Analysis and Synthesis of Speaker Age

Schötz, Susanne

2007

Citation for published version (APA):
Schötz, S. (2007). *Analysis and Synthesis of Speaker Age*. Paper presented at International Congress of Phonetic Sciences.

Total number of authors:
1

ANALYSIS AND SYNTHESIS OF SPEAKER AGE

Susanne Schötz

Dept. of Linguistics and Phonetics, Centre for Languages and Literature, Lund University
susanne.schotz@ling.lu.se

ABSTRACT

Speaker age is an important speaker-specific quality, which was investigated in the two studies presented here. The first study automatically extracted 161 acoustic features from six words produced by 527 speakers, and used normalised mean values to compare the features. Segment duration and sound pressure level (SPL) range were identified as two important acoustic correlates of age. The second study developed a research tool for analysis of speaker age by data-driven formant synthesis and age-weigthed linear interpolation to simulate an age between the ages of any two of four female differently-aged reference speakers. Evaluation of the tool revealed that speaker age may in fact be simulated using formant synthesis. Both studies will be used in further attempts to model and simulate speaker age.

Keywords: speaker age, acoustic analysis, acoustic correlates, data-driven, formant synthesis.

1. INTRODUCTION

Numerous acoustic features of speech undergo change with ageing. Age-related variation has been found in duration, F_0, sound pressure level (SPL), acoustic correlates of voice quality and spectral energy distribution (phonatory and resonance) [9, 1, 13, 7]. Still, few attempts have been made to establish the relative importance of the various features. Moreover, despite the growing need for voice variation in terms of speaker-specific qualities in speech synthesis applications like spoken dialogue systems and voice prostheses, very few attempts have been made to simulate age using speech synthesis.

This paper briefly describes two studies of speaker age. The first one analysed and compared a large number of potential age cues, and the second study developed a research tool for analysis and synthesis of speaker age.

2. STUDY 1: ACOUSTIC ANALYSIS OF SPEAKER AGE

The purpose of the first study was to automatically extract and analyse a large number of acoustic features in various segments from a large speech corpus. The aim was to identify the most important acoustic correlates of female and male speaker age.

2.1. Speech Material

The speech samples consisted of 810 female and 836 male versions of the six Swedish isolated words *käke* [ˈɕɛːkə] (jaw), *saker* [ˈsɑːkəʁ] (things), *själen* [ˈɧɛːlən] (the soul), *sot* [suːt^h] (soot), *typ* [t^{h}yːp^h] (type (noun)) and *tack* [t^{h}ak^h] (thanks). These words had been used in an earlier study [11], and were selected because they contained phones with tendencies to contain age-related information (/p/, /t/, /k/, /s/, /ɕ/ and /ɧ/) [10]. The words were produced by 259 female and 268 male speakers, taken from the SweDia 2000 corpus [3] as well as from new recordings using similar equipment and conditions.

2.2. Method and procedure

First, a number of Praat [2] scripts and an automatic aligner (originally developed by Johan Frid, Centre for Languages and Literature, Lund University) were used to normalise all words for SPL and transcribe them into phoneme, plosive closure, voice onset time (VOT) and aspiration segments. The words were concatenated into six-word sound files.

Then, another Praat script extracted 161 acoustic features (in 7 groups) from the concatenated words. Some (e.g. syllables and phonemes per second, jitter and shimmer) were extracted only for the whole file, while others (e.g. F_0, formant frequencies and segment duration) were extracted for several segments, e.g. the 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. A comprehensive descriptions of the features used is given in [12].

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/sec.	whole file (i.e. all six words)
	segment duration (ms)	whole file, words, str. vowels, fricatives, plosives (+ VOT)
2	sound pressure level (dB)	
3	F_0 (Hz, semitones)	whole file, words, str. vowels
4	jitter, shimmer	whole file
5	sp. tilt, sp. emphasis, inverse-filtered SB, LTAS	
6	HNR, NHR, other voice measures	whole file, str. vowels
7	formant frequencies (F_1–F_5) sp. balance (SB)	str. vowels fricatives and plosives

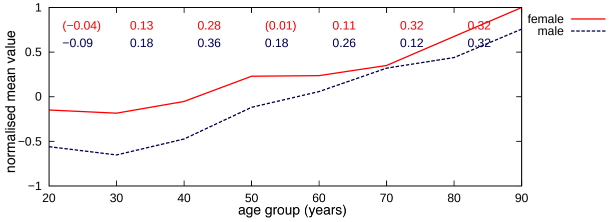
The corpus analysis toolkit m3iCAT, developed by Christian Müller at DFKI, Saarland University, was used to calculate normalised mean values for each age class. The advantage of using normalised means is that variation can be studied across features regardless of differences in their original scaling and units. Line graphs with the age classes on the x-axis and the normalised mean values on the y-axis were generated to show the tendencies of the age-related variation. In addition, the differences between the normalised means of all pairs of adjacent age classes were 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 except the ones within parentheses are statistically significant ($p \leq 0.01$). A detailed description of m3iCAT is given in [8].

