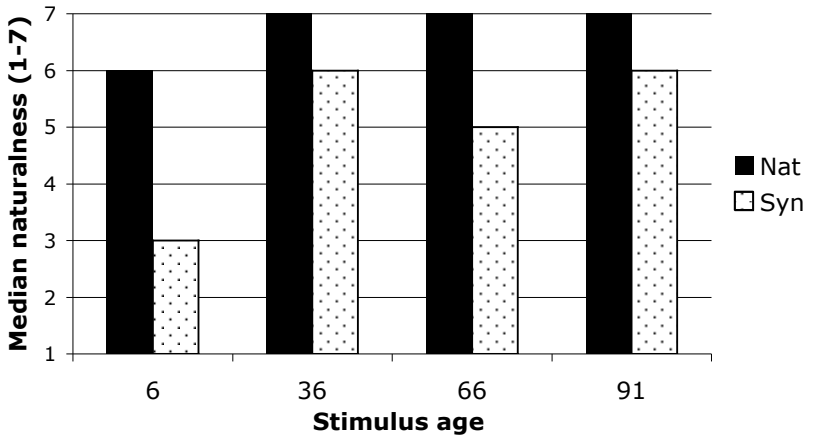


Figure 5.12: Median naturalness for the natural and synthesised stimuli in the first evaluation (seven-point scale)

50 years are close to the curve representing the synthesised words.

In the naturalness evaluation task, the eight natural and synthesised versions were used as stimuli. Figure 5.12 shows that the natural words were always judged to be more natural than the synthesised ones. Speakers 2, 3 and 4 were judged to be very natural (7), while the child speaker was found to sound somewhat less so (6). As for the synthesised words, the median perceived naturalness for the 36- and 91-year-olds was 6, the rating was a bit lower for the 66-year-old (5), and the synthesised version of the 6-year-old was judged to be the least natural, with a median naturalness of 3.

5.4.6 Improvements and second evaluation

After the first evaluation, development of the research tool continued in several steps to improve its performance. First, an attempt was made to avoid distortion. All parameters except F_0 were smoothed using linear interpolation of parameters representing segment boundaries. Next, a simple low-pass filter generated smoother parameter curves. This was used to improve the voice parameters (except F_0) as well as the formant frequencies and bandwidths. During each step, audio-visual comparison of the natural, the previously synthesised and the new synthesised words was used, for instance to decide which filter values yielded the greatest improvement. After that, the additional recordings of the four speakers (made three years after the first recordings) were included. The new words were synthesised in the same way and integrated fully into the tool, so that a total of eight speakers could be chosen as reference speakers in the synthesis of a new age by interpolation. Finally, Praat's built-in pre-emphasis filter was used to set the spectral balance of the synthesised words.

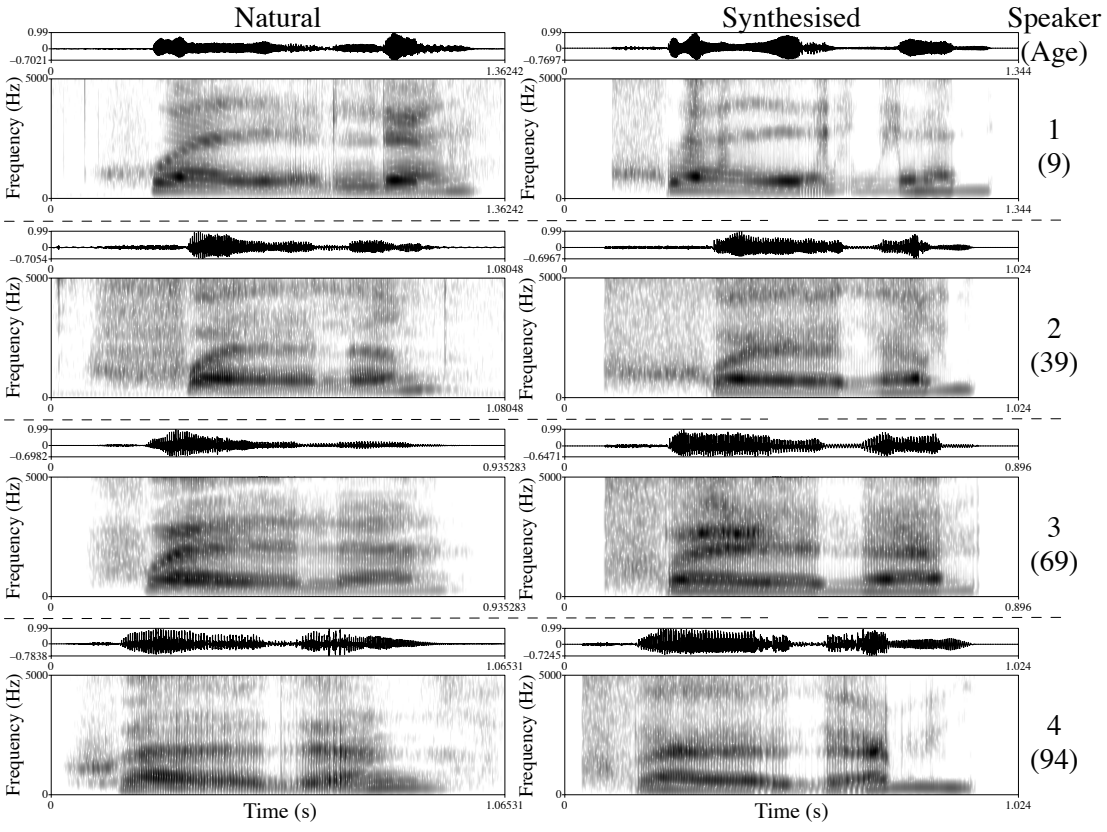


Figure 5.13: Waveforms and spectrograms for the four natural (left) and the four improved synthesised (right) words for the 6-, 36-, 66- and 91-year-old speakers

Figure 5.13 shows waveforms and spectrograms for the improved synthesised words and their corresponding natural versions for the 6-, 36-, 66- and

91-year-old speakers. Again, since the parameter extraction script extracted values only once every 10 ms, word durations differ between the natural and synthesised versions. Comparison of the spectrograms of the improved versions with those from the first results (see Figure 5.9) shows that there is now more energy at the higher frequencies than before, because of the pre-emphasis filter. Also, the synthesised formants have become more clear and prominent, and are somewhat closer to the ones in the natural words (at least for Speaker 1). One may also detect a slight improvement in the synthesised phoneme transitions, which appear to be more gradual and hence more similar to those of the natural words. However, the waveform amplitudes of the improved words, although they sometimes seem to bear a greater resemblance to the natural words, are generally higher, and therefore it is hard to visually distinguish any improvements in amplitude.

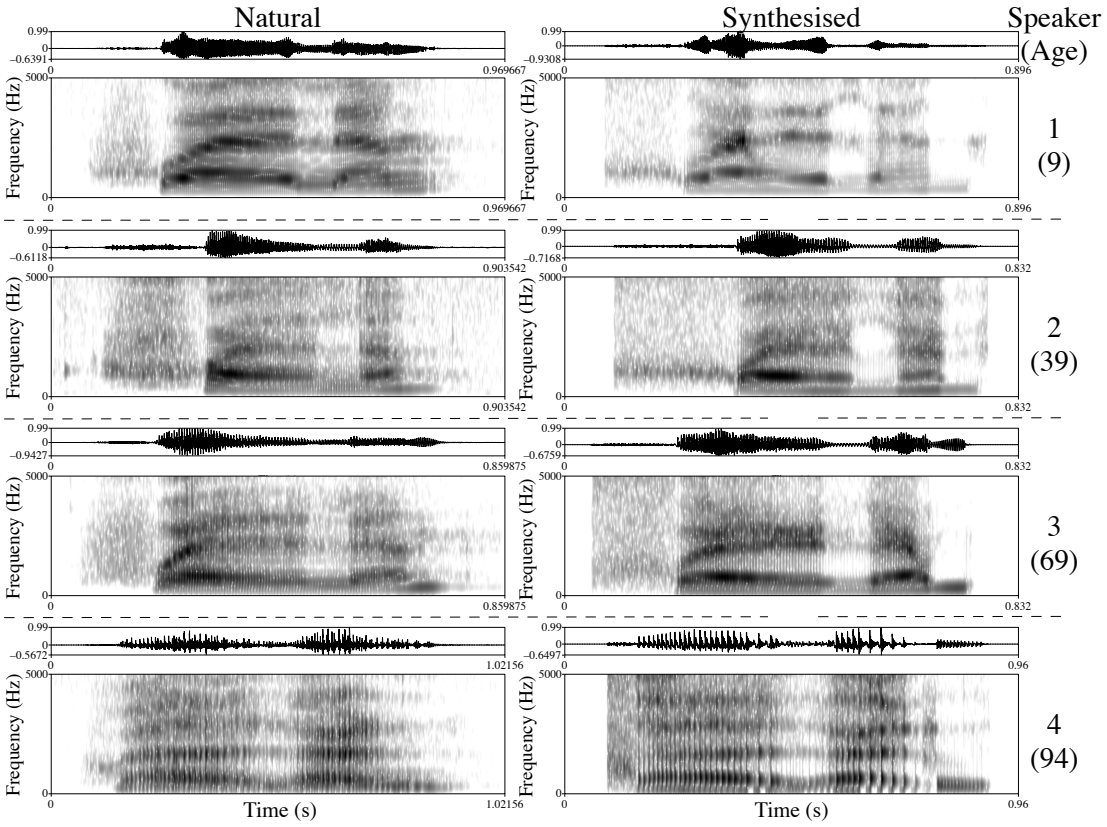


Figure 5.14: Waveforms and spectrograms for the four natural (left) and the four synthesised (right) words for the 9-, 39-, 69- and 94-year-old speakers

