Speaker Age: A First Step From Analysis To Synthesis

Susanne Schötz
Department of Linguistics and Phonetics, Lund University, Sweden
E-mail: susanne.schotz@ling.lu.se

ABSTRACT

A first tentative step towards synthesis of age was made in three pilot studies of speaker age. The first was a perceptual study, where results showed that some speakers are more typical for their age than others, and that listeners are able to age estimate single word stimuli considerably better than chance. In the second study an auditive and acoustic analysis of the same material indicated that older voices contain more variability, that some cues, including spectral quality and VOT, may be more important than others (e.g. F_0), and that combinations of several cues are probably used for age perception. In the third study, natural, synthesized and resynthesized stimuli were used in a second listening test. Results indicated that spectral information and duration are more important than F_0 for age perception. The results from these studies will be used in future research with more data and formant synthesis.

1. INTRODUCTION

One way to increase the naturalness of synthetic speech would be to integrate models and parameters for paralinguistic phonetic variation including age, physical and emotional state in a TTS-system. In order to successfully synthesize such variation, more phonetic knowledge about the acoustic and perceptual correlates of these phenomena is needed. This paper describes three small pilot studies building on previous research in this area and. intended to gain better understanding of the various cues to speaker age The studies were also meant to serve as a first step towards synthesizing speaker age.

1.1. Acoustic and perceptual correlates to age

In studies carried out on the aging voice, the acoustic and perceptual dimensions associated with speaker age typically are: (1) *fundamental frequency (F_0)/pitch (average, 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* [3, 11]. F_0 varies due to physiological age-related changes of the larynx and vocal folds [5], leading to increased variability and lower range [3]. Up to the age of 50, average F_0 is normally lowered [5], but is often raised again at very old age for male speakers, while female F_0 may be constant, lowered or first lowered and then increased [3]. Moreover, vocal fry and breathiness are more common in old voices [3]. Intensity is either lowered due to reduced vital capacity and vocal fold vibration, or increased [3]. Jitter and shimmer are more frequent in both female and male older voices [6]. Formant frequencies are lowered [6], and spectral tilt increases with age, except at 0-5 kHz for some vowels [1]. Segment durations increase with age, causing lower speech rates and higher maximum vowel durations in fluent speech [8]. Perception of age is further influenced by prosodic patterns, grammar, sentence structure and choice of words, especially in longer sequences of speech.

1.2. Previous studies

It is argued that perception of speaker age is based on both laryngeal and supralaryngeal cues [3, 11] and that listeners are able to estimate speaker age to within ±5 years of chronological age. Perceptual studies have been carried out using several kinds of stimuli ranging from phonated, whispered or filtered prolonged vowels to longer read passages played both forward and backwards [8]. A few studies refer to typical and atypical speakers for their age [7, 9, 10], partly due to the individual aging process.

F_0 has shown to be an important correlate to age in several studies. The results of a study by Jacques and Rastatter [4] indicated that F_0 was the dominant cue for age perception of prolonged vowels produced by both female and male speakers. Linville [6] used only female speakers and listeners and also restrained F_0 to an interval of 40 Hz in her vowel stimuli. Despite these limitations, F_0 was still considered to be an important cue in age perception. F_0 is also considered an important cue for children's age [12]. Moreover, F_0SD increases with age [3, 6]. While no clear difference in intensity between older and younger speakers at normal speech levels have been found [7], F_0, jitter, shimmer, formant frequencies, spectral tilt and segment duration seem to play a more important role [6, 7, 9]. Stimuli durations of 0,5-2 seconds are usually enough for age estimation tasks [1, 4]. Speech rate has been reported significantly slower [8], and segment durations of vowels and consonants were considerably longer for older speakers [7], while VOT has been reported shorter for older female speakers [7], but longer for older male speakers [1] in comparison to the younger ones.