2.3. Results

Though most of the features were found not to vary consistently with age, some did. As it would be impossible to present every result within the scope of this paper, only the features which varied consistently with age are described here. A more detailed presentation of the results is given in [12].

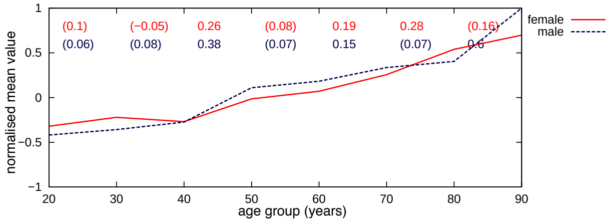
Segment duration increased with advancing age in most segments. The tendencies were less clear for the female than the male speakers. Figure 1 shows the results for all six words.

Figure 1: Normalised tendencies (for all speakers) for *duration (all six words).*



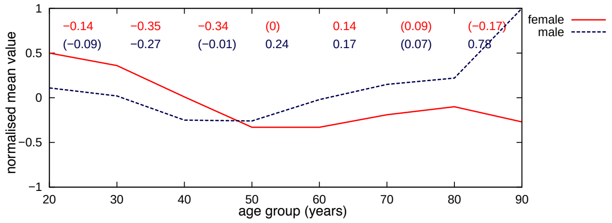
Average relative SPL generally either decreased slightly or remained constant with increased female and male age. The SPL range either increased or remained relatively stable with advancing age for both genders. Figure 2 shows the results for SPL range in the word *käke*. Similar tendencies were found for all words, including *själen* which contains no plosives.

Figure 2: Normalised tendencies (for all speakers) for *SPL range (käke).*



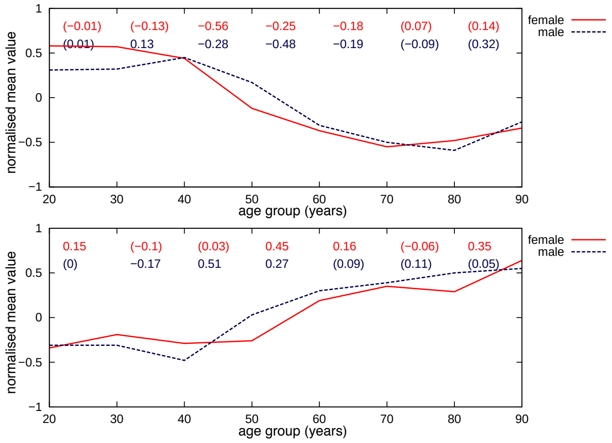
Female F_0 decreased until age group 50 and then remained relatively stable. Male F_0 lowered slightly until age group 50, but then rose into old age. Due to the gender-related differences in F_0, the results for mean F_0 (Hz, all six words) are presented in Figure 3 as normalised separately for each gender to show clearer tendencies.

Figure 3: Normalised tendencies (separately for each gender) for *mean F_0 (Hz, all six words).*



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 age in [ɑː] and [ɛː] for both genders, but decreased slightly in [a] and [uː], interrupted by increases and peaks at age group 40. Figure 4 shows normalised tendencies for F_1 and F_2 in the vowel [ɛː].

Figure 4: Normalised tendencies (separately for each gender) for *mean F_1* (top) *and F_2* (bottom) *([ɛː] in the word själen [ˈɧɛːlən])*



2.4. Discussion and conclusion

The speech material as well as the method influenced the results. Six words may constitute too short samples to capture all aspects of ageing. Moreover, automatic analyses have to be checked manually. This was done only to some extent. Still, the following tentative conclusion is drawn: The relatively most important correlates of adult speaker age seem to be speech rate and SPL range. F_0 may also provide variation with speaker age, as may F_1 and F_2 in some segments. These features may be used in combination with other features as cues to speaker age.