As for the newly added ages of 9, 39, 69 and 94 years, Figure 5.14 displays the corresponding waveforms and spectrograms, while Figure 5.15 shows F_0 values for the natural and synthesised versions. The spectrograms for the four last added versions of the word also display similarities between the natural and the synthesised words, especially in the vowels. Differences can, however, still be found in the frequency distribution of the initial [ʃ]

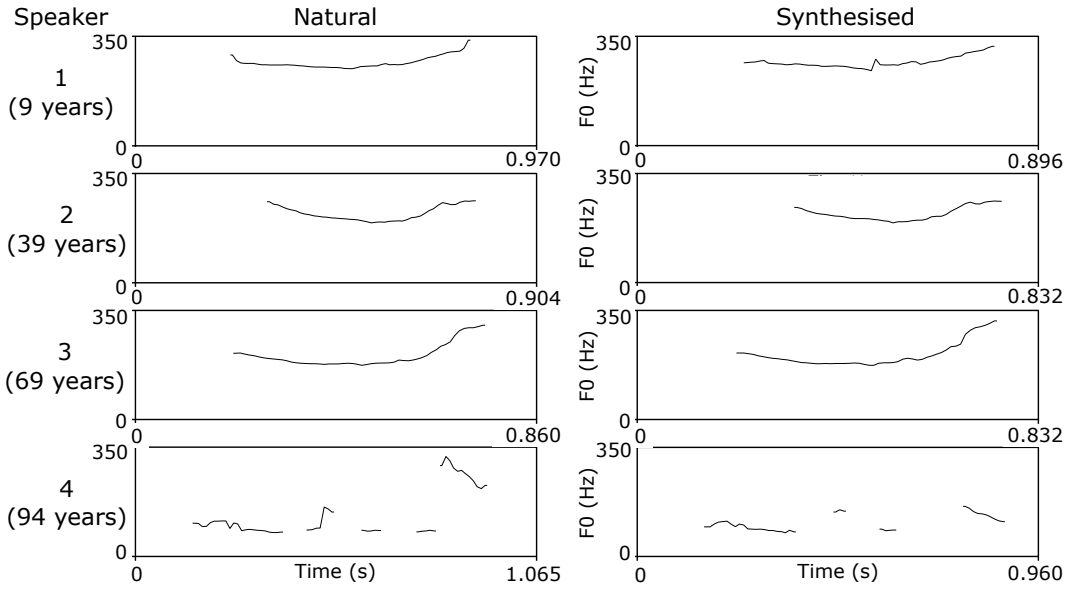


Figure 5.15: Natural and synthesised F_0 contours for the 9-, 39-, 69- and 94-year-old speakers

and the final nasal, as well as in the waveform amplitudes. Here too, the word durations of the natural and corresponding synthesised words differ (see p. 148). The F_0 contours of the synthesised versions strongly resemble the natural F_0 contours, except perhaps for the 94-year-old whose voice quality was very harsh and creaky. Praat's pitch extraction algorithm obviously had trouble identifying the correct F_0 values, even in the natural word (see the last part of the F_0 contour, which is much higher than the rest of the contour). However, the F_0 halving function of the extraction script lowered these high values, and there are practically no audible differences in pitch between the natural and synthesised versions for the 94-year-old speaker. Other minor differences include the fact that for Speaker 1, the small initial F_0 fall is not captured properly by the parameter extraction algorithm, which extracted values only once every 10 ms. There is also a minor peak in the synthesised version which cannot be observed in the natural F_0 contour.

Second evaluation

After the improvements described above, the research tool was evaluated in a second listening test, carried out essentially in the same way as the first evaluation. Eight natural and eight synthesised words of different ages were used as stimuli in the naturalness part, while the age estimation part comprised the natural and synthesised words as well as seven interpolations of ages representing even decades from 20 to 80 years. Just as in the first evaluation, the natural version of the 66-year-old was presented twice to test judging consistency. This time, 21 other students of phonetics (20 female, 1

male; mean age: 24 years) participated in the listening test. The results are shown in Figures 5.16 and 5.17.

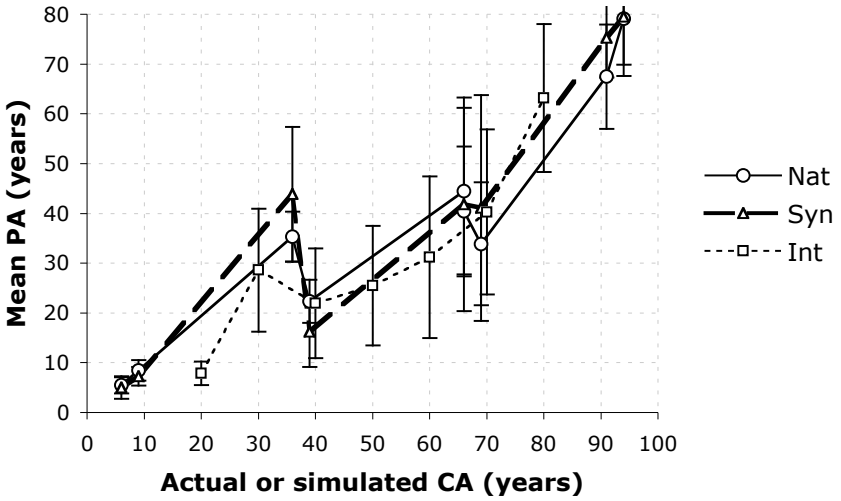


Figure 5.16: Scatter plot (points connected with lines) of chronological age (CA) and mean perceived age (PA) for natural, synthesised and interpolated stimuli in the second evaluation (error bars = standard deviation)

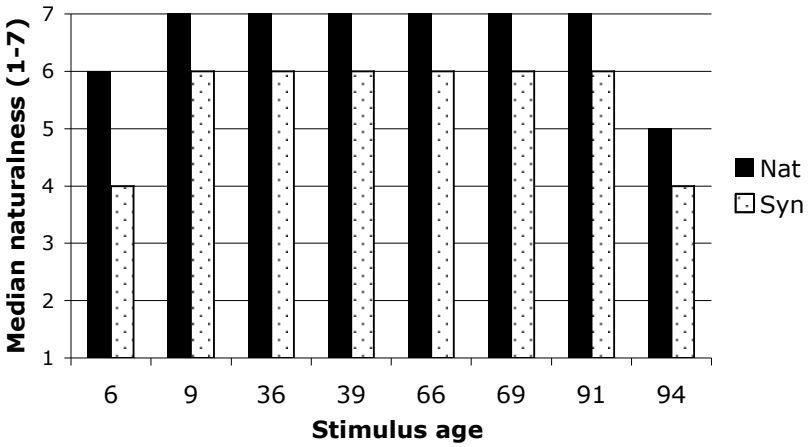


Figure 5.17: Median naturalness for the natural and synthesised stimuli in the second evaluation (seven-point scale)

In the second evaluation, the curves (i.e. the lines connecting the points in the scatter plot) for the natural and synthesised words were considerably more similar than in the first evaluation, and the interpolation curve was also much closer to the natural and synthesised ones. However, one clear exception was that the interpolation of 20 years was judged to be 9 years old, i.e. it was judged to be exactly the same age as the natural and synthesised versions of the 9-year-old speaker. Furthermore, the mean PAs of some of the stimuli – both natural and synthesised – were rather remarkable. Although the 36-year-old was judged to have a mean PA of 35 (natural) and

44 (synthesised), the 39-year-old was estimated to be much younger (natural PA: 22, synthesised PA: 16). Similar underestimations were found for the 66-year-old (natural PA: 40 and 44, synthesised PA: 42) and the 69-year-old (natural PA: 34, synthesised PA: 41). Moreover, Speaker 4 (91 and 94 years old) was consistently judged to be much younger than her CA (or simulated CA).

In the naturalness part, all the natural words were perceived as more natural than the synthesised ones. The natural stimuli were again judged as very natural (7), except for the 6-year-old (6) and the 94-year-old (5). All synthesised words except those of the 6- and 94-year-olds obtained a mean naturalness value of 6. Improvements were observed for the 6-year-old (from 3 to 4) and for the 91-year-old (from 5 to 6). The 94-year-old was perceived as a 4 on the naturalness scale.

5.4.7 Comparison of the age estimates of the natural and synthesised words

One interesting observation can be made regarding the listeners' perception of age from the natural and synthesised stimuli in the first and second evaluation. Figure 5.18 shows that the correlation between the listeners' mean PA for the natural versions and the equally aged synthesised versions was high in both the first ($r = 0.977$) and the second ($r = 0.988$) evaluation.

