Perceived age: a distracter for voice disguise and speaker identification?

Erik Eriksson1, James Green2, Maria Sjöstrom1, Kirk P H Sullivan1 and Elisabeth Zetterholm1
1Department of Philosophy and Linguistics, Umeå University, Sweden
2Department of Psychology, The University of Otago, New Zealand

Abstract

Listeners assign an age to a speaker based on their voice alone. The accuracy of estimates of speakers’ ages may not be as great as has been suggested. This experiment suggests that listeners may be better able to rank-order voices by perceived age, than providing perceived ages, per se. However, whether perceived age can act as a distracter for voice disguise and lead to correct speaker identification remains an open question.

Introduction

The ability of listeners to estimate a given speaker’s age is, according to Braun and Cerrato (1999), something that is viewed “rather optimistically” (p. 1369). This was also the position taken in Eriksson et al (2003) when chronological age was one of a number of factors examined as an explanation for the rate of selection of the various foil distracter voices used in a two-alternative forced-choice voice line-up perception experiment for which the training voice had been imitations of the target voice. Eriksson et al (2003) reported a selective listener preference for the voice, Foil 4; a preference that could not be easily explained by mean F0, the most prominent phonetic features in the individual phrase of each foil voice, dialect closeness to the target voice nor chronological age. Foil 4 was second nearest in age to the age of the target voice (the voice that had been imitated), yet was selected more frequently than the voice nearest in age (the natural voice of the imitator). Further, Foil 3, was selected more frequently than the voice nearest in age, even though he was 20 years younger. Foil 3 was however, dialectally closest to the target voice.

It is possible, however, that the ages of the voices in the line-up are not perceived in the order of their chronological age, but are consistently perceived in a different order. Perceived age may provide part of the explanation for the preference of voices presented in a voice line-up.

This research question considered in this paper asks if listeners place voices in the same age order irrespective of whether the individual listener assigns perceived ages to the voices that have a large or a small discrepancy between the perceived and the chronological age of the voices.

This paper begins by briefly overviewing research on estimating speaker age before progressing to defining the problem posed in Eriksson et al. (2003). Thereafter, the experimental design and the data analysis approach used to investigate the perceived age of the voices used in the experiments reported in Eriksson et al (2003) are presented, before the results and their explanatory ability are discussed.

Estimating speaker age

The ability of listeners to estimate a given speaker’s age has been investigated in studies, such as Neiman and Applegate (1990), that calculate the correlations between the calendar age of the speaker and the perceived age, and studies, such as that conducted by Cerrato, Falcone and Paolini (1998), that have used percentage correctly identified age groups to assess listener accuracy.

Schötz (2001) in her study of perceived age in Swedish used 18 age groups with five-year age spans. She found that between 50% and 92% of listeners could estimate a speaker’s age within a maximum error of ±10 years (p. 44). Braun (1996) wrote that it appeared “better to use age spans or even general descriptions like ‘very young’, ‘middle-aged’ etc. rather than to attempt absolute age estimations for forensic purposes” (p.72). However, using age groups has the inherent problem of where to place the boundaries; moving the boundaries slightly can cause differences in the percentage correct.

Braun (1996) demonstrated a limitation with correlations between calendar age and perceived age. She argued, “a constant error, such as the overestimation of the age of all speakers by a constant amount (10 years say), would still result in a high correlation even though the as-
both of these approaches implicitly assume (1) that all voices age at the same rate, (2) that the acoustic changes that are perceived as part of the aging process are due to aging and not external feature such as smoking, and (3) that all age-groups are equally able to estimate the age across the entire human being age range. The literature (e.g. Ryan and Burk (1974), Horii and Ryan (1981), Braun (1996), Jacques and Rastatter (1990), and Wright and Stroud (2002)) shows the importance of considering these factors when assess age perceptually. For example, in relation to smoking Braun (1996) wrote: “if there is extralinguistic evidence in a recording that a speaker smokes, several years may have to be deducted from his perceived age in order to assess his chronological age with accuracy” (p. 72).