3. STUDY 2: FORMANT SYNTHESIS OF SPEAKER AGE

The purpose of the second study was to develop a research tool for analysis of speaker age using data-driven formant synthesis. The aim was to simulate speaker age using the research tool.

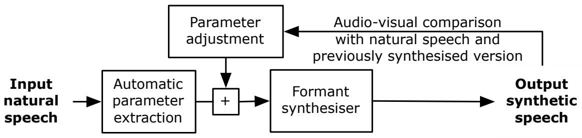
3.1. Speech material

Four female non-smoking native Swedish speakers of the same family and dialect were selected to represent different ages, and recorded twice over a period of 3 years: a girl (aged 6 and 9), her mother (aged 36 and 39), her grandmother (aged 66 and 69), and her great grandmother (aged 91 and 94). The isolated word ‘själen’ [ˈɧɛːlən] (the soul), was selected as a test word, and the recordings were segmented into phonemes and normalised for intensity.

3.2. Method and procedure

The formant synthesiser GLOVE [4], an extension of OVE III [6] with an expanded LF voice source model [5] was used in the study by kind permission of CTT, KTH. A Praat script extracted 23 GLOVE parameters from the words every 10 ms and generated synthesised copies with GLOVE. An overview of the tool is shown in Figure 5. For a detailed description of the parameters and the method, see [12].

Figure 5: Schematic overview of the research tool.



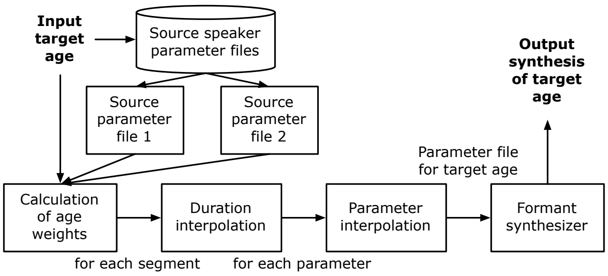
Parameter adjustment
Audio-visual comparison with natural speech and previously synthesised version
Input natural speech
Automatic parameter extraction
+
Formant synthesiser
Output synthetic speech

The parameters were then adjusted to generate more natural-sounding synthesis. Another Praat script was developed, which called the parameter extraction script, and then displayed waveforms and spectrograms of the original, the synthesised, as well as the previously synthesised words. Audio-visual comparison of the three versions allowed the user to determine whether a newly added parameter or adjustment had improved the synthesis. Whenever this was the case, the adjustment was added to a set of rules. Formants, amplitudes and voice source parameters (except F_0) caused the most serious problems, which were first solved using fixed values, then by parameter smoothing.

The synthesised versions of the words were then used to simulate other ages by age-weighted linear interpolation between two reference parameter files. A Java program was developed to calculate the weights and to perform the interpolations.

For each target age provided as input by the user, the program 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 36 and 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. 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 synthesised reference words for comparison. A schematic overview of the procedure is shown in Figure 6.

Figure 6: Schematic overview of the interpolation method.



3.3. Evaluation

To evaluate the tool’s performance, two direct age estimation perception tests were carried out. Stimuli in the first evaluation consisted of natural and synthesised versions of the 6, 36, 66 and 91 year old speakers. The second evaluation was carried out at a later stage when the 9, 39, 69 and 94 year olds had been included, and when parameter smoothing and pre-emphasis filtering had improved the synthesis. Thirty-one students participated in the first evaluation test, also including interpolations for eight decades (10 to 80 years), while 21 students took part in the second, which also comprised interpolations for seven decades (10 to 70 years).

In the first evaluation, the correlation curves between chronological age (CA, or simulated “CA” for the synthesised words) and perceived age (PA) displayed some similarity for the natural and synthesised words, though the synthesised ones were judged older in most cases, as seen in Figure 7. The interpolations were mostly judged as much older than both the natural and synthesised words.

Figure 7: Correlation between chronological age (CA) and perceived age (PA) for natural, synthesised and interpolated words (first evaluation).

