



LUND UNIVERSITY

Some Acoustic Cues to Human and Machine Estimation of Speaker Age

Schötz, Susanne

Published in:
Proc. of Fonetik 2004

2004

[Link to publication](#)

Citation for published version (APA):

Schötz, S. (2004). Some Acoustic Cues to Human and Machine Estimation of Speaker Age. In *Proc. of Fonetik 2004* (pp. 40-43). Department of Linguistics, Stockholm University.

Total number of authors:

1

General rights

Unless other specific re-use rights are stated the following general rights apply:

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: <https://creativecommons.org/licenses/>

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Some Acoustic Cues to Human and Machine Estimation of Speaker Age

Susanne Schötz

Dept. of Linguistics and Phonetics, Lund University

Abstract

Two experiments were carried out in order to learn more about the relation between the various acoustic cues to speaker age. The first included listening tests with resynthesized stimuli, and the second comprised automatic estimation of age using the CART (Classification And Regression Trees) technique. In the first experiment, results indicate that human listeners seem to rely more on spectral cues than on F_0 and duration when judging age. The results of the second experiment seem to agree with the first, as formant frequencies outperformed the other features in the CART tests. The acoustics and perception of speaker age will be studied further using a larger material and additional methods.

Introduction

When estimating the age of a speaker, we probably use a combination of several cues present in the speech signal, but which cues are the most important ones? Furthermore, which acoustic cues would an automatic age estimator need in order to make fairly correct judgements? Would they be the same as the ones used by humans?

This paper describes two experiments – one with human listeners and one with a machine learning technique – aiming at identifying some important cues to speaker age for humans as well as for machines.

Background

Researchers agree that humans are usually able to estimate speaker age to within 10 years. Age cues have been found in almost every phonetic dimension, but the relationship between the various cues has not been fully explored yet. Several studies have found F_0 and F_0SD to be dominant cues to age perception (Hollien, 1987; Jacques & Rastatter, 1990; Linville, 1987). However, some recent studies have failed to find strong correlations between F_0 and age, suggesting that other factors, including speech rate and spectral features are more important to perception of speaker age (Schötz, 2003; Winkler et al., 2003). Even the performance of an

automatic estimator of age was improved when speech rate and shimmer were included as cues (Minematsu et al., 2002). However, studies of speaker age are not easy to compare due to differences in speaker gender and age distribution as well as in the types of speech material used in the experiments.

Purpose and Aim

The purpose of these two studies is to investigate the relationship between several acoustic cues to age and try to identify the most important ones, by studying human perception of age as well as an automatic estimation technique. In the human listener study, F_0 and duration are contrasted with the rest of the speech signal containing the spectral qualities, and in the machine experiments, 51 acoustic feature values are compared. The aim of the two studies is to increase our understanding of the acoustic cues used in both human and automatic estimation of speaker age.

Experiment I (Human listeners)

The experiment with human listeners will be explained only briefly here. A more detailed description is given in Schötz (2004). It consisted of two almost identical perception tests – one with only female speaker stimuli and one with male speaker stimuli. The purpose was to investigate if F_0 and duration are more important to age perception than other qualities in speech, and if there were any differences between perception of female and male speaker age.

Material

Twenty-four elicitations from twelve female and twelve male natural speakers, taken from the Swedish dialect project SweDia 2000 (Bruce et al., 1999), and two female and two male MBROLA-based concatenation synthesis versions (Filipsson & Bruce, 1997, Svensson, 2001) of the word *rasa* (collapse) were used in the listening tests. Twelve of the natural speakers were older speakers (60-82 years) and the other twelve as well as the speakers who had recorded the diphones of the synthetic versions

were younger speakers (18-31 years). Based on these 28 productions of the word, resynthesized stimuli were created by switching the F_0 and word duration values for two input words A (always an older speaker) and B (always a younger speaker), so that output stimulus AB consisted of the spectral quality (i.e. the whole signal except F_0 and duration) of the older input A, but with the duration and F_0 of the younger input B, while output stimulus BA consisted of the spectral quality of the younger input B, except for the F_0 and duration, which was from the older input A, as shown in Figure 1. All stimuli were normalized for intensity.

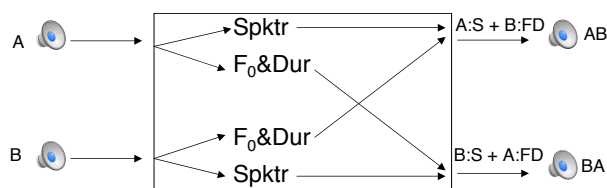


Figure 1. Schematic diagram of how the resynthesized stimuli were created.