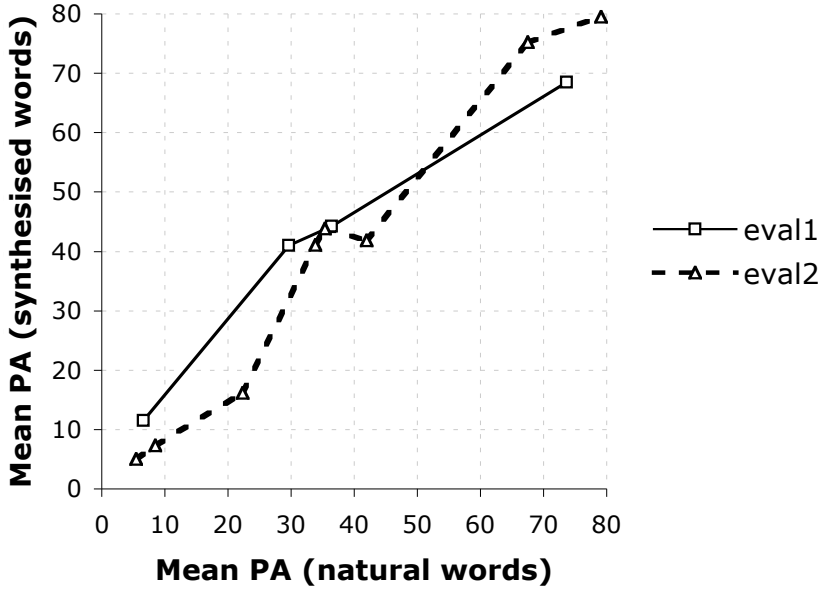


Figure 5.18: Mean perceived age (PA) for the natural words plotted against mean PA of the equally aged synthetic words in the two evaluations (points connected with lines)

5.4.8 Discussion

Because of the limited amount of time devoted to development, the research tool for analysis by synthesis of speaker age is far from complete. A number of problems remain to be solved, and several ideas for improving the tool have not yet been implemented. More time should be devoted to the fine-tuning of the parameter extraction function, the adjustment rules and the interpolation functions used for the age synthesis experiments. Some of these problems are discussed below, along with additional reflections and suggestions for the further development of the tool.

Parameter extraction and adjustment

All parameters except duration and F_0 caused problems in the extraction functions. More work is needed to reduce the number of adjustment rules. One problem with post-adjustment of automatically extracted parameters is that qualities other than the one intended may be influenced as well. A small change in one parameter could affect the whole word in an unwanted and unpredictable way, e.g. in the form of digital distortion due to increased amplitude, or a change of voice quality. Improving the parameter extraction functions as well as the parameter adjustment rules are main objectives in the further development of the tool. Special care must be taken to improve the automatic parameter extraction functions first, as the parameters should represent natural speech as much as possible. Adjustment rules that use ad hoc and fixed parameter values should be considered only as temporary solutions; they should be avoided in favour of values extracted from natural speech whenever possible.

The formant extraction function needs special attention. Formant values are notoriously difficult to measure in female and child voices because of the more widely spaced harmonics compared with those of male speakers. Perhaps better formant values could be obtained using the formant extraction algorithm developed by Öhlin and Carlson (2004). One goal is to synthesise better voiceless fricatives.

Problems with the voice source parameters may explain the muffled sound quality of the synthesis, which was adjusted by the pre-emphasis filter in Praat. Since this was a very first tentative attempt, and a crude one, at automatic extraction of the *RK* and *RG* parameters, a mistake in the implementation may have been the reason for the dullness of the resulting synthesised versions. This part of the extraction scripts needs to be further checked for errors.

Formant synthesis

Rather good versions of the natural words were obtained using data-driven formant synthesis. However, it is hard to tell whether other synthesis techniques, such as articulatory or HMM-based synthesis, may be a better choice in the future if the tool is to be expanded to handle a larger number of speakers and longer as well as more varied speech. Studies of other synthesis techniques are needed to answer that question. Still, the fact that formant synthesis did succeed in synthesising speaker age proves that formant synthesis can in fact be used in experiments with voice variation, including paralinguistic or speaker-specific variation, if controlled in the correct way.

The synthesised words were perceived as being of fairly similar age to their corresponding natural versions in most cases, and this resemblance was improved after the first evaluation. However, some of the age estimations were unexpected. For instance, the 39-, 66- and 69-year-olds were judged to be much younger than their CA. This may be because these voices were atypical of their age (see e.g. Chapter 2, p. 18). Moreover, the listeners' mean PA never exceeded 80 years for Speaker 4 (91 and 94 years old). This may not be such a strange result after all, since the listeners were university students in their mid-twenties and not likely to have had much experience with 90-year-old speakers. In fact, "people older than 70 just sound old to me" was a comment made by one of the students.

Simulation of speaker age

Although the first evaluation of the interpolated ages strongly suggested that the research tool still needed some developmental work, the second evaluation showed that it is possible to simulate speaker age using formant synthesis. The results of the evaluations also prove that the parameters chosen can be used to describe speaker age. Still, it has to be emphasised that linear interpolation is a crude simplification of the age-related changes in speech, which are far from linear. Also, some age parameters may change during one period in life while others remain constant. Better interpolation techniques will thus have to be tested in the future. One should also bear in mind that the research tool is likely to interpolate not only between two ages but also between a number of individual characteristics, even when the speakers are closely related.

An important point is that the synthesised and interpolated words were judged as closer to the corresponding natural words in perceived age than in chronological age. In future applications of age synthesis, this is not likely to be a problem, because if a natural speaker can be regarded as atypical of her or his age, so can a synthetic one. The main goal in this case must be

for a synthesised voice to be perceived as being of the target age (without necessarily being synthesised on the basis of voice input from people whose chronological age is identical with the target age). However, it would be wise to choose "age-typical" speakers as input material for simulation of age by synthesis.

Naturalness

Not once did the synthesised words reach the same high perceived naturalness as the natural ones. This shows that we still do not possess enough knowledge to generate natural-sounding formant-synthesised speech. However, the naturalness of the synthesised words was better in the second evaluation. In future experiments, other high-quality speech synthesis types should be compared with data-driven formant synthesis in order to create as natural-sounding talking machines as possible.

In both evaluations, the natural 6-year-old word was judged to sound slightly less natural than most of the others. This may have been caused by the little girl's nervousness during the first recording session, and the unnatural impression made by her voice was perhaps also reinforced by its somewhat creaky quality.

5.4.9 Conclusions and future work

As has been shown, the approach with age-weighted interpolation of parameters to simulate age can be used to generate synthetic voices of different ages, which are perceived as being of similar age to the corresponding natural ones. From these experiments, the conclusion is drawn that speaker age may be synthesised using data-driven formant synthesis. Future experiments with the research tool will involve more speakers of both genders and a larger and more varied speech material, including read and spontaneous speech. Moreover, attempts to simulate other speaker characteristics such as attitude, emotion and dialect with the research tool are also being planned.

5.5 Summary

This chapter described the development of a research tool for analysis of speaker age by data-driven formant synthesis. The tool was designed to automatically extract 23 parameters every 10 ms from one isolated word spoken by four differently aged female speakers of the same dialect and family, and then generate synthetic versions of the words. Functions for parameter adjustment as well as playback and visual display of waveforms and spectrograms of the natural and synthesised words enabled systematic change of

parameters and immediate audio-visual feedback. These functions were used to improve the resemblance between the natural and synthesised words.

Next, age-weighted linear interpolation of parameters between two reference speakers was used in a first crude attempt to simulate speaker age. An evaluation at an early stage of the development process revealed a fair amount of resemblance between the natural and synthesised words but also several weaknesses, such as distorsion owing to large differences in the adjacent formant frequency and voice parameter values. Further development included parameter smoothing to avoid distortion. Moreover, the tool was expanded with new recordings of the same speakers made three years later. A second evaluation showed improved performance: the synthesised and age-interpolated stimuli were perceived as being almost the same age as the corresponding natural ones. Despite the encouraging results, it should be kept in mind that linear interpolation is not optimal, given that ageing is far from linear. The research tool needs to be developed further with improved parameter extraction for formants, better interpolation algorithms and expansion of the tool to handle more speakers (including male voices), as well as a larger and more varied speech material.

Chapter 6

Concluding summary

The purpose of this thesis was to investigate several aspects of speaker age with the aim of gaining more phonetic knowledge about age-related variation in speech. In six different studies, various aspects of speaker age have been examined using a number of phonetic and speech technology methods. This chapter summarises the major findings and contributions of this work, following the structure of the thesis. It also recapitulates the conclusions drawn from the experiments presented in the previous chapters, and offers some ideas for future research.

6.1 Human perception of speaker age

Two studies of human perception of age were described in Chapter 2. Study 1 investigated if prosodic (F_0 and speech rate (word duration)) or non-prosodic (everything but F_0 and duration, e.g. formant frequencies) cues were more important in human perception of age. Listening tests were conducted with natural, synthesised and resynthesised versions of the isolated Swedish word *rasa* ['ʁɑ̀ːsa] ('collapse'), and F_0 and word duration were measured automatically. It was found that F_0 and duration are probably less important than non-prosodic cues. Moreover, while word duration seems equally important for both genders, F_0 seems somewhat more important in female than in male age perception. Owing to the short stimulus duration of a single disyllabic word, the findings of this study have to be supplemented by work using a larger speech material.

In Study 2, the effect of stimulus type and length (exposure time) was addressed. Listening tests with four different stimuli (one word, six words, three and ten seconds of spontaneous speech) were carried out. Results indicated that listeners seem to use different cues and listening strategies when estimating female and male age: stimulus duration (regardless of speech type) appears to be more important when judging female speakers, while

spontaneous speech (regardless of duration) seems to be more important for male speakers. Optimal stimulus durations and types for making accurate age judgements may thus depend on the gender of the speaker. Further studies systematically varying a larger number of stimulus durations and types are needed to verify this hypothesis.

6.2 Machine perception of speaker age

Chapter 4 described two attempts to estimate speaker age automatically. In Study 3, the aim was to identify acoustic-phonetic features which could be used as cues in machine estimation of age. Fifty features were automatically extracted (using a Praat script) from phoneme segments of 2,048 versions of the word *rasa* produced by 214 female and 214 male speakers. The feature vectors were then used to build automatic estimators of gender, age group (old or young) and direct chronological age (CA) using the CART (classification and regression trees) technique. It was found that the important features were mainly spectral ones, e.g. formant frequencies and HNR (harmonics-to-noise ratio). Comparison of the performance of the CART estimator of CA (mean error ± 14.45 years) with the results of the human age perception experiment described in Study 1 (mean error ± 8.89 years) showed that the human listeners were better at estimating speaker age.

