Testing the Relevance of Speech Rate, Pitch and a Glottal Chink for the Perception of Age in Synthesized Speech Using Formant Synthesis

Ralf Winkler
Institute of Language and Communication, Technical University Berlin, Germany
ralf.winkler@tu-berlin.de

Abstract

Listeners are able to rate a speaker’s age with reasonable accuracy. However, it is still controversial which features reliably signal a speaker’s age. This paper presents results of a synthesis study, where speech rate, pitch, and a glottal chink were varied systematically over a range that effectively occurs in natural speech to shift the mean perceived age.

The strongest impact on age judgements was found for (i) speech rate, followed by (ii) the glottal chink, while the impact of pitch was only marginal. Some interactions (iii) between the parameters were observed as well.

Results regarding (i) and (ii) show, that formant synthesis is capable of producing speech considerably varying in its mean perceived age even if only a small number of features are manipulated. Regarding (iii), results indicate, that in the study of the impact of selected features their interactions should be considered too.

Index Terms: aged voice, elderly, speech perception, speech synthesis

1. Introduction

Previous research has shown that listeners can estimate a talker’s age quite accurately based on listening to speech sounds alone [[1],[2]]. Several features such as \( F_0 \), jitter, shimmer and spectral tilt as well as temporal features like segment durations and pauses have been identified as markers of chronological and perceived age (for details see [3]).

The influences of single speech features as well as combinations of them on the perception of a speaker’s age is still controversial. While elderly speakers tend to speak slower [4], listener’s judgement is sometimes strongly influenced by tempo [1], but sometimes only weak [5]. Although \( F_0 \) seems to vary systematically between chronologically young speakers and the elderly (e.g. [6]), the impact on listener’s judgement sometimes is strong [7], but another time weak [8].

Increased vowel spectral noise was found for chronologically old men with less physiological fitness [9]. The spectral noise may be partly explained by a glottal chink. The increased incidence of a glottal chink for chronologically old men has been shown in [3]. A glottal chink is an opening in the glottis during vocal fold adduction and allows oral pressure to increase, leading to an increase in friction. The presence of a glottal chink further prevents the flow derivative from changing abruptly when the vocal folds close, leading to a loss of high-frequency energy [10]. Influenced by this observation, in a previous study we analyzed the relation between the EGG Open Quotient (EGG OQ) and the mean perceived age of natural vowels [11]. Results have shown, that for male more than for female speakers the EGG OQ was significantly correlated with the mean perceived age. Higher EGG OQ-values were associated with higher mean perceived age judgements.

Synthesis experiments could give some insights into the impact of single speech features on the perception of age. Shrivastav et al. [12] re-synthesized young and old male natural speech samples by systematically manipulating pitch and speech rate to shift the perceived age of the groups towards each other. A significant shift was observed for the the older, but not for the younger, voices.

An approach to synthesize speech with an intended mean perceived age was introduced by Schötz [13]. She used four female speakers to be representative for an age episode. By weighted linear interpolation of 23 formant-synthesizer parameters all at once she aimed to produce speech with a perceived age ranging from 10 to 80 years. She successfully demonstrated that speech considerably varying in the mean perceived age can be synthesized using data-driven formant synthesis. However, her approach to manipulate all involved parameters at once does not allow for determining the relevance of selected features and interactions.

In the present study we combined two aspects of the studies mentioned above. Our synthesis approach should lead to synthesized speech varying in the mean perceived age. Furthermore, the impact of selected features on the age judgements should remain traceable.

2. Material & Methods

2.1. Natural speech database

A database of 23 single words spoken by 30 female and 30 male subjects was recorded with speakers young and old one half each. The mean age of the female speakers was 26.3 (young) and 69.6 (elderly) years. For the group of male speakers the mean age was 25.5 (young) and 66.8 (elderly) years.

Several acoustic and temporal features were measured in order to restrict the pitch and speech rate variations to a realistic range. Other features (e.g. group mean formant values) were used in the specification of the high-level synthesis parameters to avoid a possible judgement bias that could be regarded to e.g. low static formant targets.

2.2. Synthesis approach

For synthesis the commercially available synthesizer HLsyn (Sensimetrics) was used. HLsyn is a high-level formant synthesizer that is based on a hybrid articulatory-auditory model of speech production [10]. A Matlab environment was developed to calculate the HL-parameter trajectories. Every phone was defined in terms of articulatory events, that influence at least one of the 13 variable parameters. In an initial step the single articulatory events were concatenated. The next step was to apply
rules for adapting the formant transitions to the corresponding consonantal environment. Finally $F_0$ and subglottal pressure were manipulated to generate the appropriate prosody.

Values for the articulatory events (e.g. formant targets) were taken from the natural stimuli (see sec. 2.1) if possible. Speed values for the articulators were adopted from [14]. To synthesize stimuli with a female voice, the pitch values were adapted based on the natural speech database. Male formant targets were multiplied by a factor of 1.2 to account for different vocal tract lengths of men and women.