Considering the impact of these variables upon accuracy and Schötz’s finding of an perceived accuracy of ±10 years for her Swedish listeners, chronological age of the speakers used in the experiments reported in Eriksson et al (2003) is less relevant to foil selection by listeners than perceived age.

The perception test

In order to investigate whether the listeners’ non-target voice selections reported in the data analysis presented in Eriksson et al. (2003) and Zetterholm et al. (2003b) can be (partically) explained by perceived age and to examine whether listeners place voices in the same age order irrespective of ‘actual’ perceived ages, a new experiment was carried out. The design of the experiment differed only in the task the experimental participants were to undertake; the speech material and the order of presentation were identical to the experiments described in Sullivan et al (2002) and Zetterholm et al (2002; 2003a). The participants’ task in this perception test was to write down the age of the speakers.

Participants

Two groups of participants took part in the investigation: One group from Umeå and one group from Lund. The participants were randomly selected and no listener reported any hearing damage. Tables 1 and 2 present the descriptive statistics for the participant groups.

### Table 1. The descriptive statistics for the participants who took part in the perception experiment in Umeå.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Age (mean)</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>24</td>
<td>31.1</td>
<td>13.6</td>
</tr>
<tr>
<td>Male</td>
<td>25</td>
<td>31.7</td>
<td>12.3</td>
</tr>
<tr>
<td>Total</td>
<td>49</td>
<td>31.4</td>
<td>12.8</td>
</tr>
</tbody>
</table>

### Table 2. The descriptive statistics for the participants who took part in the perception experiment in Lund.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Age (mean)</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>75</td>
<td>20.7</td>
<td>10.9</td>
</tr>
<tr>
<td>Male</td>
<td>78</td>
<td>20.1</td>
<td>10.5</td>
</tr>
<tr>
<td>Total</td>
<td>154</td>
<td>20.6</td>
<td>10.7</td>
</tr>
</tbody>
</table>

The voices

The voices consisted of a set of six recordings of the same text, a political speech. The recordings were an original by Carl Bildt (PS-Bildt), the natural voice of the professional imitator who provided the imitations used in the experiments re-analyzed in Eriksson et al (2003) (PS-AM), and four other male voices, referred to hereafter, following the terminology of Eriksson et al (2003) as Foil 1–4. Table 3 reports information about these voices and the duration and F0 of the recordings.

The experiment

The experiment was constructed using the line-up used in Sullivan et al (2002) and Zetterholm et al (2002; 2003a). The line-up was constructed from the six recordings of the political passage i.e. recordings by PS-Bildt (the earlier experiment’s target voice), PS-AM (the natural voice of the imitator, who imitations of the target voice acted as the training voice in the earlier experiments) and the four other male voices (Foils 1–4). Three separate segments were spliced out from each of the six recordings. Each segment was repeated three times in the line-up, giving a total of 54 speech stimuli in the line-up (3 repetitions x 3 speech samples x 6 speakers). These stimuli were saved to CD in random order.

In the experiment the listeners did not hear a training voice as in the earlier experiments (e.g. Sullivan et al 2002); they were asked to write down the age of the speaker for each of the 54 speech stimuli they would hear.
Table 3: The voices, their age and dialect background, and each recording’s duration, F0 mean and standard deviation.

<table>
<thead>
<tr>
<th>Voice</th>
<th>Age</th>
<th>Dialect Background</th>
<th>Duration</th>
<th>F0 mean</th>
<th>F0 sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS-Bildt 45c</td>
<td>Halland</td>
<td>31 sec</td>
<td>135</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>PS-AM 40c</td>
<td>Västergötland</td>
<td>34 sec</td>
<td>123</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Foil 1 28</td>
<td>Jämtland</td>
<td>32 sec</td>
<td>132</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>Foil 2 20</td>
<td>Västerbotten</td>
<td>38 sec</td>
<td>134</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Foil 3 22</td>
<td>Halland</td>
<td>29 sec</td>
<td>137</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Foil 4 54</td>
<td>Skåne</td>
<td>34 sec</td>
<td>118</td>
<td>46</td>
<td></td>
</tr>
</tbody>
</table>

Data Analysis

Each participant’s three estimates of perceived age for each speech sample and voice were added together and averaged. The resultant averages for each participant were then rank-ordered, 1–6, for each speech sample. Kendall’s Tau was applied to the rank-ordered sequences; Kendall’s Tau is more appropriate that Spearman’s Rho when there are less than 20 ranks.