Study 4 was an attempt to improve the age estimator from Study 3 by increasing the number of speakers (to 748) and features (to 78, extracted from the word as well as from the phoneme segments), and by building separate age estimators for female and male speakers. The aim was to identify the most important single feature and the best combination of features in automatic age estimation (based on *rasa*). However, only a marginal improvement in performance (mean error ± 14.07 years) was obtained compared with Study 3, perhaps because of the larger dialectal variation in the extended material. Further results showed that the most important single acoustic features in automatic age estimation were F_0 (females) and F_2 and vowel duration (males). Combinations of prosodic features were generally more successful than spectral (resonance and inverse filtered) combinations for both genders, and the CART estimators of speaker age seemed to use different features for female and male speakers.

These two studies differ from earlier attempts at automatic recognition of speaker age in several ways. Previous studies have been concerned mainly with age classification (into age groups) or correlation with age as perceived by humans, and they have used longer speech samples (≈ 5–30 s), while Studies 3 and 4 estimated chronological age (in exact years) from single words (≈ 0.4 s), which is clearly a more difficult task. Earlier studies have

also used a small number of more typical spoken language processing features, e.g. MFCC (mel frequency cepstral coefficients), whereas the data of these two experiments consisted of a large number of (more traditional) acoustic-phonetic features. Studies 3 and 4 were also the first attempts to use the CART technique in age recognition experiments.

Despite the poor accuracy of the age estimators, the two studies yielded knowledge which may be used in future attempts to build better ones. The speech material should be larger and age- and gender-balanced with minimal dialectal variation, and the speech samples should be longer, e.g. spontaneous monologues. Better methods for automatic phonetic transcription and feature extraction, e.g. better inverse filtering techniques and ways of extracting reliable values for formants, may also improve performance. In addition, different machine learning techniques (e.g. artificial neural networks or Bayesian networks) should be explored as part of the endeavour to build better gender-specific age estimators.

It should also be kept in mind that the goal when building an automatic age estimator may not be to obtain absolutely accurate estimates, but rather to obtain the same level of performance as human listeners – at least the ability to place a speaker in "her early twenties" or "his mid-seventies". Furthermore, most machine attempts to estimate speaker age so far have dealt only with adult speakers. Automatic age recognisers for children should also be explored.

It is difficult to tell whether we will have reliable objective age recognisers in the near future. However, it is likely that the results from Studies 3 and 4 can be improved substantially. The future plans of the author[1] include attempts to improve the automatic estimation of speaker age using the findings of Study 5 and a classifier based on Bayesian networks described in Müller (2005, 2006). In addition, analysis of more speakers – from the SweDia 2000 corpus as well as new recordings – using spontaneous speech samples and additional features (e.g. MFCCs) will be explored. One goal is to test whether automatic classifiers can obtain the same accuracy as human listeners.

6.3 Acoustic analysis of speaker age

Chapter 4 described an acoustic study of speaker age, motivated by the fact that the relationships among the various known acoustic correlates of speaker age had not yet been fully established by previous research. Study 5 was carried out to obtain new data about a large number of potential correlates of speaker age, and to identify the most important ones. From 1,641 versions of six Swedish concatenated words (*käke* [ˈɕɛ̀ːkə] ('jaw'), *saker*

[1]Together with Christian Müller, DFKI, Saarland University.

[ˈsɑ̀ːkəʁ] ('things'), *själen* [ˈɧɛːlən] ('the soul'), *sot* [suːt] ('soot'), *typ* [tyːp] ('type' [noun]) and *tack* [tak] ('thanks')), produced by 259 female and 268 male speakers (aged 20–89 years), 161 acoustic features were extracted from various phonetic segments. The segments were obtained using an automatic aligner. Normalised mean values for eight decade-based age groups (20–90 years) were calculated and plotted in diagrams, which enabled direct comparison of the age-related variation for different features, regardless of their original scaling and units. It was found that speech rate and intensity range seem to constitute the most important correlates of speaker age, but that F_0 as well as some spectral measures (F_1 and F_2 in certain segments and LTAS in some frequency intervals) might also be used as age correlates – at least in combination with other features. These findings were often in line with previous research on the age-related variation of various features. However, some features (e.g. jitter, shimmer and spectral noise) did not yield as much age-related variation as had been found in several earlier studies. More research is needed to clarify the effect of these features.

Study 5 has provided some interesting data on how various aspects of speech vary with the chronological age of the speaker. However, owing to the limitations in the type of speech samples used here (six concatenated isolated words) as well as in the balance regarding the ages of the speakers used, further studies (with a larger and more varied and age-balanced material) are needed to verify the findings. Moreover, the important correlates of speaker age reported in this study may be further used and analysed in the automatic estimation of speaker age, as well as in synthesis of speaker age. Thereby, better age recognisers or classifiers may be achieved, as well as better and more natural-sounding synthesis of speaker age.

The method described here could also be used to study other – linguistic as well as paralinguistic – aspects of speech. Normalised tendency diagrams may be a helpful tool in the search for important correlates of other speech characteristics.

6.4 Formant synthesis of speaker age

Chapter 5 described the development of a research tool for the analysis of speaker age by data-driven formant synthesis. Study 6 was motivated by the lack of an analysis tool enabling a large number of potential age parameters to be varied systematically and studied in detail. One aim was to successfully simulate an age between two differently aged reference speakers using automatic and relatively simple techniques, e.g. linear interpolation.

The tool was built to, first, automatically extract 23 parameters every 10 ms from the isolated Swedish word *själen* [ˈɧɛːlən] ('the soul') spoken by four

female speakers (aged 6, 36, 66 and 91 years) of the same dialect and family, and then to generate synthetic copies of the words. Functions for parameter adjustment, playback and visual display of waveforms and spectrograms of the natural and synthesised words enabled systematic parameter variation as well as immediate audio-visual feedback. These functions were used to improve the similarity between the natural and synthesised words. Next, age-weighted linear interpolation between parameters of two reference speakers was used in a first attempt to synthesise speaker age. An evaluation at an early stage of the development process showed that the research tool needed further development. It was extended with new recordings of the same four speakers made three years later, and the parameter curves were improved by smoothing. A second evaluation of the tool showed that the synthesised and age-interpolated words were now perceived as almost the same age as their corresponding natural versions, leading to the conclusion that speaker age can in fact be simulated using data-driven formant synthesis and a simple interpolation technique.

Despite the encouraging results of the second evaluation, it must be emphasised that linear interpolation is no optimal technique for simulating speaker age, since the ageing process is far from linear. Moreover, it is not clear whether the research tool actually synthesises age and not some other speaker-specific quality. The tool is thus still far from perfect, and a number of problems remain. Better techniques to extract formant and voice parameters and to interpolate between reference speakers need to be developed. The research tool should also be expanded to handle more speakers (including male voices), as well as a larger and more varied speech material (more words, phrases and utterances). If further developed, the tool could be used to analyse, model and synthesise speaker age more generally by systematic variation of parameters in order to identify the most relevant age parameters, and also to describe how these are modified to generate the impression of a certain age. This could increase phonetic knowledge of matters related to speaker age and may bring us one step closer to incorporating age into speech synthesis applications. Moreover, several other paralinguistic or speaker-specific qualities, such as emotion, attitude, dialect and personal voice quality, might be addressed using a similar approach, eventually leading to the development of more natural-sounding talking machines.

6.5 Concluding remarks

This thesis has contributed to increasing our knowledge of speaker age. The six studies presented here have investigated phonetic as well as speech technology aspects of age. They have provided some insight into human percep-

tion of age by comparing prosodic and non-prosodic cues, and have shown that the speaker's gender influences listeners' age-estimation strategies. In machine recognition of age, automatic feature extraction of a large number of features was explored, as well as a machine learning technique previously unused in age estimation. The acoustic study was the first to make direct comparisons of a large number of phonetic features in order to identify the most important correlates of speaker age. It has also been shown that speaker age can be successfully simulated using data-driven formant synthesis and age-weighted linear interpolation. However, more research is needed into age-related variation in speech, especially if we want to build reliable automatic recognisers of speaker age.

Several of the studies in this thesis addressed the question of whether prosodic or non-prosodic features constitute the most important acoustic correlates as well as perceptual cues to age. However, the various studies sometimes led to different conclusions. A possible explanation is that no single feature is more important than the others and that a combination of both prosodic and non-prosodic features is necessary to obtain good estimates of speaker age.

While either sustained vowels or read or spontaneous longer utterances have been used in most previous studies of speaker age, Studies 1, 3, 4 and 6 used single disyllabic words as speech material. One advantage of using isolated words is that they are closer to connected and natural (spontaneous) speech than sustained vowels. They were also easier to transcribe phonetically as they were pronounced in isolation and as their orthography was known. Moreover, promising results with disyllabic words had been obtained in a pilot perception study (Schötz, 2003a). Still, there were also several drawbacks of having such short speech samples. The automatic extraction of several features, e.g. formants and perturbation measures (jitter, shimmer, etc.) would have been more reliable if longer segments had been used. Longer samples of connected speech are also likely to contain more phonetic – including age-related – variation. Furthermore, human age perception generally improved when longer speech samples were used as stimuli. Thus, future studies of speaker age should include longer speech samples.

To sum up, speaker age is a very complex characteristic of speech. It leaves traces in all phonetic dimensions and it is influenced by numerous other factors, such as physiological condition. Studying it has not been a trivial task. However, this thesis has provided an excellent opportunity for trying out a number of phonetic and speech technology research methods. It is the author's hope that this work has contributed somewhat to increasing phonetic knowledge and understanding of speaker age, and that it will inspire at least a few readers to continue investigating age – and other speaker-specific qualities – in speech.