2. METHOD

2.1. Material

The material for the first two studies was taken from the speech database of the Swedish dialect project SweDia-2000 (Bank of Sweden). It consisted of the three words *tack* [tʰakː] (thanks), *rasa* [ˈʁɑːsa] (collapse) and *tusendollarsedlar* [ˈtʰʉːsɔndɔlːaʁˌseːdlaʁ] (thousand-dollar bills) spoken in isolation by four female and four male non-pathological speakers of the same dialect, and equally divided into two age groups (21-30 and 61-73 years), giving a total of 24 words. The speakers were judged by phoneticians of the SweDia-2000 project as representing

 ISBN 1-876346-48-5

both typical and atypical speakers.

In the third study only the word *rasa*, produced by the four male speakers from the first study, and also by eight other male speakers (four old and four young) of the same dialect, taken from the SweDia-2000 project as well, was used. In addition two synthetic versions of *rasa* with monotonous F_0 (80 and 110 Hz) generated by the Swedish MBROLA-based concatenative young speaker synthesizer LUKAS [2] and 24 PSOLA-resynthesized versions of the same word were used. The resynthesized or mixed stimuli were created using a script (by Johan Frid) for the speech analysis program PRAAT (www.praat.org), enabling two stimuli to switch prosody (duration and F_0) with each other, while keeping their own spectral quality. LUKAS switched prosody with all older speakers, one typical old speaker switched prosody with all younger speakers, and one typical young speaker switched prosody with all older speakers, making a total of 40 stimuli.

All of the stimuli were normalized for intensity before used in the studies.

2.2. Procedure

In the first study 38 listeners of various age, sex and phonetic experience were asked to judge the age of the 24 stimuli by choosing from 18 alternatives on a 5-year scale ranging from 10 to 95 years in an Internet-based listening test. The results were then analyzed with respect to various speaker and listener differences and tested statistically.

The second study was an acoustic analysis of the same material. Measurements of F_0 (mean, range, SD) and relative intensity, of jitter, F_1-F_4, B_1-B_4 and spectral tilt of the vowels [a ɑː ʉː ɔ eː], of spectral features of the voiceless plosives [t k] and the fricative [s] as well as of segment duration (word, phoneme and VOT) were carried out. All the acoustic analyses were made with PRAAT, and then manually checked for errors. Because of the anticipated difficulties in analyzing acoustic correlates to age, the study began with a careful auditive analysis. The material was listened to a number of times while making notes for each cue associated with old age.

In the third study the results from the two previous studies were used in a second perception test containing 36 pairs of natural as well as synthesized and resynthesized stimuli. First the synthesized stimuli were compared to the older natural speakers. Then the synthesized stimuli mixed with the natural stimuli were compared with respect to F_0 and spectral features. Finally one typical young and one typical old speaker that had switched prosody with the opposite age group were compared in the same way. 21 subjects (students and staff of the Dept. of Linguistics and Phonetics, Lund University) listened to the stimuli pairs and judged which one of each pair sounded older.

3. RESULTS

The results of the first study indicated that although only about half the stimuli were estimated correctly ±10 years, the subjects were able to judge speaker age considerably better than chance, since only 5 of the 18 alternatives of the test were correct ±10 years for each speaker. The best results were obtained for *rasa*, and the worst for *tack*. Moreover, typical speakers were significantly more often correctly age estimated than atypical speakers. No differences between listeners were found.

Table 1: The number and percentage of correct age estimations (three words, eight speakers) of the first study.

word:	*tack*		*rasa*		*tusendollarsedlar*	
(38 listeners)	no. of corr. ±10 years	%	no. of corr. ±10 years	%	no. of corr. ±10 years	%
speaker 1 (typical)	9	24	30	79	21	55
speaker 2 (atypical)	18	47	7	18	14	37
speaker 3 (typical)	10	26	25	66	31	82
speaker 4 (atypical)	23	61	24	63	17	45
speaker 5 (atypical)	8	21	15	39	13	34
speaker 6 (typical)	25	66	28	74	22	58
speaker 7 (atypical)	21	55	17	45	15	39
speaker 8 (typical)	29	76	34	89	30	79
all typical speakers	*73*	*48*	*117*	*77*	*104*	*68*
all atypical speakers	*70*	*46*	*63*	*41*	*59*	*39*
total (all speakers)	*143*	*47*	*180*	*59*	*163*	*54*

