A Study of Acoustic Correlates of Speaker Age

Susanne Schötz[1] and Christian Müller[2]

[1] Dept. of Phonetics, Centre for Languages and Literature,
Lund University, Sweden
susanne.schotz@ling.lu.se
[2] International Computer Science Institute, Berkeley, CA
cmueller@icsi.berkeley.edu

Abstract. Speaker age is a speaker characteristic which is always present in speech. Previous studies have found numerous acoustic features which correlate with speaker age. However, few attempts have been made to establish their relative importance. This study automatically extracted 161 acoustic features from six words produced by 527 speakers of both genders, and used normalised means to directly compare the features. Segment duration and sound pressure level (SPL) range were identified as the most important acoustic correlates of speaker age.

Keywords: Speaker age, Phonetics, Acoustic analysis, Acoustic correlates.

1 Introduction

Many acoustic features of speech undergo significant change with ageing. Earlier studies have found age-related variation in duration, fundamental frequency, SPL, voice quality and spectral energy distribution (both phonatory and resonance) [1,2,3,4,5,6]. Moreover, a general increase of variability and instability, for instance in F_0 and amplitude, has been observed with increasing age.

The purpose of the present acoustic study was to use mainly automatic methods to obtain normative data of a large number of acoustic features in order to learn how they are related to speaker age, and to compare the age-related variation in the different features. Specifically, the study would investigate features in isolated words, in stressed vowels, and in voiceless fricatives and plosives. The aim was to identify the most important acoustic correlates of speaker age.

2 Questions and Hypotheses

The research questions concerned acoustic feature variation with advancing speaker age: (1) What age-related differences in features can be identified in female and male speakers? and (2) Which are the most important correlates of speaker age?

C. Müller (Ed.): Speaker Classification II, LNAI 4441, pp. 1–9, 2007.

Based on the findings of earlier studies (cf. [5]), the following hypotheses were made: **Speech rate** will generally decrease with advancing age. **SPL range** will increase for both genders. **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 occurs, followed by either an increase, decrease or no change. Male F_0 will decrease until around middle age, when an increase will follow until old age. **Jitter and shimmer** will either increase or remain stable in both women and men. **Spectral energy distribution** (spectral tilt) will generally change in some way. However, in the higher frequencies (spectral emphasis), there will be no change. **Spectral noise** will increase in women, and either increase or remain stable in men. **Resonance measures** in terms of formant frequencies will decrease in both female and male speakers.

3 Speech Material

The speech samples consisted of 810 female and 836 male versions of the six Swedish isolated words *käke* [ˈɕɛ̀ːkə] (jaw), *saker* [ˈsàːkəʁ] (things), *själen* [ˈɧɛːlən] (the soul), *sot* [suːt], *typ* [tyːp] (type (noun)) and *tack* [tak] (thanks). These words were selected because they had previously been used by the first author in a perceptual study [7] and because they contained phonemes which in a previous study had shown tendencies to contain age-related information (/p/, /t/, /k/, /s/, /ɕ/ and /ɧ/) [8]. The words were produced by 259 female and 268 male speakers, taken from the SweDia 2000 speech corpus [9] as well as from new recordings. All speakers were recorded using a Sony portable DAT recorder TCD-D8 and a Sony tie-pin type condenser microphone ECM-T140 at 48kHz/16 bit sampling frequency in a quiet home or office room. Figure 1 shows the age and gender distribution of the speakers.