Appendix A

Features used in Studies 3, 4 and 5

In all three studies, the feature names were combined with labels representing the segment (using SAMPA transcription) from which they were extracted, e.g. *In_R_r*, *B3_A:*, *Dur_rA:sa* etc. Further explanations of some of the features are given in the Praat manual (Boersma and Weenink, 2005).

No	*Name*	*Description (Features used in* **Study 3***)*
1	Dur	segment duration
2–6	F[i]_Mean	mean F_1–F_5 (i = 1–5)
7–11	F[i]_Med	median F_1–F_5
12–16	F[i]_R	F_1–F_5 range
17–21	F[i]_R2	F_1–F_5 range (without 10% outliers)
22–26	F[i]_SD	F_1–F_5 standard deviation
27	F0_Mean	mean F0
28	F0_Med	median F0
29	F0_R	F0 range
30	F0_R2	F0 range (without 10% outliers)
31	F0_SD	F0 standard deviation
32	HNR	harmonics-to-noise ratio
33	In_Mean	mean intensity
34	In_Med	median intensity
35	In_R	intensity range
36	In_R2	intensity range (without 10% outliers)
37	In_SD	intensity standard deviation
38	Jitt_local	five jitter measurements available in Praat (see Praat manual)
39	Jitt_locabs	
40	Jitt_rap	
41	Jitt_ppq5	
42	Jitt_ddp	
43	Shi_local	six shimmer measurements available in Praat (see Praat manual)
44	Shi_loc_dB	
45	Shi_apq3	
46	Shi_apq5	
47	Shi_apq11	
48	Shi_dda	
49	SpEmph	spectral empasis
50	SpTilt	spectral tilt (of the inverse filtered signal)

No	*Name*	*Description (Features used in* **Study 4***)*
1	Dur	segment duration
2–6	F[i]_Mean	mean F_1–F_5 (i = 1–5)
7–11	F[i]_Med	median F_1–F_5
12–16	F[i]_R	F_1–F_5 range
17–21	F[i]_R2	F_1–F_5 range (without 10% outliers)
22–26	F[i]_SD	F_1–F_5 standard deviation
27–31	B[i]	formant bandwidths B_1–B_5
32–36	L[i]	formant levels L_1–L_5
37	SB0-0.5	spectral balance (peak level) at 0–0.5 kHz
38	SB0.5-1	spectral balance (peak level) at 0.5–1 kHz
39	SB1-2	spectral balance (peak level) at 1–2 kHz
40	SB2-4	spectral balance (peak level) at 2–4 kHz
41	SB4-6	spectral balance (peak level) at 4–6 kHz
42	COG	centre of gravity (see Praat manual)
43	SpSD	spectral standard deviation (see Praat manual)
44	Skew	spectral skewness (see Praat manual)
45	Kurt	spectral kurtosis (see Praat manual)
46	F0_Mean	mean F_0
47	F0_Med	median F_0
48	F0_R	F_0 range
49	F0_R2	F_0 range (without 10% outliers)
50	F0_SD	F_0 standard deviation
51	In_Mean	mean intensity
52	In_Med	median intensity
53	In_R	intensity range
54	In_R2	intensity range (without 10% outliers)
55	In_SD	intensity standard deviation
56	Jitt_local	five jitter measurements available in Praat (see Praat manual)
57	Jitt_locabs	
58	Jitt_rap	
59	Jitt_ppq5	
60	Jitt_ddp	
61	HNR	harmonics-to-noise ratio
62	Shi_local	six shimmer measurements available in Praat (see Praat manual)
63	Shi_loc_dB	
64	Shi_apq3	
65	Shi_apq5	
66	Shi_apq11	
67	Shi_dda	
68	SpEmph	spectral empasis
69	SpTilt	spectral tilt (of the inverse filtered signal)
70	ISB0-0.5	inverse filtered spectral balance (peak level) at 0–0.5 kHz
71	ISB0.5-1	inverse filtered spectral balance (peak level) at 0.5–1 kHz
72	ISB1-2	inverse filtered spectral balance (peak level) at 1–2 kHz
73	ISB2-4	inverse filtered spectral balance (peak level) at 2–4 kHz
74	ISB4-6	inverse filtered spectral balance (peak level) at 4–6 kHz
75	InCOG	inverse filtered centre of gravity (see Praat manual)
76	InSD	inverse filtered spectral standard deviation (see Praat manual)
77	InSkew	inverse filtered spectral skewness (see Praat manual)
78	InKurt	inverse filtered spectral kurtosis (see Praat manual)

No	*Name*	*Description (Features used in* **Study 5***)*
1	S_per_s	syllables per second
2	Ph_per_s	phonemes per second
3	Dur	segment duration (s)
4	In_Mean	mean intensity (dB)
5	In_Med	median intensity (dB)
6	In_R	intensity range (dB)
7	In_R2	intensity range (without 10% outliers) (dB)
8	In_SD	intensity standard deviation (dB)
9	F0_Mean_Hz	mean F_0 (Hz)
10	F0_Med_Hz	median F_0 (Hz)
11	F0_R_Hz	F_0 range (Hz)
12	F0_R2_Hz	F_0 range (without 10% outliers) (Hz)
13	F0_SD_Hz	F_0 standard deviation (Hz)
14	F0_Mean_ST	mean F_0 (semitones)
15	F0_Med_ST	median F_0 (semitones)
16	F0_R_ST	F_0 range (semitones)
17	F0_R2_ST	F_0 range (semitones)
18	F0_SD_ST	F_0 standard deviation (semitones)
19	Jitt_local	five jitter measurements available in Praat (see Praat manual)
20	Jitt_locabs	
21	Jitt_rap	
22	Jitt_ppq5	
23	Jitt_ddp	
24	Shi_local	six shimmer measurements available in Praat (see Praat manual)
25	Shi_loc_dB	
26	Shi_apq3	
27	Shi_apq5	
28	Shi_apq11	
29	Shi_dda	
30	SpTilt	spectral tilt (inverse filtered signal, dB/octave, see p. 105)
31	SpEmph	spectral empasis (dB, see p. 105)
32–36	ISB0-0.5–ISB4-6	inverse filtered spectral balance (SB), max intensity in five frequency bands: 0–0.5, 0.5–1, 1–2, 2–4 and 4–6 kHz (dB)
37–86	ltas160–ltas8000	mean amplitudes in LTAS measured from 0 to 8 kHz in 50 contiguous frequency bands of 160 Hz (dB)
87–136	pcltas160–pcltas8000	mean amplitudes in pitch-corrected LTAS (see p. 105) measured in the same bands as features 37–86 (dB)
137	HNR	harmonics-to-noise ratio (dB, see p. 105)
138	NHR	noise-to-harmonics ratio (dB, see p. 105)
139	Fr_uv	fraction of unvoiced pitch frames (see p. 105)
140	Nr_vb	number of voice breaks (see p. 105)
141	Dg_vb	degree of voice breaks (see p. 105)
142–146	F[i]_Mean (i = 1–5)	mean F_1–F_5 (Hz)
147–151	F[i]_Med	median F_1–F_5 (Hz)
152–156	F[i]_SD	F_1–F_5 standard deviation (Hz)
157–161	SB0-0.5–SB4-6	spectral balance (SB), max intensity in five frequency bands: 0–0.5, 0.5–1, 1–2, 2–4 and 4–6 kHz (dB)

Bibliography

Ainsworth, W. A. (1997). Some approaches to automatic speech recognition. In Hardcastle, W. J. and Laver, J., editors, *The Handbook of Phonetic Sciences*, pages 721–743. Blackwell Publ., Oxford.

Amerman, J. D. and Parnell, M. M. (1992). Speech timing strategies in elderly adults. *Journal of Voice*, 20:65–67.

Andersson, L.-G. (2006). Språket, Vetenskapsradion [radio programme]. Article retrieved 24 August 2006, from http://www.sr.se.

Batliner, A. and Möbius, B. (2005). Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground? In Barry, W. J. and van Dommelen, W. A., editors, *The Integration of Phonetic Knowledge in Speech Technology*, pages 21–44. Springer, Dordrecht.

Beck, J. M. (1997). Organic variation of the vocal apparatus. In Hardcastle, W. J. and Laver, J., editors, *The Handbook of Phonetic Sciences*, pages 256–297. Blackwell Publ., Oxford.

Benjamin, B. (1982). Phonological performance in gerontological speech. *Journal of Psycholinguistic Research*, 11:159–167.

Biemans, M. (2000). *Gender variation in voice quality.* PhD thesis, Netherlands Graduate School of Linguistics, Utrecht.

Black, A. and Lenzo, K. (2003). Building synthetic voices. [website]. Retrieved 8 October 2003, from http://www.festvox.org/festvox/index.html.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proc. of the Institute of Phonetic Sciences*, volume 17, pages 97–110. University of Amsterdam.

Boersma, P. and Weenink, D. (2005). Praat: doing phonetics by computer (version 4.3.04) [computer program]. Retrieved 8 March 2005, from http://www.praat.org/.

Bogert, B. P., Healy, M. J. R., and Tukey, J. W. (1963). The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. In Rosenblatt, M., editor, *Proc. of the Symposium on Time Series Analysis*, pages 209–243. Wiley, New York.

Braun, A. (1996). Age estimation by different listener groups. *Forensic Linguistics*, 3:65–73.

Braun, A. and Cerrato, L. (1999). Estimating speaker age across languages. In *Proc. of ICPhS 99*, pages 1369–1372. San Francisco, CA.

Braun, A. and Rietveld, T. (1995). The influence of smoking habits on perceived age. In *Proc. of ICPhS 95*, volume 2, pages 294–297. Stockholm.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Brown, W. S., Morris, R. J., Hollien, H., and Howell, E. (1991). Speaking fundamental frequency characteristics as a function of age and professional singing. *Journal of Voice*, 5:310–315.

