Umeå University, Department of Philosophy and Linguistics
PHONUM 9 (2003), *153-156*
Available online at http://www.ling.umu.se/fonetik2003/

Towards synthesis of speaker age: A perceptual study with natural, synthesized and resynthesized stimuli

Susanne Schötz

Department of Linguistics and Phonetics, Lund University

As a first step towards synthesis of speaker age the hypothesis that spectral cues may be more important for age perception than F_0 and duration was tested in a pilot listening experiment with male speaker stimuli consisting of natural, synthesized and resynthesized isolated words. Results indicate that spectral information is dominant over pitch as cues for age. Slow speech rate also seems to be an important cue for old age. The results will be used in further research with a larger material in an attempt to model typical speaker age using formant synthesis.

1. Introduction

One way to increase the naturalness of synthesized speech would be to include speaker-specific information, such as age, sex, physical and emotional state. In order to model such paralinguistic qualities, more knowledge about their acoustic and perceptual correlates is needed, as they are often present in several phonetic dimensions.

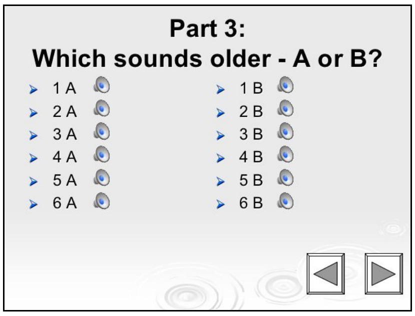
This paper describes a small pilot study of perceptual cues to speaker age. The acoustic and perceptual dimensions normally associated with age are: (1) *fundamental frequency (F_0)/pitch (mean level, range, SD)*, (2) *intensity/loudness*, (3) *jitter and shimmer/harsh voice*, (4) *formant frequencies and spectral tilt/voice quality* and (5) *duration and pausing/speech rate and rhythm* (Hollien, 1987; Schötz, 2001:b). Increased variability and lower range of F_0 is usually observed in older speakers (Hollien, 1987). Up to the age of 50, average F_0 is normally lowered, but may be raised again at very old age (Lindblad, 1992). Intensity is either lowered due to reduced vital capacity and vocal fold vibration, or increased (Hollien, 1987). Formant frequencies are lowered (Linville, 1987), and the spectral tilt normally increases with age, except at 0-5 kHz for some vowels (Decoster 1998). Vocal fry, breathiness, jitter and shimmer are more common in both female and male older voices (Hollien, 1987; Linville, 1987). Segment duration increases with age, resulting in a lower speech rate and higher maximum vowel duration in fluent speech (Ptacek & Sander 1966). Other factors influencing perception of age include prosodic patterns, grammar, sentence structure and choice of words, especially in longer sequences of speech.

F_0 and F_0SD have been shown to be dominant cues to age perception (Jacques & Rastatter, 1990; Linville, 1987; Traunmüller & van Bezooijen, 1994; Hollien, 1987). Jitter, shimmer, formant frequencies, spectral tilt and segment duration also seem to play an important role (Linville, 1987; Morris & Brown, 1987; Ringel & Chodzko-Zajko, 1987).

The aim of this study was to gain better understanding of the various cues to speaker age. In order to find out whether spectral information or F_0 is the more important cue for age estimation a small listening test was created using one-word stimuli of various types.