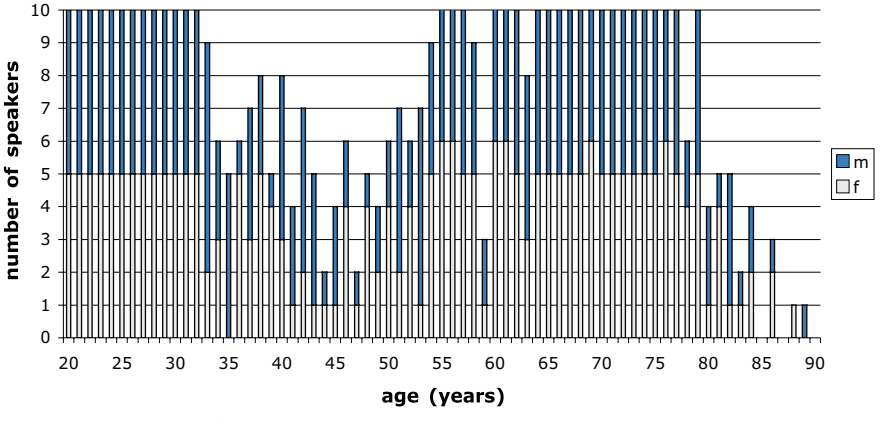

Fig. 1. Age distribution of the speakers used in this study

4 Method and Procedure

The acoustic analysis was carried out using mainly automatic methods. However, occasional manual elements were necessary in a few steps of the procedure. All words were normalised for SPL, aligned (i.e. transcribed into phoneme as well as plosive closure, VOT and aspiration segments) using several Praat [10] scripts and an automatic aligner[1]. Figure 2 shows an alignment example of the word *tack*. The alignments were checked several times using additional Praat scripts in order to detect and manually correct errors.

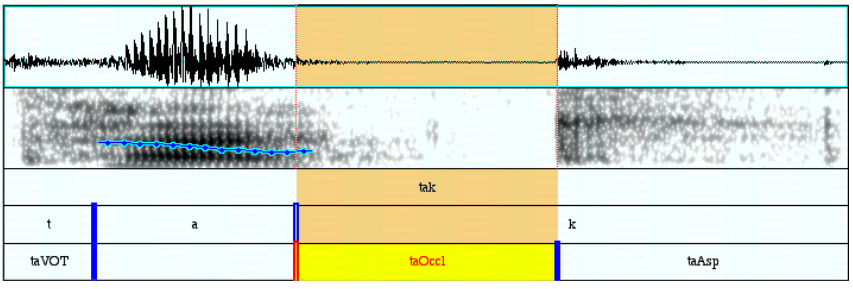

Fig. 2. Example of the word *tack*, aligned into word, phoneme, VOT, plosive closure and aspiration segments

The aligned words were concatenated; all the first word productions of a speaker were combined into one six-word sound file, all the second ones concatenated into a second file and so on until all words by all speakers had been concatenated. Figure 3 shows an example of an concatenated file.

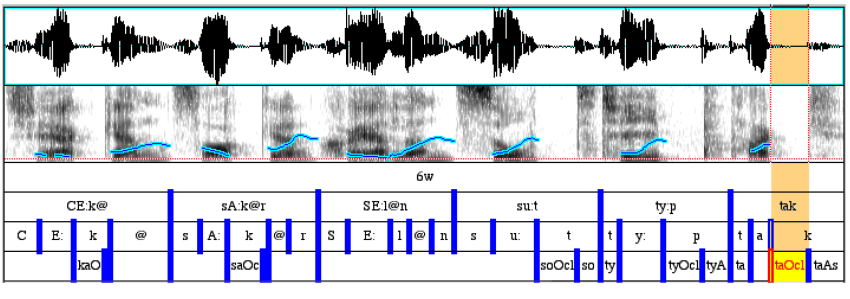

Fig. 3. Example of a concatenated file, aligned into word, phoneme, plosive closure, VOT and aspiration segments

[1] Originally developed by Johan Frid at the Department of Linguistics and Phonetics, Centre for Languages and Literature, Lund University, but further adapted and extended for this study by the first author.

A Praat script[1] extracted 161 acoustic features – divided into seven feature groups – from the concatenated words. Some features (e.g. syllables and phonemes per second, jitter and shimmer) were extracted only for all six words, while others (e.g. F_0, formant frequencies and segment duration) were extracted for several segments, including all six words and stressed vowels. Table 1 offers an overview of which segments were analysed in each feature group. Most features were extracted using the built-in functions in Praat. More detailed feature descriptions are given in [5].

Table 1. Segments analysed in each feature group (LTAS: long-term average spectra, HNR: harmonics-to-noise ratio, NHR: noise-to-harmonics ratio, sp.: spectral, str.: stressed)

Nr	*Feature group*	*Segments analysed*
1	syllables & phonemes per second	whole file
	segment duration (ms)	whole file, words, str. vowels,
2	sound pressure level (SPL) (dB)	fricatives, plosives (incl. VOT)
3	F_0 (Hz, semitones)	whole file, words, str. vowels
4	jitter, shimmer	whole file
5	sp. tilt, sp. emphasis, inverse-filtered SB, LTAS	whole file
6	HNR, NHR, other voice measures	whole file, str. vowels
7	formant frequencies (F_1–F_5)	str. vowels
	sp. balance (SB)	fricatives and plosives