Brown, W. S., Morris, R. J., and Michel, J. F. (1989). Vocal jitter in young adult and aged female voices. *Journal of Voice*, 3:113–119.

Bruce, G. (1998). *Allmän och svensk prosodi*. Praktisk Lingvistik No 16. Dept. of Linguistics and Phonetics, Lund University.

Bruce, G., Elert, C.-C., Engstrand, O., and Eriksson, A. (1999). Phonetics and phonology of the Swedish dialects – a project presentation and a database demonstrator. In *Proc. of ICPhS 99*, pages 321–324. San Francisco, CA.

Brückl, M. (2002). Altersbedingte Veränderungen von Frauenstimmen. Eine akustische und perzeptive Analyse. Master's thesis, Institute of Communications Research, Technical University of Berlin.

Brückl, M. and Sendlmeier, W. (2003). Aging female voices: An acoustic and perceptive analysis. In *Proc. of VOQUAL'03*, pages 163–168. Geneva.

Bühler, K. (1934). *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Fischer, Jena.

Burkhardt, F. (2002). *Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren*. PhD thesis, Institute of Communications Research, Technical University of Berlin.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A database of German emotional speech. In *Proc. of Interspeech 2005*. Lisbon.

Burkhardt, F. and Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In *ISCA Workshop on Speech & Emotion*, pages 151–156. Belfast.

Callan, R. (1999). *The Essence of Neural Networks*. Prentice Hall Europe.

Campbell, N. (1995). Loudness, spectral tilt, and perceived prominence in dialogues. In *Proc. of ICPhS 95*, volume 3, pages 676–679. Stockholm.

Campbell, N. (2000). Databases of emotional speech. In *Proc. of the ISCA Workshop on Speech and Emotion*, pages 34–38. Belfast.

Carlson, R. and Granström, B. (1997). Speech synthesis. In Hardcastle, W. J. and Laver, J., editors, *The Handbook of Phonetic Sciences*, pages 768–788. Blackwell Publ., Oxford.

Carlson, R. and Granström, B. (2005). Data-driven multimodal synthesis. *Speech Communication*, 47:182–193.

Carlson, R., Granström, B., and Karlsson, I. (1991). Experiments with voice modelling in speech synthesis. *Speech Communication*, 10:481–489.

Carlson, R., Granström, B., and Nord, L. (1992). Experiments with emotive speech – acted utterances and synthesized replicas. In Ohala, J. J., Nearey, T. M., Derwing, B. L., Hodge, M. M., and Wiebe, G. E., editors, *Proc. of ICSLP 92*, pages 671–674. University of Alberta.

Carlson, R., Sigvardson, T., and Sjölander, A. (2002). Data-driven formant synthesis. In *Proc. of Fonetik 2002*, volume 44, pages 121–124. TMH-QPSR.

Cavanaugh, J. C. (1999). Theories of aging in the biological, behavioural and social sciences. In Cavanaugh, J. and Whitbourne, S., editors, *Gerontology: An Interdisciplinary Perspective*, pages 346–357. Oxford University Press, Oxford.

Cerrato, L., Falcone, M., and Paoloni, A. (1998). Age estimation of telephonic voices. In *Proc. of the RLA2C conference*, pages 20–24. Avignon.

Cerrato, L., Falcone, M., and Paoloni, A. (2000). Subjective age estimation of telephonic voices. *Speech Communication*, 31, 2–3:107–113.

Clements, D. (1999). Young children and technology. Retrieved 16 August 2006, from http://www.project2061.org/publications/earlychild/online/experience/clements.htm.

Daikuhara, M. (1986). A study of compliments from a cross-cultural perspective: Japanese vs. American English. *Working Papers in Educational Linguistics*, 2:103–134.

de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research*, 36:224–266.

Decoster, W. (1998). *Akoestische kenmerken van de ouder wordene stem*. PhD thesis, Leuwen: Leuwen University Press (Summary in English).

Decoster, W. and Debruyne, F. (1997). Changes in spectral measures and voice onset time with age: A cross-sectional and a longitudinal study. *Folia Phoniatrica et Logopaedica*, 49:269–280.

Decoster, W. and Debruyne, F. (2000). Longitudinal voice changes: facts and interpretation. *Journal of Voice*, 14:184–193.

Dedivitis, R. A., Abrahão, M., Simães, M. J., Mora, O. A., and Cervantes, O. W. (2004). Aging histological changes in the cartilages of the cricoarytenoid joint. *Acta Cir Bras [serial online] 19*, Retrieved 16 August 2006, from http://www.scielo.br/acb.

Deliyski, D. and Gress, C. (1998). Intersystem reliability of MDVP for DOS and Windows 95/98. Paper presented at the 1998 Annual Convention of American Speech-Language-Hearing Association, San Antonio, Texas.

Dutoit, T. (1997). *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic, Dordrecht.

Endres, W., Bambach, W., and Flösser, G. (1971). Voice spectrograms as a function of age, voice disguise, and voice imitation. *Journal of the Acoustical Society of America*, 49:1842–1848.

Engwall, O. (2002). *Tongue Talking – Studies in Intraoral Speech Synthesis*. PhD thesis, Dept. of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm.

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

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

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

Fant, G. (2005). *Speech Acoustics and Phonetics, Selected Writings*. Kluwer Academic Publishers – Springer.

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

Ferrand, C. T. (2002). Harmonics-to-noise ratio: An index of vocal aging. *Journal of Voice*, 16:480–487.

Filipsson, M. and Bruce, G. (1997). Lukas – a preliminary report on a new Swedish speech synthesis. In *Working Papers 46*, pages 45–56. Dept. of Linguistics and Phonetics, Lund University.

Flanagan, J. (1972). *Speech Analysis, Synthesis, and Perception*. Springer, Berlin–Heidelberg–New York.

Frid, J. (2003). *Lexical and Acoustic Modelling of Swedish Prosody.* PhD thesis, Travaux de l'Institut de linguistique de Lund 45. Lund: Dept. of Linguistics and Phonetics, Lund University.

Fujisaki, H. (2004). Information, prosody, and modeling – with emphasis on tonal features of speech. In *Proc. of Speech Prosody*, pages 1–10. Nara.

Furui, S., Nakamura, M., and Iwano, K. (2006). Why is automatic recognition of spontaneous speech so difficult? In *Proc. of Symposium on Large-scale Knowledge Resources (LKR 2006)*, pages 83–90. Tokyo.

Germesin, S. (2006). Spracherkennung mit dynamisch geladenen, spezifischen Akustikmodellen. Master's thesis, Computer Science Institute, Saarland University.

Gobl, C. (2003). *The voice source in speech communication – production and perception experiments involving inverse filtering and synthesis*. PhD thesis, Dept. of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm.

González, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of Phonetics*, 32:277–287.

González, J. and Carpi, A. (2004). Early effect of smoking on voice: A multidimensional study. *Medical Science Monitor*, 10:649–656.

Granström, B. (1992). The use of speech synthesis in exploring different speaking styles. *Speech Communication*, 11:347–355.

Greenberg, S. (2001). From here to utility – melding phonetic insight with speech technology. In *Proc. of Eurospeech-2001*, pages 2485–2488.

Hartman, D. (1979). The perceptual identity and characteristics of aging in normal male adult speakers. *Journal of Communication Disorders*, 12:53–61.

Heldner, M. (2001). Spectral emphasis as an additional source of information in accent detection. In Bacchiani, M., Hirschberg, J., Litman, D., and Ostendorf, M., editors, *Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, pages 57–60. Red Bank, NJ: ISCA.

Higgins, M. B. and Saxman, J. H. (1991). A comparison of selected phonatory behaviours of healthy aged and young adults. *Journal of Speech and Hearing Research*, 13:1000–1010.

Högberg, J. (1997). Data driven formant synthesis. In *Proc. of Eurospeech 1997*, pages 565–568. Rhodes.

Hoit, J., Hixon, K., Altman, M., and Morgan, W. (1989). Speech breathing in women. *Journal of Speech and Hearing Research*, 32:353–365.

Hollien, H. (1987). Old voices: What do we really know about them? *Journal of Voice*, 1:2–13.

Hollien, H. and Shipp, T. (1972). Speaking fundamenal frequency and chronological age in males. *Journal of Speech and Hearing Research*, 15:155–159.

Horii, Y. and Ryan, W. J. (1981). Fundamental frequency characteristics and perceived age of adult male speakers. *Folia Phoniatrica (Basel)*, 33:227–233.

Huang, X., Acero, A., and Hon, H. (2001). *Spoken Language Processing*. Prentice Hall, Upper Saddle River, New Jersey.

Huntley, R., Hollien, H., and Shipp, T. (1987). Influences of listener characteristics on perceived age estimations. *Journal of Voice*, 1:49–52.

Ikeno, A., Pellom, B., Cer, D., Thornton, A., Brenier, J. M., Jurafsky, D., Ward, W., and Byrne, W. (2003). Issues in recognition of Spanish-accented

spontaneous English. In *Proc. of IEEE/ISCA Workshop on Spontaneous Speech Processing and Recognition*, Tokyo.

Jacques, R. and Rastatter, M. (1990). Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners. *Folia Phoniatrica (Basel)*, 42:118–124.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31:264–323.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. In *Proc. of IEEE 64*, pages 532–556.

Jurik, A. (1984). Ossification and calcification of the laryngeal skeleton. *Acta Radiol Diagn.*, 25:17–22.

Karlsson, I. (1992). *Analysis and Synthesis of Different Voices with Emphasis on Female Speech.* PhD thesis, Dept. of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm.

