F_0 and Segment Duration in Formant Synthesis of Speaker Age

Susanne Schötz

Linguistics and Phonetics, Centre for Languages and Literature
Lund University
susanne.schotz@ling.lu.se

Abstract

This paper describes the work with F_0 and segment duration when developing a prototype system for analysis of speaker age using data-driven formant synthesis. The system was developed to extract 23 parameters from the test words—spoken by four differently aged female speakers of the same dialect and family—and to generate synthetic copies. Audio-visual feedback enabled the user to compare the natural and synthetic versions and facilitated parameter adjustment. Next, weighted linear interpolation was used in a first crude attempt to synthesize speaker age. Evaluation of the system revealed its strengths and weaknesses, and suggested further improvements. F_0 and duration performed better than most other parameters.

1. Introduction

In speech synthesis applications like spoken dialogue systems and voice prosthesis, there is a growing need for voice variation in terms of age, emotion and other speaker-specific qualities. To contribute to the research in this area, as part of a larger study aiming at identifying phonetic age cues, a system for analysis by synthesis of speaker age was developed using data-driven formant synthesis. This paper briefly describes the developing process, with focus on the work devoted to to F_0 and duration.

Research has shown that acoustic cues to speaker age can be found in almost every phonetic dimension, e.g. in F_0, duration, intensity, resonance, and voice quality [7, 8, 11, 14]. However, the relative importance of the different age cues has still not been fully explored. One reason for this may be the lack of an adequate analysis tool where a large number of potential age parameters can be varied systematically and studied in detail.

Several phonetic cues to age have been found in previous studies, including F_0 and duration (speech rate). In adult women F_0 remains fairly constant until menopause, when a drop usually occurs [11]. After this drop, F_0 remains stable or continues to decrease. Old women and men show lower mean F_0 values and higher measurements of F_0 range and SD than middle-aged and young speakers [14]. In a study of female speaker age, Brückl and Sendlmeier [2] found that (1) decreasing F_0 was a fairly good predictor of increasing age, especially for spontaneous speech, (2) there were general, but only minor, correlations of F_0 perturbations and age and (3) slower speech rate correlated with increased age in read speech. However, children tend to speak more slowly than adult speakers [6]. A comprehensive summary of previos studies of speaker age is given in [11].

Formant synthesis generates speech from a set of rules and acoustic parameters, and is considered both robust and flexible. Still, the more natural-sounding concatenation synthesis is generally preferred over formant synthesis [12]. Lately, formant synthesis has made a comeback in speech research, e.g in data-driven and hybrid synthesis with improved naturalness [4, 13].

2. Material

The best speech material for developing the system would consist of lifelong longitudinal recordings of the same speaker. Since no such recordings were found, four very similar and closely related female non-smoking native speakers of the same Swedish dialect were selected to represent four different ages:

- Speaker 1: girl (aged 6)
- Speaker 2: mother (aged 36)
- Speaker 3: grandmother (aged 66)
- Speaker 4: great grandmother (aged 91)

The speakers were recorded in their homes with a Sony portable DAT recorder TCD-D8 and a Sony tie-pin type condenser microphone ECM-T140 at 48kHz/16 bit sampling frequency. After listening to the recordings, the isolated word 'själen' ['ɧɛ:lən] (the soul), was selected as a first test word for developing the system. The recordings were segmented into words and phonemes, resampled to 16 kHz, and normalized for intensity.

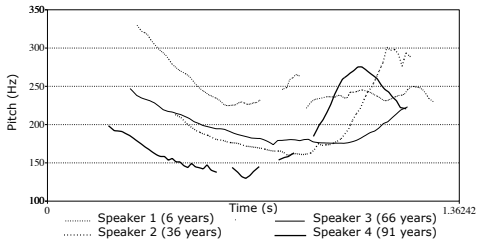

Figure 1: *F_0 contours for ['(ɧ)ɛ:lən] for the four speakers.*

Table 1: *Segment durations for the four speakers (ms)*

Segment	*Speaker 1*	*Speaker 2*	*Speaker 3*	*Speaker 4*
[ɧ]	139	207	128	106
[ɛ:]	469	291	297	320
[l]	312	89	83	105
[ə]	179	110	147	136
[n]	111	117	72	156
['ɧɛ:lən]	1209	813	728	822