Three German words were synthesized: /libanen:za/ (Lebanese), /lavi:n@/ (avalanche) and /ma:si:f/ (solid). The three words were selected because of their different length (2 to 4 syllables) and the different consonants they are comprised of. Here we focus on the results of two words: /libanen:za/ and /lavi:n@/.

### 2.3. Feature variation

The full stimulus set was produced by varying systematically pitch, speech rate, vowel lengthening in stressed syllables and a glottal chink while keeping all other parameters constant. We will here focus on the results for pitch and speech rate variations with and without a glottal chink. Pitch and speech rate dimensions were sampled at three points. Relative to the mean value, maximum pitch variation is equivalent to an increase and decrease of roughly 44%. Maximum speech rate variation is equivalent to an increase and decrease of roughly 36%.

HLsyn makes available the simulation of a glottal chink by the possibility of explicitly specifying the area of the cartilaginous portion of the glottis (ap in HLsyn). The spectral tilt (TL in Klatt-synthesis), which is defined as an additional decrease in source spectrum amplitude at 3 KHz, is increased when ap is nonzero. In our approach the glottal chink was implemented as a binary feature. The parameter space of our manipulations is given in Table 1.

### Table 1: Range of feature space spanned by speech rate ($SR$) [Phonem/s], pitch (first value in Hertz [Hz]/ second value in semitones [sm]) and glottal chink ($GC$, binary) $[mm^2]$.

<table>
<thead>
<tr>
<th></th>
<th>Low/ slow/no</th>
<th>Middle</th>
<th>High/ fast/yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_0$</td>
<td>170.0 (5.04)</td>
<td>214.3 (9.04)</td>
<td>270.0 (13.04)</td>
</tr>
<tr>
<td>$SR$</td>
<td>4.5</td>
<td>7.0</td>
<td>9.5</td>
</tr>
<tr>
<td>$GC$</td>
<td>0</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>$F_0$</td>
<td>100.0 (-4.2)</td>
<td>126.1 (-0.1)</td>
<td>159.0 (3.9)</td>
</tr>
<tr>
<td>$SR$</td>
<td>4.5</td>
<td>7.0</td>
<td>9.5</td>
</tr>
<tr>
<td>$GC$</td>
<td>0</td>
<td>-</td>
<td>4</td>
</tr>
</tbody>
</table>

### 2.4. Perception test

Perception tests were done using the software Praat [15]. The listener’s task was to listen to a single word and immediately rate the age of the simulated speaker. Listeners were asked to rate between 15 and 90 years in 5-year steps. Stimuli were presented in random order, starting with all male voices first and then going on with the female voices after a short break. All participants used earphones.

A total of 20, ten female and ten male listeners participated in our perception experiment. The mean age (and SD) of the group of female listeners was 26.7 (2.5) years. Mean age of the male listeners was 30.6 (6.5) years. All participants declared to have normal hearing ability.

#### 2.5. Statistical analysis

In order to describe the overall range of mean perceived age judgements the statistical properties of their distribution are given. The impact of the single factors on the mean perceived age was analyzed by means of an ANOVA. Where significant main effects were identified, pairwise t-tests with Bonferroni-adjustment were applied between two groups each. Statistical analysis was done using R [16].

### 3. Results

#### 3.1. Main effects

By using our approach it was possible to synthesize speech varying in the mean perceived age between 36.8 and 66.4 years. The distribution of the mean perceived age values are reported in Table 2. The mean age judgements range from 36.8 to 63.4 years for the female, and from 42.7 to 66.4 years for the male voices. The range of mean perceived age values is slightly higher for female (26.7 years) than for male (23.7 years) voices. All statistical values reported in Table 2 are slightly higher for male compared to the female voices.

The shift in the perception of age was accomplished by manipulating different speech features. Analysis by means of ANOVA reveals main effects for all three features under investigation. For the female voices we found a significant main effect for speech rate [F(2,342)=34.46, p=0.00] and a glottal chink [F(1,342)=4.63, p=0.03]. The factor pitch [F(2,342)=2.39, p=0.09] might still be interpreted as a trend. Similar main effects were found for male voices, where ANOVA reveals a significant main effect for speech rate [F(2,342)=49.15, p=0.00] and glottal Chink [F(1,342)=2.78, p=0.00], but not for pitch.

#### 3.2. Differences between levels of speech rate

With a range of approximately 20 years the factor speech rate had the biggest impact on the mean perceived age. Mean values corresponding to the different levels of speech rate separated by voice gender are given in Table 3. An increased speech rate leads to a decrease in the mean perceived age of the stimuli. Significant differences for the judgements of female voices were found between slow and normal [p=0.04], slow and fast

#### Table 2: Statistics for the distribution of the mean perceived age values [years]. In the first column “f” refers to female, “m” to male synthesized voices.