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

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

Klatt, D. (1987). Review of text-to-speech conversions for English. *Journal of the Acoustical Society of America*, 82:737–793.

Klatt, D. and Klatt, L. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87:820–857.

Kreiman, J. and Gerratt, B. R. (2005). Perception of aperiodicity in pathological voice. *Journal of the Acoustical Society of America*, 117:2201–2211.

Kreiman, J. and Papcun, G. (1985). Voice discrimination by two listener populations. Paper presented at ASA, Austin, TX.

Künzel, H. J. (2001). Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics*, 8:80–99.

Laver, J. (1980). *The Phonetic Description of Voice Quality.* Cambridge University Press, Cambridge.

Laver, J. (1991). *The gift of speech*. Edinburgh University Press, Edinburgh.

Laver, J. (1994). *Principles of phonetics*. Oxford University Press, Oxford.

Li, J., Zheng, T. F., Byrne, W., and Jurafsky, D. (2006). A dialectal Chinese speech recognition framework. *Journal of Computer Science and Technology*, 21:106–115.

Liljencrants, J. (1968). The OVE III speech synthesizer. *IEEE Trans AU-16*, No 1:137–140.

Lin, Q. (1990). *Speech production theory and articulatory speech synthesis*. PhD thesis, Dept. of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm.

Lindblad, P. (1992). *Rösten*. Studentlitteratur, Lund.

Linville, S. E. (1987). Acoustic-perceptual studies of aging voice in women. *Journal of Voice*, 1:44–48.

Linville, S. E. (1996). The sound of senescence. *Journal of Voice*, 10:190–200.

Linville, S. E. (2000). The aging voice. In Kent, R. D. and Ball, M. J., editors, *Voice Quality Measurement*, pages 359–376. Singular Thomson Learning, San Diego, CA.

Linville, S. E. (2001). *Vocal Aging*. Singular Thomson Learning, San Diego, CA.

Linville, S. E. (2002). Source characteristics of aged voice assessed from long-term average spectra. *Journal of Voice*, 16:472–479.

Linville, S. E. (2004). The aging voice. *The American Speech-Language-Hearing Association (ASHA) Leader, Oct. 19*, pages 12, 21.

Linville, S. E. and Fisher, H. (1985). Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females. *Journal of the Acoustical Society of America*, 78:40–48.

Linville, S. E. and Rens, J. (2001). Vocal tract resonance analysis of aging voice, using long-term average spectra. *Journal of Voice*, 15:323–330.

Malfrère, F. and Dutoit, T. (1997). High quality speech synthesis for phonetic speech segmentation. In *Proc. of Eurospeech*, pages 2631–2634. Rhodes.

Marasek, K. (1997). EGG & voice quality [web page]. Retrieved 31 July 2006, from http://www.ims.uni-stuttgart.de/phonetik/EGG/.

Markel, J. D. and Gray, A. H. (1976). *Linear Prediction of Speech*. Springer Verlag, Berlin.

McAllister, A., Sundberg, J., and Hibi, S. (1998). Acoustic measurements and perceptual evaluation of hoarseness in children's voices. *Logopedics Phoniatrics Vocology*, 23:27–38.

McGehee, F. (1944). An experimental study of voice recognition. *Journal of General Psychology*, 31:53–65.

Mendoza, E., Valencia, N., Munoz, J., and Trujillo, H. (1996). Differences in voice quality between men and women: use of the long-term average spectrum (LTAS). *JV*, 10:59–66.

Minematsu, N., Sekiguchi, M., and Hirose, K. (2002a). Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In *Proc. of ICASSP 2002*, pages 137–140. Orlando, FL.

Minematsu, N., Sekiguchi, M., and Hirose, K. (2002b). Performance improvement in estimating subjective agedness with prosodic features. In *Proc. of Speech Prosody 2002*, pages 507–510. Aix-en-Provence.

Minematsu, N., Yamauchi, K., and Hirose, K. (2003). Automatic estimation of perceptual age using speaker modeling techniques. In *Proc. of Eurospeech*, pages 3005–3008. Geneva.

Möbius, B. (2000). Corpus-based speech synthesis: methods and challenges. In *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS) 6 (4)*, pages 87–116. University of Stuttgart.

Möbius, B. and van Santen, J. (1996). Modeling segmental duration in German text-to-speech synthesis. In *Proc. of the International Conference on Spoken Language Processing*, pages 2395–2398. Philadelphia, PA.

Montero, L. M., Gutierrez-Arriola, J., Palazuelos, S., Enriquez, E., Aguilera, S., and Pardo, J. M. (1998). Emotional speech synthesis: From speech database to TTS. In *Proc. of ICSLP 98*, pages 923–926. Sydney.

Morris, R. J. and Brown, W. S. (1987). Age-related voice measures among adult women. *Journal of Voice*, 1:38–43.

Morris, R. J. and Brown, W. S. (1994a). Age-related differences in speech intensity among adult females. *Folia Phoniatrica (Basel)*, 46:64–69.

Morris, R. J. and Brown, W. S. (1994b). Age-related differences in speech variability among women. *Journal of Communication Disorders*, 27:49–64.

Mulac, A. and Giles, H. (1996). You're only as old as you sound: Perceived vocal age and social meanings. *Journal of Health Communication*, 8:199–216.

Müller, C. (2005). *Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht.* PhD thesis, Computer Science Institute, Saarland University.

Müller, C. (2006). Automatic recognition of speakers' age and gender on the basis of empirical studies. In *Proc. of Interspeech 2006*, Pittsburgh, PA.

Müller, C., Wittig, F., and Baus, J. (2003). Exploiting speech for recognizing elderly users to respond to their special needs. In *Proc. of Eurospeech 2003*, pages 1305–1308. Geneva.

Mupparapu, M. and Vuppalapati, A. (2005). Ossification of laryngeal cartilages on lateral cephalometric radiographs. *The Angle Orthodontist*, 75:196–201.

Murry, T. and Singh, S. (1980). Multidimensional analysis of male and female voices. *Journal of the Acoustical Society of America*, 68:1294–1300.

Mysak, E. (1959). Pitch and duration characteristics of older males. *Journal of Speech and Hearing Research*, 2:46–54.

Nagao, K. and Kewley-Port, D. (2005). The effect of language familiarity on age perception. Presented as poster at the International Research Conference on Aging and Speech Communication, Bloomington, IN.

Narayanan, S. and Alwan, A. (2004). *Text to Speech Synthesis: New Paradigms and Advances.* Prentice Hall PTR, Upper Saddle River, NJ.

Neiman, G. and Applegate, J. (1990). Accuracy of listener judgements of perceived age relative to chronological age in adults. *Folia Phoniatrica (Basel)*, 42:327–330.

Neiman, G., Kluch, R., and Shuey, E. (1983). Voice onset time in young and 70-year-old women. *Journal of Speech and Hearing Research*, 26:118–123.

Öhlin, D. and Carlson, R. (2004). Data-driven formant synthesis. In *Proc. of Fonetik 2004*, pages 160–163. Dept. of Linguistics, Stockholm University.

Öhman, S. (2001). Why current speech technology is false phonetics. In *Proc. of Fonetik 2001, Working Papers 49*, pages 180–183. Dept. of Linguistics and Phonetics, Lund University.

O'Leidhin, E. and Murphy, P. (2005). Analysis of Spectral Measures for Voiced Speech with Varying Noise and Pertubation Levels. In *Proc. of ICASSP*, volume 1, pages 869–872.

Orlikoff, R. (1990a). Heartbeat-related fundamental frequency and amplitude variation in healthy young and elderly male voices. *Journal of Voice*, 4:322–328.

Orlikoff, R. (1990b). The relationship of age and cardiovascular health to certain acoustic characteristics of male voices. *Journal of Speech and Hearing Research*, 33:450–457.

Oyer, E. and Deal, L. (1985). Temporal aspects of speech and the aging process. *Folia Phoniatrica (Basel)*, 37:109–112.

Pemberton, C., McCormack, P., and Russell, A. (1998). Have women's voices lowered across time? A cross sectional study of Australian women's voices. *Journal of Voice*, 12:208–213.

Pereira Jotz, G., Cervantes, O., Abrahao, M., Settanni, F. A. P., and de Angelis, E. C. (2002). Noise-to-harmonics ratio as an acoustic measure of voice disorders in boys. *Journal of Voice*, 16:28–31.

Petrosino, L., Colcord, R. D., Kurcz, K. B., and Yonker, R. (1993). Voice onset time of velar stop productions in aged speakers. *Journal of Perceptual and Motor Skills*, 76:83–88.

Pols, L. (1990). Does improved performance of a rule synthesizer also contribute to more phonetic knowledge? In *Proc. of the ESCA Tutorial Day on Speech Synthesis*, pages 50–54. Autrans.

Ptacek, P. H. and Sander, E. K. (1966). Age recognition from voice. *Journal of Speech and Hearing Research*, 9:273–277.

Ptacek, P. H., Sander, E. K., Maloney, W. H., and Jackson, C. C. R. (1966). Phonatory and related changes with advanced age. *Journal of Speech and Hearing Research*, 9:350–360.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kauffman, San Mateo, CA.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proc. of the IEEE, 77 (2)*, pages 257–286.

Ramig, L. (1986). Aging speech: Physiological and sociological aspects. *Language and Communication*, 6:25–34.

Ramig, L. A. (1983). Effects of physiological aging on vowel spectral noise. *Journal of Gerontology*, 38:223–225.

Ramig, L. A. and Ringel, R. L. (1983). Effects of physiological aging on selected acoustic characteristics of voice. *Journal of Speech and Hearing Research*, 26:22–30.