Results from the auditive part of the second study were only the subjective judgment of one person, but might still provide hints on cues to speaker age. The older typical voices contained a larger number of cues related to old age (e.g. tremor, harshness, tiredness) than the atypical older voices, whereas the atypical younger voices had more old age cues than the typical younger voices. The results from the auditive analysis were mainly used for comparisons with results from the first study and the acoustic analysis.

Table 2: The number of old age-related cues (three words, eight speakers) in the auditive part of the second study.

speakers:	*tack*	*rasa*	*tusendollarsedlar*	**total**
1 (typical older man)	1	11	13	**31**
2 (atypical older man)	8	8	8	**24**
3 (typical older woman)	17	16	24	**57**
4 (atypical older woman)	8	7	6	**21**
5 (atypical younger man)	9	15	15	**39**
6 (typical younger man)	1	6	4	**11**
7 (atypical younger woman)	5	6	15	**26**
8 (typical younger woman)	1	6	6	**13**
all typical older speakers	*18*	*27*	*37*	**82**
all atypical older speakers	*16*	*15*	*14*	**45**
all atypical younger speakers	*14*	*21*	*30*	**65**
all typical younger speakers	*2*	*12*	*10*	**24**
total (all speakers)	***50***	***75***	***91***	***216***

The results of the acoustic analyses were relatively poor, showing only possible trends. No support for previous studies was found with regard to F_0 being an important cue to age. However, all the female stimuli had wider F_0 range and higher F_0SD than the male ones. As intensity had been normalized, only relative measures were carried out. The results indicated a smaller intensity range for the older speakers compared to the younger speakers. Jitter could not be measured in all of the words because of the sometimes very short vowel durations, and the only trend in the obtained measures was that older sounding male speakers tended to have higher jitter levels than younger sounding speakers. The formant frequencies and

bandwidths measured showed no pattern or trends, except the obvious; that male formant frequencies were lower than female ones. In the spectra for [t], the older sounding speakers displayed higher intra-speaker variability than the younger sounding speakers. Spectra for [s] showed that the typical energy platform on higher frequencies began around 4 kHz for younger-sounding speakers, but already between 3,5 - 4 kHz for older-sounding speakers.

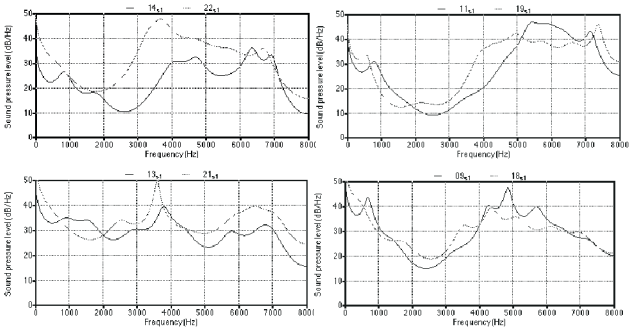

Figure 1: Spectra for [s] in *tusendollarsedlar* (dotted lines) and *rasa* (plain lines) produced by female (top), male (bottom), typical old (left) and typical young (right) speakers.

Spectral tilt ranged from about –15 to –20 dB/octave, being far below the values considered normal (about –12 dB/octave). This implied methodological errors and therefore unreliable results, but the tendency was that the older sounding speakers had somewhat higher tilts than the younger sounding ones. Word and phoneme duration correlated with neither chronological nor perceived age, but the older sounding speakers frequently displayed longer VOT than the younger sounding speakers.