An acoustic pre-analysis of F_0 and duration for the test words are shown in Figure 1 and Table 1. As expected, Speaker 1 displayed the highest and Speaker 4 the lowest F_0 values. These speakers also had produced the longest word durations. However, Speaker 3, who reported that her voice was often judged as being younger than her chronological age, displayed a slightly higher F_0 as well as a shorter word duration than the younger Speaker 2. Although these values would be likely to influence the results, it was decided to use this material anyway, keeping in mind that Speaker 3 might sound atypical for her age.

3. Method

3.1. Tools

For the acoustic analyses, the speech analysis software Praat [1] was used. Because of its ability to display waveforms, spectrograms and spectra on the computer screen, Praat also served as the graphical user interface, and was the program from where the other tools were called. The synthesis was generated with an internal and non-public software version of the GLOVE formant synthesis system along with the small script Dat-convert, which converted parameter files to the GLOVE format. GLOVE, which is an extension of the cascade formant synthesizer OVE III [10], with an expanded LF voice source model [5], has been used for experiments with voice variations since the late 1980s [3, 9]. For a more detailed description, see Carlson et al. [3]. GLOVE and Dat-convert were used by kind permission of the Centre for Speech Technology at the Royal Institute of Technology in Stockholm. Additional programs were developed in the Java and Perl programming languages.

3.2. Procedure

The prototype system was developed in several steps. First, parameters were extracted from the natural words and used to generate synthetic copies. Next, the parameters were adjusted to generate more natural-sounding synthesis. A schematic overview of the system can be seen in Figure 2. When acceptable synthetic versions had been obtained from all four natural speakers, the system was used in an initial experiment to synthesize speaker age by interpolation between parameters of two speakers. Below, the steps are explained in more detail.

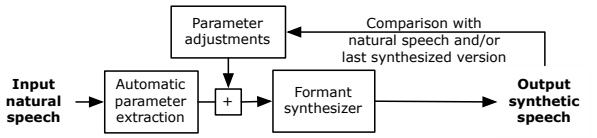

Figure 2: *Schematic overview of the protoype system.*

3.2.1. Parameter extraction

A Praat script was developed to extract 23 parameters (see Table 2) every 10 ms, to store the values in files, and to use them as input to the GLOVE synthesizer. Formants and F_0 were extracted first, followed by amplitudes, RG, RK, and FA. NA was then added to introduce a small amount of pitch synchronous noise for breathy voice, and DI, which simulates creaky voice, was the last parameter to be integrated.

To be able to compare the natural speech to the synthetic versions, another Praat script was developed, which first called the parameter extraction script, and then displayed waveforms and spectrograms of the original word, the resulting synthetic word, as well as the previous synthetic version. By auditive and visual comparison of the three files, the user could easily determine whether a newly added parameter or adjustment had improved the synthesis. Figure 3 shows such a screenshot for Speaker 2, where the user also has applied the spectral slice function in Praat to compare [fj].

3.2.2. Parameter adjustment

Several adjustments—sometimes systematically and sometimes using ad hoc methods—were made to improve the synthesis.

Table 2: *The 23 GLOVE parameters used in the system*

Parameter	*Description*
F1-F4, B1-B4	Formant frequencies and their bandwidths
FH	Higher pole correction with 3 double poles (FH, FH*1.2 and FH*1.4) with fixed bandwidths (set to F5)
K1-K2, C1-C2	Fricative formant frequencies and their bandwidths (K1 set to F2, K2 to F3, C1 to B2, C2 to B3)
AK	A zero for fricatives
F0	Fundamental frequency
AC, AH	Noise amplitudes for frication (AC) and aspiration (AH)
A0	Voice amplitude
RG	Glottal shape
RK	Glottal pulse skewness factor
FA	Frequency above which an extra 6dB/octave is added to the spectral tilt
NA	Noise Added, mixing of noise into the voice source
DI	Simulation of diplophonia, creak or laryngalisation