Ramig, L. A., Scherer, R. C., and Titze, L. R. (1985). The aging voice. In *Symposium on Care of the Professional Voice*, pages 1–10. Denver: The Voice Foundation.

Rastatter, M. and Jacques, R. (1990). Formant frequency structure of the aging male and female vocal tract. *Folia Phoniatrica (Basel)*, 42:118–124.

Rastatter, M., McGuire, R., Kalinowski, J., and Stuart, A. (1997). Formant frequency characteristics of elderly speakers in contextual speech. *Folia Phoniatrica et Logopaedica*, 49:1–8.

Riley, M. D. (1991). A statistical model for generating pronunciation networks. In *Proc. IEEE Conference on Acoustics, Speech and Signal Processing*, pages 737–740.

Ringel, R. L. and Chodzko-Zajko, W. J. (1987). Vocal indices of biological age. *Journal of Voice*, 1:31–37.

Roach, P. (2001). *Phonetics*. Oxford University Press, Oxford.

Roach, P., Stibbard, R., Osborne, J., Arnfield, S., and Setter, J. (1998). Transcription of prosodic and paralinguistic features of emotional speech. *Journal of the International Phonetic Association*, 128:83–94.

Ryan, W. J. (1972). Acoustic aspects of the aging voice. *Journal of Gerontology*, 27:256–268.

Ryan, W. J. and Burk, K. W. (1974). Perceptual and acoustic correlates of aging in the speech of males. *Journal of Communication Disorders*, 7:181–192.

Sataloff, R. T., Rosen, D. C., Hawksha, M., and Spiegel, J. R. (1997). The three ages of voice: the aging adult voice. *Journal of Voice*, 11:156–160.

Schaeffler, F. (2005). *Phonological Quantity in Swedish Dialects: Typological aspects, phonetic variation and diachronic change.* PhD thesis, Dept. of Philosophy and Linguistics, Umeå University.

Schneider, K., Lintfert, B., Dogil, G., and Möbius, B. (2006). Phonetic grounding of prosodic categories. In Sudhoff, S., Lenertov, D., Meyer, R., Pappert, S., Augurzky, P., Mleinek, I., Richter, N., and Schlieer, J., editors, *Methods in Empirical Prosody Research*, pages 335–361. Gruyter, Berlin.

Schötz, S. (2001a). A perceptual study of speaker age. In *Proc. of Fonetik 2001, Working Papers 49*, pages 136–139. Dept. of Linguistics and Phonetics, Lund University.

Schötz, S. (2001b). Röstens ålder – en auditiv och akustisk studie. Master's thesis, Dept. of Linguistics and Phonetics, Lund University.

Schötz, S. (2003a). Speaker age: A first step from analysis to synthesis. In *Proc. of ICPhS 03*, pages 2528–2588. Barcelona.

Schötz, S. (2003b). Towards synthesis of speaker age: A perceptual study with natural, synthesized and resynthesized stimuli. In *PHONUM, Reports in Phonetics 9*, pages 153–156. Umeå University.

Schötz, S. (2004). The role of F0 and duration in perception of female and male speaker age. In *Proc. of Speech Prosody*, pages 379–382. Nara.

Schötz, S. (2005a). Automatic prediction of speaker age using CART. In Svantesson, J.-O., editor, *Working Papers 51*, pages 155–168. Dept. of Linguistics and Phonetics, Lund University.

Schötz, S. (2005b). Stimulus duration and type in perception of female and male speaker age. In *Proc. of Interspeech 2005*. Lisbon.

Schötz, S. (2006a). F0 and segment duration in formant synthesis of speaker age. In *Proc. of Speech Prosody*, pages 515–518. Dresden.

Schötz, S. (2006b). Prosodic cues in human and machine estimation of female and male speaker age. In Bruce, G. and Horne, M., editors, *Nordic Prosody. Proc. of the IXth Conference, Lund, 2004*, pages 215–223. Frankfurt am Main: P. Lang.

Schroeder, M. (1993). A brief history of synthetic speech. *Speech Communication*, 13:231–237.

Schweitzer, A., Braunschweiler, N., and Morais, E. (2002). Prosody generation in the smartkom project. In *Proc. of Speech Prosody 2002*, pages 639–642. Aix-en-Provence.

Shafran, I., Riley, M., and Mohri, M. (2003). Voice signatures. In *Proc. of The 8th IEEE Automatic Speech Recognition and Understanding Workshop*. St. Thomas, U.S. Virgin Islands.

Shipp, T. and Hollien, H. (1969). Perception of the aging male voice. *Journal of Speech and Hearing Research*, 12:703–710.

Shipp, T., Qi, Y., Huntley, R., and Hollien, H. (1992). Acoustic and temporal correlates of perceived age. *Journal of Voice*, 6:211–216.

Shuey, E., Herr-McCauley, J., and Anders, M. (2004). Indices of turbulence in aging voice. Presented at the annual convention of the American Speech-Language-Hearing Association (ASHA), November 17-20, Philadelphia, PA.

Shuey, E., Herr-McCauley, J., Prohaska, C., and Martin, K. (2003). Perturbation measures and chronologic age. Presented at the annual convention of the American Speech-Language-Hearing Association (ASHA), November 13–15, Chicago, IL.

Sigvardson, T. (2002). Datadrivna metoder för parametersyntes – beskrivning av ett system och experiment med CART-analys. Master's thesis, Dept. of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm.

Sjölander, K. (2001). Automatic alignment of phonetic segments. In *Proc. of Fonetik 2001, Working Papers 49*, pages 140–143. Dept. of Linguistics and Phonetics, Lund University.

Slawinski, E. B. (1994). Acoustic correlates of [b] and [w] produced by normal young to elderly adults. *Journal of the Acoustical Society of America*, 95(4):2221–2230.

Sluijter, A. M. C. and van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100:2471–2485.

Smith, S. W. (1999). *The Scientist and Engineer's Guide to Digital Signal Processing, 2nd edition*. California Technical Publishing, San Diego.

Sprott, R. L. and Roth, G. S. (1992). Biomarkers of aging. *Generations*, 16:11–15.

Statistiska centralbyrån (2005). Från folkökning till folkminskning. Befolkningsutvecklingen i världen 1950 till 2050 . Retrieved 16 August 2006, from http://www.scb.se.

Stölten, K. and Engstrand, O. (2002). Effects of sex and age in the Arjeplog dialect: A listening test and measurements of preaspiration and vot. In *Proc. of Fonetik 2002*, volume 44, pages 29–32. TMH-QPSR.

Stölten, K. and Engstrand, O. (2003). Effects of perceived age on perceived dialect strength: A listening test using manipulations of speaking rate and F_0. In *In PHONUM, Reports in Phonetics 9*, pages 29–32. Umeå University.

Svensson, A. (2001). Ofelia – en ny syntesröst. Master's thesis, Dept. of Linguistics and Phonetics, Lund University.

Taylor, P., Caley, R., Black, A., and King, S. (1999). Edinburgh speech tools library. Retrieved 8 October 2003, from http://festvox.org/docs/speech_tools-1.2.0/book1.htm.

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

Traunmüller, H. (1996). Manipulations in speaker age and sex [website]. Retrieved 12 April 2006, from http://www.ling.su.se/staff/hartmut/manipul.htm.

Traunmüller, H. (1997). Perception of speaker sex, age, and vocal effort. In *PHONUM, Reports in Phonetics 4*, pages 183–186. Umeå University.

Traunmüller, H. (2000). Evidence for demodulation in speech perception. In *Proc. of ICSLP 00*, volume 44, pages 790–793. TMH-QPSR.

Traunmüller, H. (2005). Paralinguale Phänomene. In Ammon, U., Dittmar, N., Mattheier, K., and Trudgill, P., editors, *SOZIOLINGUISTIK Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft*, pages 653–665. Walter de Gruyter, Berlin/New York.

Traunmüller, H., Branderud, P., and Bigestans, A. (1989). Paralinguistic speech signal transformations. In *PERILUS X*, pages 47–64. Dept. of Linguistics, Stockholm University.

Traunmüller, H. and Eriksson, A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. [manuscript]. Retrieved 2 January 2006, from http://www.ling.su.se/staff/hartmut/aktupub.htm.

Traunmüller, H. and Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America*, 107:3438–3451.

Trubetzkoy, N. S. (1958). *Grundzüge der Phonologie*. Vandenhoeck & Ruprecht., Göttingen.

Wagner, A. and Braun, A. (2003). Is voice quality language-dependent? Acoustic analyses based on speakers of three different languages. In *Proc. of ICPhS 03*, pages 651–654. Barcelona.

Wang, C. C. and Huang, H. T. (2004). Voice acoustic analysis of normal Taiwanese adults. *J Chin Med Assoc*, 67:179–184.

Wang, M. Q. and Hirschberg, J. (1992). Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175–196.

Wells, J. C. (1997). Sampa computer readable phonetic alphabet. In D. Gibbon, R. M. and Winski, R., editors, *Handbook of Standards and Resources for Spoken Language Systems,*. Mouton de Gruyter, Part IV, section B. Berlin and New York.

Wells, J. C. (2006). Sampa computer readable phonetic alphabet [website]. Retrieved 17 January 2006, from http://www.phon.ucl.ac.uk/home/sampa/.

Winkler, R., Brückl, M., and Sendlmeier, W. (2003). The aging voice: an acoustic, electroglottographic and perceptive analysis of male and female voices. In *Proc. of ICPhS 03*, pages 2869–2872. Barcelona.

Winkler, R. and Sendlmeier, W. (2006). Open quotient and breathiness in aging voices – changes with increasing chronological age and its perception. *Logopedics Phoniatrics Vocology*, 31:51–56.

Xue, S. A. and Deliyski, D. (2001). Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications. *Educational Gerontology*, 21:159–168.