The analysis was performed with m3iCAT, a toolkit especially developed for corpus analysis [11]. It was used to calculate statistical measures, and to generate tables and diagrams, which displayed the variation of a certain feature as a function of age. The speakers were divided into eight overlapping "decade-based" age classes, based on the results (mean error ±8 years) of a previous human listening test [7]. There were 14 ages in each class (except for the youngest and oldest classes): 20, aged 20–27; 30, aged 23–37; 40, aged 33–47; 50, aged 43–57; 60, aged 53–67; 70, aged 63–77; 80, aged 73–87; 90, aged 83–89.

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:

$$a_i = \frac{(v_i - mean)}{stdev} \tag{1}$$

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 shows an example of a tendency diagram where the normalisations were carried out using all speakers (top), and the same tendencies but normalised separately for each gender (bottom).

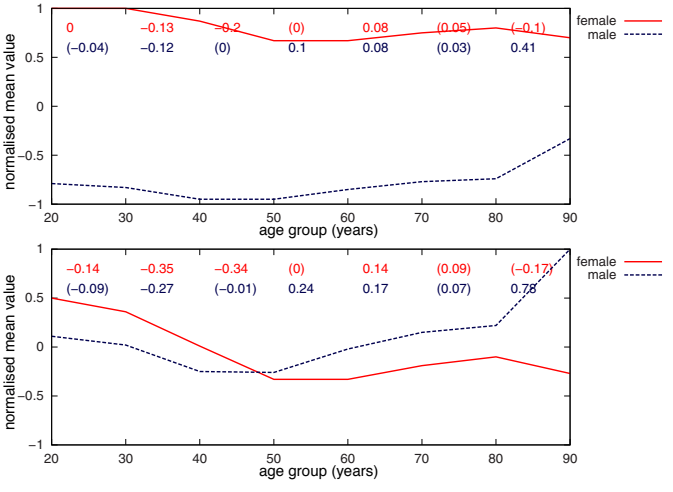

Fig. 4. Normalised tendencies for *mean F_0 (Hz) (all six words)*, 8 overlapping age classes, normalised for all speakers (top) and normalised separately for female and male speakers (bottom)

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).

5 Results

Due to the large number of features investigated, the results are presented by feature group (see Table 1). Moreover, only a few interesting results for each

feature group will be described, as it would be impossible to present the results for all features within the scope of this article. A more comprehensive presentation of the results is given in [5].

The number of syllables and phonemes per second generally decreased with increased age for both genders, while segment duration for most segments increased. The tendencies were less clear for the female than the male speakers. Figure 5 shows the results for all six words.

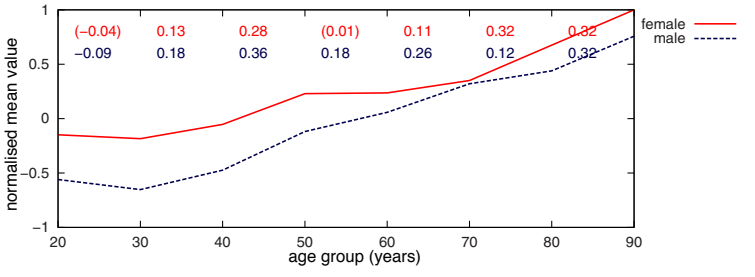

Fig. 5. Normalised tendencies for *duration (all six words)*

Average relative SPL generally either decreased slightly or remained constant with increased female and male age. The SPL range either increased or remained relatively stable with advancing age for both genders. Figure 6 shows the results for SPL range in the word *käke*. Similar tendencies were found for the other words, including the one without plosives; själen ['ʃɛːlən].

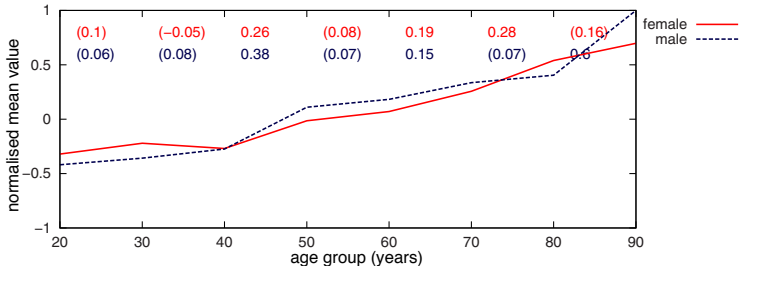

Fig. 6. Normalised tendencies for *SPL range (käke)*

Female F_0 decreased until age group 50 and then remained relatively stable. Male F_0 lowered slightly until age group 50, but then rose into old age. Due to the gender-related differences in F_0, the results for mean F_0 (Hz, all six words) are presented in Figure 7 as normalised separately for each gender to show clearer tendencies.