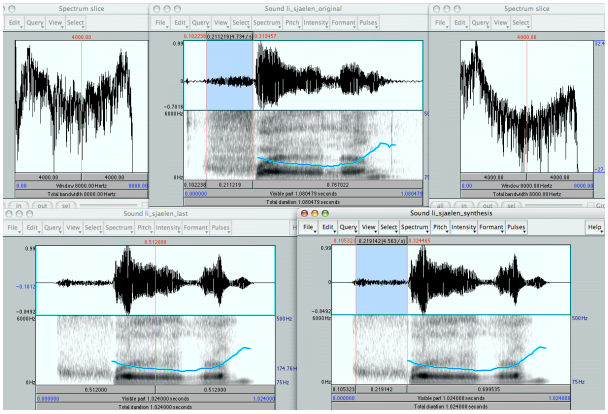

Figure 3: *Display of waveforms and spectrograms for the natural (top), the previous (left) and latest (right) synthetic words, as well as natural (left) and synthetic (right) spectra for [fj].*

After each adjustment, the sound and acoustic diagrams of the resulting synthesis were compared to the natural and previously synthesized versions. Whenever an adjustment had improved the synthesis, it was added to the set of adjustment rules.

F_0 needed several adjustments. Occasionally, F_0 was identified in voiceless segments, and creaky segments were often analyzed as voiceless or as having a very high F_0. This was solved by adjusting the arguments to the pitch analysis in Praat, and by adapting a second pitch analysis especially for F_0 contours below 150 Hz. This additional analysis was used whenever the analysis failed to find reasonable F_0 values in voiced segments. Also, the DI parameter was activated to simulate creak.

Duration was determined by the number of 10 ms frames extracted from the natural words. It did not need any adjusting.

Formants and amplitude parameters caused the most serious problems, including distortion. By smoothing the parameter curves, the synthesis was improved considerably.

3.2.3. Parameter interpolation and synthesis of age

A first attempt to synthesize speaker age was carried out using the system. The basic idea was to use the synthetic versions of the four words to synthesize new words of other ages by age-weighted linear interpolation between two source parameter files. A small Java program was developed to calculate the weights and to perform the interpolations. For each target age provided as input by the user, the program selects the parameter files of two source speakers (the older and younger speakers

closest in age to the target age), and generates a new parameter file from the interpolations between the two source parameter files. For instance, for the target age of 51, i.e. exactly half-way between the ages of Speaker 2 (aged 36) and Speaker 3 (aged 66), the program selects these two speakers as source speakers, and then calculates the age weights to 0.5 for both source speakers. Next, the program calculates the target duration for each phoneme segment using the age weights and the source speaker durations. If the duration of a particular segment is 100 ms for source speaker 1, and 200 ms for source speaker 2, the target duration for the interpolation is 200 x 0.5 + 100 x 0.5 = 150 ms. All parameter values are then interpolated in the same way. Finally, the target parameter file is synthesized using GLOVE, and displayed (waveform and spectrogram) in Praat along with the two input synthetic words for comparison. An overview of the procedure is shown in Figure 4.

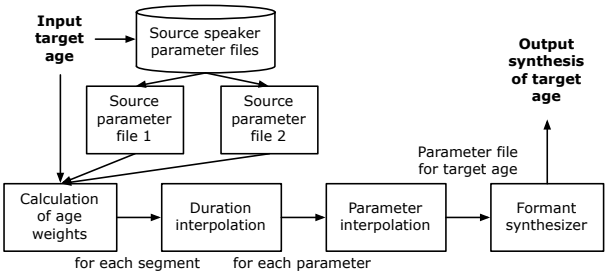

Figure 4: *Schematic overview of the age interpolation method.*

4. Results and evaluation

A first look at the results of the system showed that although there were similarities between the natural and synthetic versions, there were differences as well. The synthetic words generally sounded more muffled than the natural ones, and problems with formants and amplitudes influenced the results. However, F_0 and duration turned out well, as can be seen in Figure 5.