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>36.8</td>
<td>46.8</td>
<td>54.0</td>
<td>52.2</td>
<td>57.0</td>
<td>63.4</td>
</tr>
<tr>
<td>m</td>
<td>42.8</td>
<td>49.9</td>
<td>55.7</td>
<td>55.1</td>
<td>61.3</td>
<td>66.4</td>
</tr>
</tbody>
</table>

#### Table 3: Mean perceived age [years] for the different levels of speech rate.

<table>
<thead>
<tr>
<th></th>
<th>slow</th>
<th>normal</th>
<th>fast</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>39.3</td>
<td>34.3</td>
<td>42.9</td>
</tr>
<tr>
<td>male</td>
<td>62.9</td>
<td>56.4</td>
<td>46.0</td>
</tr>
</tbody>
</table>
3.3. Differences between levels of pitch

Mean age judgements for the three levels of pitch are reported in Table 4. Although ANOVA did not reveal a significant main effect, there are still trends to be observed. An increasing pitch leads to a small decrease in the mean perceived age in female voices, while leading to an increase in male voices.

3.4. Differences between levels of a glottal chink

The mean perceived age for the female voices synthesized without a glottal chink was 50.4 years. For the male voices with a glottal chink the mean perceived age was 57.4 years and 52.8 years without. Obviously a glottal chink leads to increased mean perceived age values regardless of the gender of the synthesized voice. No additional t-tests were performed because the factor glottal chink has two levels only. ANOVA results already revealed significant differences between the judgements of the voices with and without a glottal chink for female and male voices (see section 3.1).

3.5. Interactions

Although no interaction in a statistical sense occurred, the influence of one feature seems to vary dependent on the value of the remaining cues. Considering Figure 1(a) for female voices the increase of $F_0$ leads to lower judgements most strongly when speaking fast, while judgements remain relatively constant when speaking normal or slow. Contrary, for male voices (see Figure 2(a)) judgements seem to be influenced by $F_0$, primarily when speaking slow. Regardless of the simulated voice gender this effect is apparent in the stimuli synthesized without a glottal chink only (cp. left vs. right panel of Fig. 1 and Fig. 2).

4. Discussion & Conclusion

With our approach it was possible to synthesize speech with a mean perceived age ranging from approximately 37 to 66 years. This age variation was obtained by varying speech rate, pitch, and a glottal chink.

In the study of male voices by Shrivastav et al. [12] the absolute difference between the unmodified old voices and those with modified $F_0$ and speaking rate was roughly six years. In our study a range of roughly 24 years was captured. Several reasons may account for that difference. Firstly, Shrivastav et al. have manipulated $F_0$ and speaking rate in only one direction to shift perceived age downwards. Thus, half of our range, roughly twelve years, should be compared with the range of six years by Shrivastav et al. Our wider range of perceived age values could be further explained by the magnitude of cue modifi-cations. While in [12] $F_0$ and speaking rate was modified by 20% each, in our study the magnitude of modification was approximately twice as large. Finally, the considerably wide range of mean perceived age values in our study may rely on the nature of the synthesized speech samples. While in our study the stimuli vary in pitch, speaking rate, and a glottal chink only, the stimuli in the study of Shrivastav et al. might contain several other features (e.g. roughness) actually not under investigation but probably still important for signaling a speaker’s age.

Schötz [13] successfully demonstrated that formant synthesis is principally capable of synthesizing speech with an intended mean perceived age. In addition to the ability to synthesize speech considerably varying in the perceived age our approach further allows for easily tracing the impact of single cue modifications on the perceived age.

In our study a very limited set of features has been manipulated. This features were selected because they are frequently specified in the literature to be characteristic of a talker’s age.
However, in future work feature selection should be based on discriminant analysis as a way of selecting key properties that distinguish age episodes in natural speech. Up to now no conclusion can be drawn whether our cues make up an optimal feature set to synthesize a speaker’s age.

Our results show, that synthesizing speech considerably varying in the mean perceived age is possible using formant synthesis even by manipulating only three speech features known to change with a speaker’s chronological age. The strongest impact on the mean perceived age was measured for speech rate, followed by the simulation of a glottal chink. There was a marginal but systematic impact of pitch.

5. Acknowledgements
Parts of this work were supported by the Deutsche Forschungsgemeinschaft (German Research Fund, SE 462/5-1 and SE 462/5-2). The author would like to thank the reviewers for many helpful comments and suggestions on a previous version of this paper.

6. References

Figure 2: Means (with 95% confidence interval) of the age judgements (male voices) for the three levels of pitch. Lines represent the levels of speech rate. Left figure shows judgements for the speech synthesized without, right picture with a glottal chink.