Method

In the two perception tests, stimuli pairs of type AB and BA were presented in randomized order and the task was to decide which stimulus sounded older. 31 students of phonetics (age: 18-36, μ : 21.6) participated in the female speaker test and 29 (age: 18-28, μ : 22.3) took part in the male speaker test.

Results

Table 1 shows that the listeners more often judged older speech with younger F_0 and duration (stimulus type AB) as older than younger speech with older F_0 and duration (stimulus type BA).

Table 1. The number and percentage of spectral quality as well as F_0 & duration judged older by the listeners for female (a) and male (b) speakers.

(a) female stimuli pairs judged older:	no. of results	spectral		F_0 & dur	
		no. of	%	no. of	%
all older + two synthetic	372	258	69%	114	31%
one older + all younger	183	121	66%	62	34%
all older + one younger	181	104	57%	77	43%

(b) male stimuli pairs judged older:	no. of results	spectral		F_0 & dur	
		no. of	%	no. of	%
all older + two synthetic	347	322	93%	25	7%
one older + all younger	174	120	69%	54	31%
all older + one younger	174	139	80%	35	20%

The results were somewhat better for the male speakers, but all significant, as shown in Table 2.

Table 2. χ^2 -results for the female and male tests.

part	all older + two synthetic		one older + all younger		all older + one younger	
	$\chi^2(I)$	$p <$	$\chi^2(I)$	$p <$	$\chi^2(I)$	$p <$
female	55.742	.001	19.022	.001	4.028	.045
male	254.205	.001	25.034	.001	62.161	.001

Experiment II (Machine approach)

For the automatic age estimation experiments, the CART (Breiman et al., 1984) technique was employed. In this method, both statistical learning and expert knowledge is used to construct binary decision trees, formulated as a set of ordered yes-no questions about the features in 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 mislabeled data samples, and efficient prediction of categorical (classification) as well as continuous (regression) feature data (Huang et al., 2001). In this study, *Wagon*, a CART implementation from the Edinburgh Speech Tools package (Taylor et al., 1999), was used. It consists of two independent applications: *wagon* for building (i.e. training) the trees, and *wagon_test* for testing the trees with new data.

Material

The material comprised 7696 feature vectors containing information from 428 natural speakers (from SweDia 2000) of various ages (17-84 years), each having produced between 3 and 14 elicitations of the word *rasa*. A number of scripts (developed by Johan Frid, Dept. of Linguistics and Phonetics, Lund University) for the speech analysis software PRAAT (Boersma & Weenink, 2004) were extended and adjusted to automatically extract, and store in data files, vectors of 51 acoustic feature values from the four segments of the words, including mean, median, range, range 2 (excluding the top and bottom 5%) and SD (standard deviation) for F_0 and for F_1 - F_5 , as well as measurements of relative intensity, duration, HNR (Harmonics-to-Noise Ratio), spectral emphasis, spectral tilt, jitter and shimmer. 90% of the vectors were used for training, and the remaining 10% were used for testing the CARTs.

Method

The CART experiments were carried out in three sets. First, only one feature value at a time was used to build trees for age estimation. Second, all values (i.e. mean, median, range etc.) for each of the six features which had performed best in the first set, were further tested to determine their relative order. Finally tests were run with the 21 best feature values and with all of the 51 feature values of the vectors.

Results

From the tests with one feature value at a time, the 21 features with higher correlations than 0.4 between chronological and estimated age are shown in Table 3. The mean and median values for the formant frequencies performed best, followed by their range values and the mean and median values for F_0 . Except for HNR ($r = 0.2033$), none of the other features reached correlations over 0.2.

Table 3. The 21 best correlations between chronological and estimated age for the CART tests using only one feature value at a time.

Nr	Feature	Corr (r)	Nr	Feature	Corr (r)
1	F4 (mean)	0.5195	11	F1 (median)	0.4374
2	F4 (median)	0.5163	12	F3 (range)	0.4356
3	F3 (median)	0.5162	13	F1 (mean)	0.4348
4	F3 (mean)	0.5070	14	F5 (range)	0.4269
5	F2 (median)	0.4977	15	F4 (range 2)	0.4252
6	F2 (mean)	0.4855	16	F0 (mean)	0.4232
7	F5 (mean)	0.4819	17	F0 (median)	0.4220
8	F5 (median)	0.4817	18	F2 (range)	0.4207
9	F4 (range)	0.4455	19	F1 (range)	0.4169
10	F3 (range 2)	0.4446	20	F1 (range 2)	0.4069
			21	F2 (range 2)	0.4021

Figure 2 shows the correlations between chronological and estimated age for the CARTs using all values (mean, median, range, SD) for F_0 and for F_1 - F_5 , and also shows correlations for the CART using only the best 21 feature values as well as the CART for all of the 51 feature values. The best single feature results were obtained by F_3 followed by F_4 , and there was only a slight improvement in performance when using all 51 features ($r = 0.8752$) compared to the CART using only the 21 best features ($r = 0.8535$).