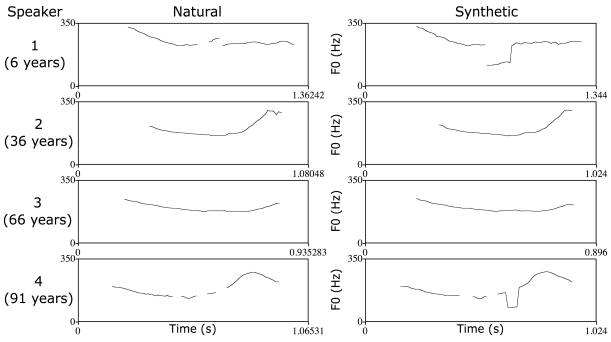

Figure 5: *Natural and synthetic F_0 contours for all speakers.*

There is a considerable resemblance between the natural and synthetic F_0 contours for all four speakers. Only a few minor differences can be observed, since the F_0 values were extracted only once every 10 ms. Also note the halved F_0 in the creaky parts of the synthetic versions for Speakers 1 and 4, which successfully simulated creak.

The sound file durations shown in Figure 5 are slightly shorter in the synthetic versions, mainly because 10-20 ms at the beginning and end were lost in the parameter extraction.

To evaluate the system's performance, a listening test was carried out. 31 students of phonetics listened to the stimuli and judged direct age (in years) and naturalness (on a 7-point scale, where 1 is very unnatural and 7 is very natural). Stimuli for the age estimation task consisted of the four natural and the four corresponding synthetic versions along with age interpolations for eight decades from 10 to 80 years. The natural version of Speaker 3 was used twice to test judging consistency.

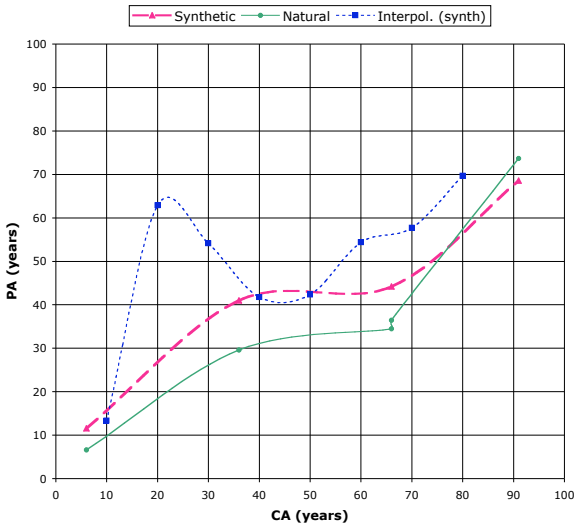

Figure 6: *Correlations between chronological (CA) and perceived (PA) age for natural, synthetic and interpolated stimuli.*

Figure 6 shows correlations between chronological age (CA) and mean perceived (PA) age for the natural, synthetic and interpolated stimuli. The curves for the natural and synthetic words are quite similar, though the synthetic versions were judged to sound older in most cases. Speaker 3 (66 years) was always estimated to be younger than her chronological age. The interpolations were mostly judged as older than both the natural and synthetic words. Especially striking is that the interpolations for 20 and 30 years were both judged to be older than 50 years.

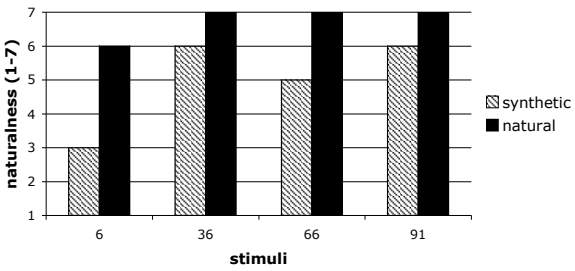

Figure 7: *Evaluated naturalness for natural and synthetic stimuli (median, 7-point scale).*

In the naturalness evaluation task, the eight natural and synthetic versions were used as stimuli. Figure 7 shows that the natural words were always judged as more natural than the synthetic ones. Moreover, both the natural and the synthetic versions of Speaker 1 (6 years) were judged to be the least natural.