2. Material and method

The listening test was carried out using stimuli consisting of 12 natural, 2 synthesized and 48 resynthesized versions of the single Swedish word *rasa* [ˈʁɑːsa] (collapse). This word was chosen because it had performed best out of three isolated words of various lengths in a previous age estimation experiment (Schötz, 2001:a). The natural stimuli were produced by six older (60-76 years) and six younger (20-29 years) non-pathological male speakers of the Swedish dialect Småländska, taken from the Swedish dialect project SweDia-2000 (Bank of Sweden). Two synthesized versions of *rasa* with monotonous F_0's of 80 and 110 Hz respectively were created using the young (30 year old) male Swedish MBROLA-based concatenation synthesis LUKAS (Filipsson & Bruce, 1997). The two synthesized and the twelve natural versions were then used to produce new PSOLA-resynthesized stimuli using a 'prosody switching' script (developed by Johan Frid, Dept. of Linguistics and Phonetics, Lund University) for the speech analysis software PRAAT (www.praat.org). This script made it possible to create two new output stimuli out of two input stimuli A and B; one with the spectral quality of A, but the F_0 and duration of B, and one with the spectral quality of B but the F_0 and duration of A. The listening test was divided into four parts based on the various stimuli types, and the stimuli were further organized into pairs to facilitate comparison. In the first part no resynthesized, but only older natural and synthesized stimuli were used to test if the LUKAS versions would be judged to be younger than the older natural stimuli by the listeners. The following parts comprised resynthesized stimuli pairs organized so that stimulus AB (with spectral features of A but with F_0 and duration of B) was compared to stimulus BA (with spectral features of B but with F_0 and duration A) in order to test which cue was more dominant for age. For the second part 24 resynthesized stimuli from the two LUKAS versions of *rasa* and the six older natural versions were used. The third part consisted of 12 resynthesized stimuli where the most typical of the older speakers had switched F_0 and duration with the six younger speakers, and the fourth part of 12 resynthesized stimuli where the most typical younger speaker had switched F_0 and duration with the six older speakers. All stimuli were normalized for intensity, and the stimuli pairs were presented in a random order in each part of the test.



Part 3:
Which sounds older - A or B?

- 1 A
- 2 A
- 3 A
- 4 A
- 5 A
- 6 A

- 1 B
- 2 B
- 3 B
- 4 B
- 5 B
- 6 B

Figure 1. The layout of part 3 of the listening test, created with Microsoft PowerPoint.

21 subjects (students and staff of the Dept. of Linguistics and Phonetics, Lund University) were asked to listen to the 36 stimuli pairs on a Macintosh PowerBook G4 using the same headphones with adjustable volume and then judge which stimulus out of each pair sounded older.

3. Results

In the first part of the listening test, the listeners judged the natural stimuli produced by older speakers to sound older than LUKAS in 90% of the stimuli pairs. Results of the following three parts of the test showed that stimuli with the older speaker spectral features in combination with the F_0 and duration of a younger speaker were more often judged to be older by the listeners than stimuli with spectral features of a younger speaker in combination with the the F_0 and duration of an older speaker. Stimuli containing the spectral quality of older speakers mixed with the F_0 and duration of LUKAS were judged older than stimuli with the spectral quality of LUKAS mixed with the F_0 and duration of older speakers in 89% of the stimuli pairs. In 75% of the cases the spectral quality of the typical old speaker mixed with the F_0 and duration of the younger speakers was judged older than the opposite combination, and 83% of the cases were judged older when the spectral quality of older speakers were mixed with the F_0 and duration of the typical young speaker.

Table 1. The number and percentage of spectral quality judged older than F_0 and duration in the 2nd, 3rd and 4th parts of the listening test.

Part	*Stimuli pairs judged by the 21 listeners*	*No. of results*	*Spectral quality*		*F_0 and duration*	
			No. of	*%*	*No. of*	*%*
2	2 LUKAS + 6 older speakers = 12 pairs	252	224	89%	28	11%
3	1 typical older speaker + 6 younger speakers = 6 pairs	126	95	75%	31	25%
4	1 typical younger speaker + 6 older speakers = 6 pairs	126	105	83%	21	17%

4. Discussion

The aim of the study was to see whether spectral cues were more important for age perception than F_0 and duration, and in this test spectral information was indeed dominant over F_0 and duration. Sometimes very long segment duration was judged older than spectral quality; one possible explanation being that slow speech rate also is an important cue to old age. Other "errors" involved two relatively atypical speakers for their age. This could be expected as typical speakers are more easily age-judged than atypical ones (Schötz; 2001:a).

