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 F0 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 F0 and F0SD 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 F0 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, F0 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 F0 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 F0 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 F0 and duration) of the older input A, but with the duration and F0 of the younger input B, while output stimulus BA consisted of the spectral quality of the younger input B, except for the F0 and duration, which was from the older input A, as shown in Figure 1. All stimuli were normalized for intensity.

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

### Table 2. $\chi^2$-results for the female and male tests.

<table>
<thead>
<tr>
<th>part</th>
<th>all older + two synthetic</th>
<th>one older + all younger</th>
<th>all older + one younger</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>$\chi^2$ (I)</td>
<td>$p &lt;$</td>
<td>$\chi^2$ (I)</td>
</tr>
<tr>
<td>female</td>
<td>55.742</td>
<td>.001</td>
<td>19.022</td>
</tr>
<tr>
<td>male</td>
<td>254.205</td>
<td>.001</td>
<td>25.034</td>
</tr>
</tbody>
</table>

### 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 F0 and for F1-F5, 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 F0. 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.

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

Figure 2 shows the correlations between chronological and estimated age for the CARTs using all values (mean, median, range, SD) for F0 and for F1-F5, 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 F3 followed by F4, 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).

Discussion
For human perception of speaker age, it seems that F0 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 F3 and F4, 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 B1-B5, L1-L5, 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 Markow Models) and NN (Neural Networks), and studies of potentially important age cues using formant synthesis in attempts to synthesize speaker age.
References