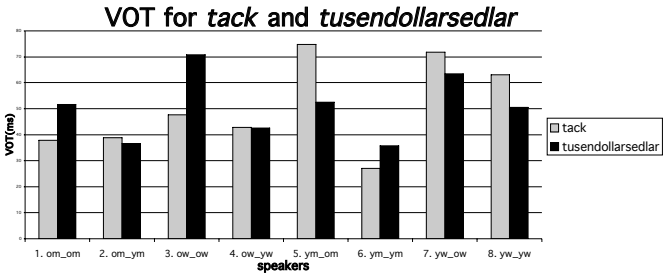

Figure 2: VOT values for *tack* and *tusendollarsedlar* (all speakers, for explanation of speaker numbers, see table 2).

Results from the third study showed that stimuli with spectral features of an older speaker in combination with the prosody (F_0 and duration) of a younger speaker were mostly judged to be older by the listeners than the opposite combination, i.e. stimuli with spectral features of a younger speaker in combination with the prosody of an older speaker. Stimuli containing the spectral quality of older speakers mixed with the prosody of LUKAS were judged older than stimuli with the spectral quality of LUKAS mixed with the prosody of older speakers in 89% of the stimuli pairs. Spectral quality of one typical old speaker mixed with the prosody of young speakers was judged older than the opposite combination in 75% of the cases, and 83% of the cases were judged older when the spectral quality of older speakers where mixed with the prosody of one typical young speaker. Two atypical speakers accounted for most of the "errors" of the test.

Table 3: The number and percentage of spectral quality and F_0 & duration being judged older in the third study.

stimuli pairs judged older:	*no. of results*	*spectral quality*		*F_0 and duration*	
		no. of	*%*	*no. of*	*%*
LUKAS + old speakers (mixtures)	252	224	89%	28	11%
one typical older speaker + younger speakers (mixtures)	126	95	75%	31	25%
one typical younger speaker + older speakers (mixtures)	126	105	83%	21	17%

4. DISCUSSION

Several earlier studies have only divided speakers into older and younger age groups. The results of this first study show that listeners are able to age estimate speakers in more detail (±10 years). Also, it seems that a single isolated word may be enough to get correct age estimations. Some of the speakers seemed more typical for their age than others, and the listeners were significantly better at age estimating typical speakers than atypical speakers, suggesting that there is a subdivision of speakers into a typical and an atypical group. This may be an important piece of knowledge when attempting to model and synthesize (typical) age.

The auditive part of the second study indicated that both the number of cues related to old age and the overall deviation from modal or "normal" voice may be important when judging age. Older sounding speakers had more variability in their speech and deviated more from modal voice than younger sounding speakers. Overall variability is an important cue according to the listeners' comments to the listening tests [8, 11]. The large number of different cues related to old age might indicate that even "normal" old voices sound more pathological than young voices.

The small size of the material is probably one reason why the results of the acoustic analyses were relatively poor, but some trends managed to emerge. Although reported important in previous studies, F_0 did not seem to be a dominant cue in the material of this study. One reason for this might be that isolated words instead of prolonged vowels were studied, and the listeners had to rely on cues provided by consonants and prosody as well. The results on intensity, where older speakers had slightly smaller ranges should be regarded merely as trends, as age differences in intensity at conversational effort have shown to be minimal [7]. The high inter- and intra-speaker variability in vowel quality made it hard to interpret the results of the formant frequency analyses. This suggests that although formant frequencies of vowels are important cues, the spectral quality of voiceless segments are too, as they also seem to vary with age, e.g. that [s] produced by older sounding speakers sounded more muffled. Although word and phoneme duration measures correlated with neither chronological nor perceived age, probably due to

ISBN 1-876346-48-5

high inter-speaker variability in speech rate, VOT still seemed a good indicator of speaker age. Thus, it is likely that listeners use a combination of several cues when judging speaker age, and that the choice of cues depend on the type and duration of the stimuli, but perhaps also on individual experiences with voices of various ages.