An analysis of variance was conducted for each of the three speech samples individually and overall with the within-subjects factors the rank for each of the six voices (rank foil 1, rank foil 2 etc.), and the between-subject factors gender (male, n = 97: female, n = 93) and age (18 and under, n = 119, mean age = 16.3, sd = 0.62, and 19+, n = 71, mean age = 34.82, sd = 13.3).

Results

Table 4 presents the overall mean perceived age and standard deviation, and the average rank mean and standard deviation. Kendall’s Tau correlation coefficient for the ranked voices is $t(191)=0.82$, $p<0.001$.

The main effect for the within-subjects variable of ‘voice’ is significant, Multivariate Pillai’s $F(5,182)=671.1$, $p<0.001$. There are linear, $F(1,186)=2934.9$, $p<0.001$, and quadratic, $F(1,186)=30.423$, $p<0.001$, trends to the data.

The ‘voice x age’ interaction is significant, Multivariate Pillai’s $F(5,182)=13.4$, $p<0.001$. There is no linear trend to the data, $F(1,182)=0.04$, $p=0.835$, however there is a quadratic trend, $F(1,182)=31.2$, $p<0.001$. Figure 1 shows the ‘voice x age’ interaction.

Examining the three speech samples separately, the following differences in significance are found. For sample 1, there is no significant linear trend to the main effect for the within-subjects variable of ‘voice’; for samples 2 and 3, no differences in significance are found.

The correlations between perceived and actual ages for the participants ranged from $r=0.17$ to 1.0 with a mean of 0.816 and a standard deviation of 0.167.

Discussion and conclusion

Table 4 shows that the average ranking of the voices’ ages was correct. This together with a Kendall’s Tau for the ranked voices of $t(191)=0.82$, $p<0.001$, indicates consistent ranking of the voices between the participants. Table 4 also shows that Foil 1 has a perceived age that is greater than his chronological age and that this perceived age is only slightly lower than the perceived age of PS-AM. However, this small difference in perceived age leads to a strong difference in ranking, 3.28 cf 4.04.
Investigating the rankings in more detail reveals no gender or age effect, yet reveals an interaction of age and voice. Figure 1 shows how both groups of listeners have similar rankings for Foils 2, 3 and 1; these are the chronologically younger speakers. However, the younger group has placed PS-AM and Foil 4 with similar average ranks that are younger than the rank given to PS-Bildt. The older listeners clearly rank these voices in their chronological age order. This agrees with Wright and Stroud’s (2002) finding that participants are better at identifying individuals of their own age.

The high selection rate for Foil 4 that was reported in Eriksson et al (2003) is, however, no more easily explained by the ranked order of the perceived age of the voices as this concurs with the chronological order of the speakers’ ages. For the younger age group Foil 4 has an average ranking that differs little from PS-AM, the natural voice of the imitator — the voice heard prior to the voice recognition tasks reported in Eriksson et al (2003). This could account for some element of Foil 4’s selection. PS-AM’s low selection rate, in spite of his age position, suggests he succeeded in disguising the non-perceived age aspects of his voice.

The selection of Foil 2 has to be attributed to his dialectal similarity to the target voice as his perceived age and its rank in the voices in the line-up motivate this voices non-selection.

This experiment suggests that listeners may be better able to rank-order voices by perceived age, than providing perceived ages, per se. Whether perceived age can act as a distracter for voice disguise and lead to correct speaker identification remains an open question.

Acknowledgement

This paper was supported by The Bank for Sweden Tercentenary Foundation through their funding of the project “Imitated voices: A research project with applications for security and the law”. Dnr K2002-1121:1, Dnr K2002-1121:2.

References