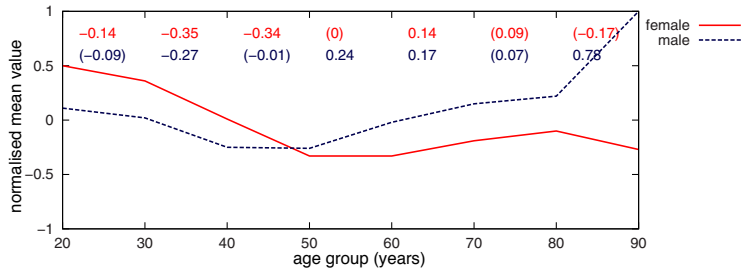

Fig. 7. Normalised (separately for each gender) tendencies for *mean F_0 (Hz, all six words)*

Although generally higher for male than female speakers, no continuous increase with age was found in either gender for jitter and shimmer. Female 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. Figure 8 shows local shimmer for all six words.

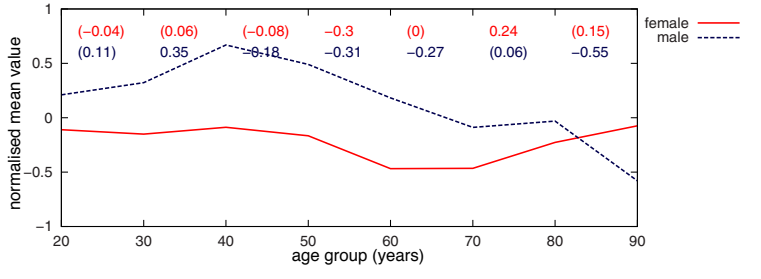

Fig. 8. Normalised tendencies for *local shimmer (all six words)*

Spectral energy distribution displayed varying results, though most measures did not change much with increased age. Figure 9 shows the LTAS amplitudes at 320 Hz, which generally increased with advancing age for both genders.

Few age-related changes were found in female NHR. Male NHR increased slightly until age class 50, where a decrease followed. Figure 10 displays the results for NHR in [ɑː].

Resonance feature results varied with segment type in both genders. F_1 decreased in [ɛː] (and in female [yː]), but remained stable in [a], [ɑː] and [uː]. F_2 was stable in [yː] and increased slightly with advancing age in [ɑː] and [ɛː] for both genders, but decreased slightly in [a] and [uː], interrupted by increases and peaks at age group 40. In F_3 and F_4, a decrease was often observed from age class 20 to 30, followed by little change or a very slight increase. Figure 11 shows normalised tendencies for F_1 and F_2 in the vowel [ɛː].

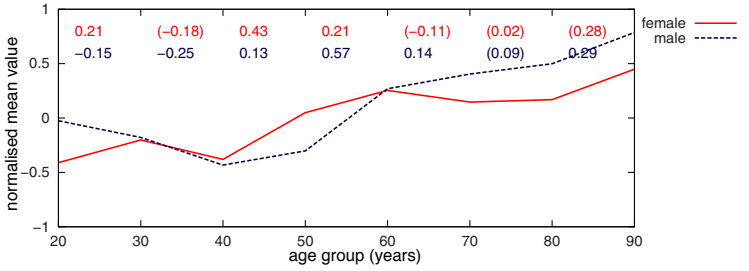

Fig. 9. Normalised tendencies for *LTAS amplitudes at 320 Hz (dB, all six words)*

