Title: Determining the Impact of Selected Parameters on the Perception of Age from Voice

Author: Ralf Winkler, Institute of Language and Communication, Technical University Berlin, Germany

Previous research has shown that listeners can estimate a talker's age quite accurately based on listening speech sounds alone. Several parameters have been identified as markers of chronological and perceived age. However, the influence 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, listener's judgment is sometimes strongly influenced by speech rate, but sometimes only weak. In perception studies the same ambiguity is found for fundamental frequency, although it seems to vary systematically with speakers' chronological age.

In our previous work we recorded a database of 23 single words spoken by 15 female and 15 male subjects for each age group (young adult vs. elderly). While significant differences were found for speech rate, the differences in fundamental frequency, commonly discussed together with voice changes related to aging, were marginal.

In previous work increased spectral noise was found in the voice of elderly men, at least for the speakers with less fitness. The increase in spectral noise could be partly explained by a glottal chink during phonation. An increased incidence of the occurrence of a glottal chink has already been shown by means of visually inspection. Given the assumption that the spectral noise is a consequence of a glottal chink during phonation, an increased Open Quotient (QO) should be determined for the elderly as well. To test this assumption in our previous study we analyzed the relation between the OQ (EGG) and the mean perceived age in natural sustained vowels varying considerably in their OQ-values. The mean perceived age was significantly correlated with the OQ-values; for the males somewhat stronger than for the female voices.

In a previous work young and old male natural speech samples were re-synthesized 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 older, but not younger, voices. This approach allows for tracing the effect of the single parameter changes as well as the combined impact. A more general approach was previously done by formant-synthesizing speech with an intended target age by linearly interpolating 23 parameters at a time, determined from natural speech of four female talkers. Her results indicate, 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.

The aim of this work is determining the relevance of speech rate, fundamental frequency and a glottal chink in the perception of age from voice. We combine the conveniences of the two studies mentioned above. Our synthesis approach leads to formant-synthesized speech varying in the mean perceived age. Furthermore, the impact of selected features on the age judgments remains traceable.

Our approach consists of synthesizing single word stimuli by simultaneously varying selected speech features using the limits determined form natural speech recordings mentioned in our previous work. The stimuli were then rated by a group of listeners regarding the perceived age. Finally, the mean perceived age values were analyzed with respect to the feature variations to determine their impact on the perception of age.

For synthesis the commercial available synthesizer HLsyn (Sensimetrics) was used. HLsyn is a high-level formant synthesizer that is based on a hybrid articulatory-acoustic model of speech production. A Matlab environment was developed to calculate the HL-parameter trajectories. Each phoneme was defined in terms of articulatory events that influence at least one of the 13 variable parameters of HLsyn. 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 fundamental frequency and subglottal pressure were manipulated to generate an appropriate prosody. Values for the articulatory events (e.g. formant targets) were taken from the natural stimuli if possible. Speed values for the articulators were adopted from the HLsyn manual. To produce 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: /libane:z@/ (Lebanese), /lavi:n@/ (avalanche) and /masi:f/ (solid). Here we focus on the results of two words: /libane:z@/ and /lavi:n@/, while results of the remaining word will be reported elsewhere.

The stimulus set was produced by varying systematically pitch, speech rate, lengthening 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 dimension was sampled at three points. Relative to the center, 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. The spectral tilt, which is defined as the additional decrease in the source spectrum amplitude at 3 kHz, is increased when the cartilaginous portion of the glottis is nonzero. In our approach the glottal chink has been implemented as a binary feature.

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 of the listeners was around an age of 27-30 years.

The impact of the single parameters (factors) on the mean perceived age was analyzed by means of ANOVA. Where significant main effects were identified pairwise t-tests with Bonferroni adjustment were applied between two groups each.

The strongest impact on age judgments was found for (i) speech rate, followed by the existence of (ii) a 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) demonstrate, that formant synthesis is capable of producing speech considerably varying in their mean perceived age even if only a small number of features were manipulated. Regarding (iii), results indicate, that in the study of the impact of selected features their interactions should be considered as well.