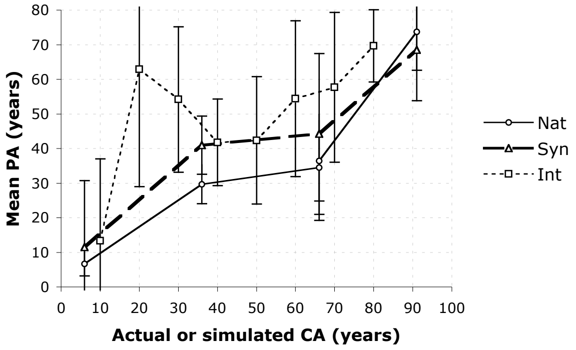
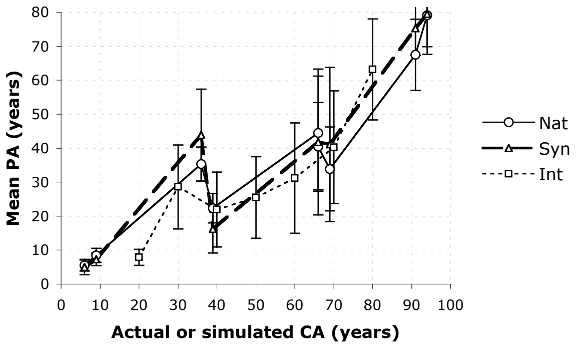


Figure 8 shows that the correlations for the natural, synthesised and interpolated words are much more similar in the second evaluation compared to the first. However, the natural and synthesised words for 39, 66 and 69 years were quite underestimated.

Figure 8: Correlation between chronological age (CA) and perceived age (PA) for natural, synthesised and interpolated words (second evaluation).



3.4. Discussion and conclusion

The synthesised words often obtained a good resemblance with the natural words, and the similarity was improved in the second evaluation. However, some of the age estimations were quite unexpected. For instance, the ages 39, 66 and 69 were judged as much younger than their CA, perhaps because these natural voices were rather atypical for their age. Still, the conclusion is drawn that speaker age may be successfully simulated using data-driven formant synthesis.

4. FUTURE WORK

The two studies presented here have contributed to our knowledge about acoustic cues to speaker age as well as provided a tool, which if developed further, may be used in future studies with systematic variation and detailed study of potential age parameters. However, additional acoustic studies with a larger material (more speakers as well as longer and more varied speech samples) are needed to verify the results of the first study and to develop further the research tool of the second study. Since ageing is far from linear, the interpolation method would benefit from more reference speakers of different ages in the material. Future work also involves further attempts to model and simulate speaker age as well as other speaker-specific qualities, including dialect and attitude. The phonetic knowledge gained from such experiments may then be used in future speech synthesis applications to generate more natural-sounding synthetic speech.

5. REFERENCES

[1] Amerman, J. D., Parnell, M. M. 1992. Speech timing strategies in elderly adults. *Journal of Voice* 20, 65–67.

[2] Boersma, P., Weenink, D. 2005. Praat: doing phonetics by computer (version 4.3.04) [computer program]. http://www.praat.org/ visited 8-Mar-05.

[3] Bruce, G., Elert, C-C., Engstrand, O., Eriksson, A. 1999. Phonetics and phonology of the Swedish dialects – a project presentation and a database demonstrator. *Proc. ICPhS 99,* San Francisco, 321–324.

[4] Carlson, R., Granström, B., Karlsson, I. 1991. Experiments with voice modelling in speech synthesis. *Speech Communication*, 10:481–489.

[5] Fant, G., Liljencrants, J., Lin, Q. 1985. A four-parameter model of glottal flow. *STL-QPSR*, 4:1–13.

[6] Liljencrants, J. 1968. The OVE III speech synthesizer. *IEEE Trans AU-16*, No 1:137–140.

[7] Linville, S. E. 2001. *Vocal Aging*. San Diego: Singular Thomson Learning

[8] Müller, C. 2005. *Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht*. PhD thesis, Computer Science Institute, Saarland University.

[9] Ryan, W. J. 1972. Acoustic aspects of the aging voice. *Journal of Gerontology*, 27:256–268.

[10] Schötz, S. 2003. Speaker age: A first step from analysis to synthesis. *Proc. 15th ICPhS* Barcelona, 2528–2588.

[11] Schötz, S. 2005. Stimulus duration and type in perception of female and male speaker age. *Proc. Interspeech 2005*. Lisbon.

[12] Schötz, S. 2006. *Perception, Analysis and Synthesis of Speaker Age.* PhD thesis, Travaux de l'Institut de linguistique de Lund 47. Lund: Dept. of Linguistics and Phonetics, Lund University.

[13] Xue, S. A., Deliyski, D. 2001. Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications. *Educational Gerontology*, 21:159–168.