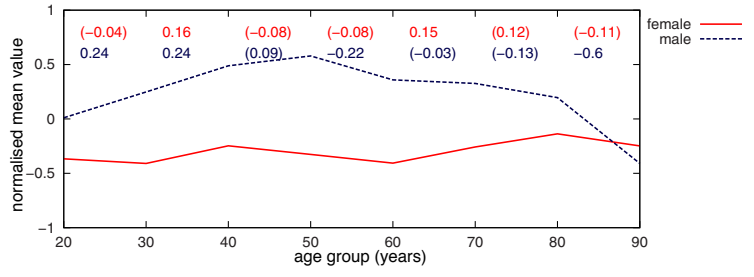

Fig. 10. Normalised tendencies for *NHR ([aː])*

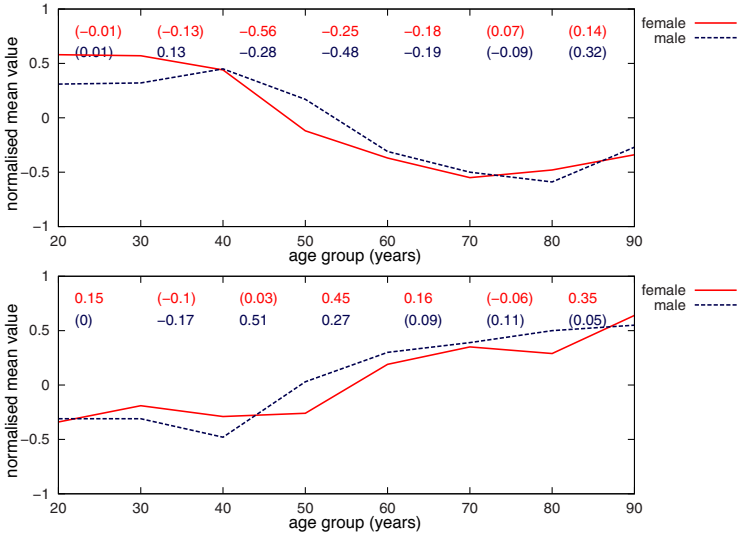

Fig. 11. Normalised (separately for each gender) tendencies for *mean* F_1 (top) *and* F_2 (bottom) (*[ɛː] in the word själen ['ɧɛːlən]*)

6 Discussion and Conclusions

Many of the hypotheses were confirmed. However, some were contradicted. Possible explanations for this include differences in the speech material compared to previous studies and the fact that mainly automatic methods were used in this study. Still, the study provided some interesting results which may be used when building automatic estimators of speaker age: (1) Automatic methods can be used to analyse large speech data sets in relation to speaker age, and may yield similar results as manual studies, (2) The relatively most important correlates of adult speaker age seem to be speech rate and SPL range. F_0 also may provide consistent variation with speaker age, as may F_1, F_2 and LTAS in some segments and frequency intervals. These features may be used in combination with other features as cues to speaker age and (3) The type of speech material used in acoustic analysis of speaker age is very likely to influence the results.

These findings will be used in future studies to improve the automatic classification of speaker age.

References

1. Ryan, W.J.: Acoustic aspects of the aging voice. Journal of Gerontology 27, 256–268 (1972)
2. Amerman, J.D., Parnell, M.M.: Speech timing strategies in elderly adults. Journal of Voice 20, 65–67 (1992)
3. Xue, S.A., Deliyski, D.: Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications. Educational Gerontology 21, 159–168 (2001)
4. Linville, S.E.: Vocal Aging. Singular Thomson Learning, San Diego (2001)
5. Schötz, S.: Perception, Analysis and Synthesis of Speaker Age. PhD thesis, Travaux de l'institut de linguistique de Lund 47. Lund: Department of Linguistics and Phonetics, Lund University (2006)
6. Schötz, S.: Acoustic analysis of adult speaker age. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol. 4343, Springer, Heidelberg (2007)
7. Schötz, S.: Stimulus duration and type in perception of female and male speaker age. In: Proceedings of Interspeech 2005, Lisbon (2005)
8. Schötz, S.: Speaker age: A first step from analysis to synthesis. In: Proceedings of ICPhS 03, Barcelona, pp. 2528–2588 (2003)
9. Bruce, G., Elert, C.C., Engstrand, O., Eriksson, A.: Phonetics and phonology of the Swedish dialects - a project presentation and a database demonstrator. In: Proceedings of ICPhS 99, San Francisco, CA, pp. 321–324 (1999)
10. Boersma, P., Weenink, D.: Praat: doing phonetics by computer (version 4.3.04) [computer program]. Retrieved (March 8, 2005) (2005), from `http://www.praat.org/`
11. Müller, C.: Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht [Two-layered Context-Sensitive Speaker Classification on the Example of Age and Gender]. PhD thesis, Computer Science Institute, University of the Saarland (2005)