For several reasons this study should be regarded only as a first step towards finding the dominant phonetic cues for age in preparation for synthesis of speaker age. Firstly, the voice used for the LUKAS synthesis produced [ˈɹɑːsa], which is less dialectal than the versions produced by the natural speakers. This may have influenced the results as dialectal speech is considered a cue for old age. Secondly, the dialectal variation for both the older and younger natural speakers was considerable, ranging from the standard Swedish [ˈɹɑːsa] to [ˈʁɑːsa], [ˈʁɑːsə] and even [ˈwɑːsə], but younger speakers were equally dialectal as older speakers. Moreover, duration was neither normalized nor separated from the F_0 part of the stimuli. This could also have influenced the results.

Listeners are normally easily able to separate and identify various linguistic and paralinguistic information in speech, but it is extremely difficult if not impossible to separate acoustic correlates for age from other acoustic features. The limited material used in these pilot studies was only intended to provide results implying possible trends and tendencies. Future research containing a larger material including female speakers and additional methods of analysis will be necessary to confirm the results found in this experiment.

However, some ideas about which correlates to use when attempting to model and synthesize age have emerged. The indication that spectral information holds important cues further supports the idea of using formant synthesis when trying to synthesize age, as it is easier to manipulate than concatenative synthesis, which “inherits” speaker-specific information from the voice used to record its units. Moreover, it is probably a good idea to start by trying to model and synthesize typical or even stereotypical age, based on phonetic information of typical speakers, and try to include other features associated with age, such as dialectal variations and choice of words at a later stage.

5. Conclusions

Based on the age judgements made by the listeners of the present study it is concluded that spectral cues are probably more important than the prosodic cues of F_0 and duration. The only exceptions found were when segment duration was very long, indicating that speech rate is an important age-related cue, and when atypical speakers were judged, implying that attempts to synthesize speaker age probably should be based on a phonetic model of typical or even stereotypical age.

Studies with larger material are needed to verify the tentative results presented in this paper. A future attempt to synthesize age using formant synthesis will be based on the results of this study, as spectral cues seem to be important cues to perception of speaker age.

6. References

Decoster, W. (1998) *Akoestische kenmerken van de ouder wordene stem*, Leuwen: Leuwen University Press. (Summary in English)

Filipsson, M. & Bruce, G. (1997) LUKAS - a preliminary report on a new Swedish speech synthesis. In *Working Papers 46*, Department of Linguistics and Phonetics, Lund University.

Hollien, H. (1987) ”Old voices”: What do we really know about them?, *Journal of Voice*, vol. 1, no 1, pp. 2-13.

Jacques, R. D. & Rastatter, M. P. (1990) Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners, *Folia Phoniatrica*, vol. 42, pp. 118-124.

Lindblad, P. (1992) *Rösten*, Lund: Studentlitteratur.

Linville, S. E. (1987) Acoustic-perceptual studies of aging voice in women, *Journal of Voice*, vol. 1, no 1, pp. 44-48.

Morris, R. J. & Brown, W. S. Jr. (1987) Age-related voice measures among adult women, *Journal of Voice*, vol. 1, no 1, pp. 38-43.

Ptacek, P. H. & Sander, E. K. (1966) Age recognition from voice, *Journal of Speech and Hearing Research*, vol. 9, pp. 273-277.

Ringel, R. L. & Chodzko-Zajko, W. J. (1987) Vocal indices of biological age, *Journal of Voice*, vol. 1, no 1, pp. 31-37.

Schötz, S. (2001:a) A perceptual study of speaker age. In *Working Papers 49*, Department of Linguistics and Phonetics, Lund University.

Schötz, S. (2001:b) *Röstens ålder – en auditiv och akustisk studie (Speaker age – an auditive and acoustic study)*. M.A. thesis in phonetics, Department of Linguistics and Phonetics, Lund University. (available online at http://www.ling.lu.se/persons/Suzi/#unpublished)

Traunmüller, H. & Bezooijen, R. van. (1994) The auditory perception of children’s age and sex. In *Proceedings ICSLP-94*, vol. 3, pp. 1171-1174. The Acoustical Society of Japan.