The aim of the third study was to see whether spectral cues were more important for age perception than F_0 and duration. Indeed spectral information was more often used as cues for speaker age than F_0 and duration. The cases where prosody was judged older than spectral quality may be explained. Two atypical speakers accounted for the major part of these estimates, and when one stimulus was significantly longer than the other, this stimulus was often regarded older, indicating that speech rate also is an important cue to age. The results on intensity in the second study, where the range seemed smaller in older speakers have to be investigated further to check if other factors, such as F_0 may have influenced the intensity of the third study stimuli.

There are a large number of factors to be considered when trying to synthesize age. 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 and more methods of analysis will be necessary. However, some ideas of what 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 (as opposed to concatenative synthesis, which "inherits" speaker-specific information from the voice used to record its units) when trying to synthesize age. Finally, maybe the saying "you are only as old as you feel" also applies to speech and even synthesized speech.

5. CONCLUSIONS

Listeners are able to age estimate speakers ±10 years considerably better than chance according to the results of the first study. Moreover, the listeners were significantly better at age estimating typical speakers than atypical speakers, suggesting that there is a subdivision of speakers into a typical and an atypical group.

The older sounding voices analyzed in the auditive part of the second study contained larger variability and deviated more from modal or "normal" voice quality than did the younger speakers analyzed, suggesting that even normal old voices sound more pathological than young voices.

Although the material in the acoustic analysis was too small to show reliable results, tendencies of some age-related correlates were found. F_0, intensity and formant frequencies did not seem to be important cues for age in this material. However, jitter levels were higher in older sounding speakers, spectra for [s] and [t] showed different patterns for older and younger sounding speakers, and VOT values were longer for older sounding speakers. When less vowel information is present, as in single word stimuli, the cues of the voiceless segments may become more important, as listeners are still able to make correct age judgements.

From the age judgements made by the listeners of the third study it was concluded that spectral cues were more important than the prosodic cues of F_0 and duration, except when judging atypical speakers or when duration was extremely long, indicating that speech rate is an important age-related cue.

Studies with larger material are needed to verify the tentative results presented in this paper. Future attempts to synthesize age will be made using formant synthesis, as spectral cues appear to be important.

6. REFERENCES

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

[2] M. Filipsson, and G. Bruce, "LUKAS - a preliminary report on a new Swedish speech synthesis," in *Working Papers* 46, Department of Linguistics and Phonetics, Lund University. 1997.

[3] H. Hollien, ""Old Voices": What Do We Really Know About Them?," *Journal of Voice*, vol. 1, no 1, pp. 2-13, 1987.

[4] R.D. Jacques and M.P. Rastatter, "Recognition of Speaker Age from Selected Acoustic Features as Perceived by Normal Young and Older Listeners," *Folia Phoniatrica*, vol. 42, pp. 118-124, 1990.

[5] P. Lindblad, *Rösten*, Lund: Studentlitteratur, 1992.

[6] S.E. Linville, "Acoustic-Perceptual Studies of Aging Voice in Women," *Journal of Voice*, vol. 1, no 1, pp. 44-48, 1987.

[7] R.J. Morris and W.S. Brown, Jr., "Age-related Voice Measures Among Adult Women," *Journal of Voice*, vol. 1, no 1, pp. 38-43, 1987.

[8] P.H. Ptacek and E.K. Sander, "Age recognition from Voice," *Journal of Speech and Hearing Research*, vol. 9, pp. 273-277, 1966.

[9] R.L. Ringel and W.J. Chodzko-Zajko, "Vocal Indices of Biological Age," *Journal of Voice*, vol. 1, no 1, pp. 31-37, 1987.

[10] S. Schötz, "A perceptual study of speaker age," in *Working Papers* 49, Department of Linguistics and Phonetics, Lund University. 2001.

[11] S. Schötz, *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. 2001.

[12] H. Traunmüller and R. van Bezooijen, "The auditory perception of children's age and sex," in *Proceedings ICSLP-94*, vol. 3, pp. 1171-1174. The Acoustical Society of Japan, 1994.