Correlations between chronological and estimated age for the best feature CARTs

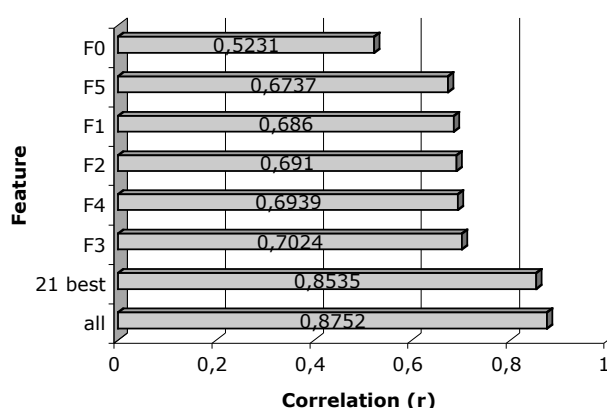


Figure 2. Correlations between chronological and estimated age for the best feature CARTs.

Discussion

For human perception of speaker age, it seems that F_0 and duration are less important than the spectral cues (i.e. the rest of the speech signal). However, which cues the listeners actually did use in their judgements still remains unclear. Formants and other spectral information, including spectral tilt and glottal features, may all provide cues to speaker age. Since some previous studies have failed to find strong correlations between specific spectral features and age (Schötz, 2003), it is possible that listeners use combinations of several cues.

Of the 51 features used in the experiments with the automatic age estimator, the formant frequencies, especially F_3 and F_4 , performed best. This is in line with the human study, so it is not impossible that humans and machines rely on similar acoustic cues in order to judge speaker age.

As both the material and the methods used in these two experiments are likely to have influenced the results, larger studies with more varied material are needed in further pursuit of the most important acoustic cues to age. Moreover, additional supralaryngeal and laryngeal features, including B_1 - B_5 , L_1 - L_5 , source spectra (using inverse filtering techniques) and LTAS, which might influence both human and machine estimation of speaker age, will be analyzed. Future work also includes experiments with other machine learning techniques, including HMM (Hidden Markov Models) and NN (Neural Networks), and studies of potentially important age cues using formant synthesis in attempts to synthesize speaker age.

References

- Boersma, P. and Weenink, D. (2004) PRAAT: doing phonetics by computer. Website: <http://www.praat.org>.
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) Classification and regression trees. Wadsworth and Brooks. Pacific Grove.
- 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. Proceedings of the XIVth ICPhS, San Francisco, 321-324.
- Filipsson, M. and Bruce, G. (1997) LUKAS – a preliminary report on a new Swedish speech synthesis Working Papers (Dept. of Linguistics, Lund University) 46, 45-56.
- Hollien, H. (1987) “Old Voices”: What Do We Really Know About Them? *Journal of Voice* 1(1), 2-13.
- Huang, X., Acero, A. and Hon, H. (2001) *Spoken Language Processing*. Prentice Hall. Upper Saddle River, New Jersey.
- Jacques, R.D. and Rastatter, M.P. (1990) Recognition of Speaker Age from Selected Acoustic Features as Perceived by Normal Young and Older Listeners. *Folia Phoniatrica* vol. 42, 118-124.
- Linville, S.E. (1987) Acoustic-Perceptual Studies of Aging Voice in Women. *Journal of Voice* 1(1), 44-48.
- Minematsu, N., Sekiguchi, M. and Hirose, K. (2002) Performance Improvement in Estimating Subjective Agedness with Prosodic Features. *Proceedings of Speech Prosody 2002, Aix en Provence*.
- Schötz, S. (2003) Speaker Age: A First Step From Analysis To Synthesis. *Proceedings of the XVth ICPhS, Barcelona, 2585-2588*.
- Schötz, S. (2004) The Role of F_0 and Duration in Perception of Female and Male Speaker Age. *Proceedings of Speech Prosody 2004, Nara*.
- Svensson, A. (2001) *Ofelia – en ny syntesröst*. M.A. thesis in Computational Linguistics. Dept. of Linguistics and Phonetics, Lund University.
- Taylor, P., Caley, R., Black, A. and King, S. (1999) *Edinburgh Speech Tools Library System. Documentation Edition 12*. Website: http://festvox.org/docs/speech_tools-1.2.0/book1.htm
- Winkler, R., Brückl, M. and Sendlmeier, W. (2003) The Aging Voice: an Acoustic, Electroglottographic and Perceptive Analysis of