5. Discussion and future work

5.1. F_0 and duration

While developing the prototype system, F_0 and duration generated fewer problems and performed better than most other parameters, including formants and amplitudes.

F_0 caused initial pitch analysis problems in the parameter extraction, especially for very low frequencies. A number of adjustment rules were added, and in the current version creaky voice quality is simulated successfully with a combination of a halved F_0 value and the diplophonia (DI) parameter. However, the two different intonation types produced by the four speakers (i.e. declarative and list intonation) led to interpolations between dissimilar intonation patterns, which may not be optimal. This problem has to be solved when expanding the system to handle more speakers and longer speech samples.

Duration was determined exclusively by the number of 10 ms frames in the parameter extraction, and neither this nor the segment duration interpolation did generate any problems. However, it is difficult to tell to what extent segment duration reflects age in isolated words. Additional data are needed to learn more about segment duration in relation to the aging process. Once longer speech samples or words with very short segments, e.g. stop consonant releases, are integrated in the system, duration needs to be handled more carefully. Such short sounds may not be captured correctly if parameter values are extracted only once every 10 ms.

5.2. Synthesis and interpolation

The synthetic words obtained a reasonable resemblance with the natural words in most cases. Remaining problems include muffled sound quality and uneven formant trajectories in the synthesis, which may be solved using a pre-emphasis filter and a better formant analysis and/or smoothing algorithm. After these improvements it is likely that listeners will judge the synthetic words to be closer to the age of the natural versions.

The interpolated versions were often judged as older than the intended age. Several explanations for why this happened are possible. Formant error in the parameter extraction for Speaker 1 generated tremor and rough voice quality in the synthesis, which may have led to an older impression of her voice when it was used in the interpolations. That Speaker 3 was judged as much younger than her chronological age may also have influenced the results. A third very important explanation is that linear interpolation is indeed a crude simplification of the human aging process, which is far from linear. Moreover, some parameters may change during a certain period of aging, while others remain constant. Therefore, the interpolation method described in this paper should be considered only a starting point for further analysis of speaker age. Better interpolation techniques will have to be tested. One should also bear in mind that the system is likely to interpolate not only between two ages, but between a number of individual characteristics, even when the speakers are closely related. Further research with a larger speech material is needed to identify and rank the most important age-related parameters.

5.3. Future work

Future work involves (1) improved parameter extraction for formants, (2) pre-emphasis filtering to avoid muffled synthesis, (3) better interpolation algorithms, and (4) expansion of the system to handle more speakers, as well as a larger and more varied speech material.

Although a number of problems remain with the prototype system, it is not unlikely that once its performance is improved, the system may well be used in studies of speaker age for analysis, modelling and synthesis. For instance, the system's parameter adjustment function could also serve as a tool for detailed studies of potential age-parameters by systematic variation and continuous audio-visual feedback. The phonetic knowledge gained from such experiments may then be used in future speech synthesis applications to include age—and other speaker-specific qualities as well—leading to more natural-sounding synthetic speech.

6. References

[1] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 4.3.04) [computer program]. Retrieved March 8, 2005, from http://www.praat.org/, 2005.

[2] M. Brückl and W. Sendlmeier. Aging female voices: An acoustic and perceptive analysis. In *Proceedings of VOQUAL'03*, pages 163–168. Geneva, Switzerland, 2003.

[3] R. Carlson, B. Granström, and I. Karlsson. Experiments with voice modelling in speech synthesis. *Speech Communication*, 10:481–489, 1991.

[4] R. Carlson, T. Sigvardson, and A. Sjölander. Data-driven formant synthesis. In *Proceedings of Fonetik 2002*, volume 44, pages 121–124. TMH-QPSR, 2002.

[5] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4:1–13, 1985.

[6] S. Hawkins. On the development of motor control in speech: Evidence from studies of temporal coordination. In N. J. Lass, editor, *Speech and Language: Advances in Basic Research and Practice*, pages 317–374. Academic Press., New York, 1984.

[7] H. Hollien. Old voices: What do we really know about them? *Journal of Voice*, 1:2–13, 1987.

[8] R. Jacques and M. Rastatter. Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners. *Folia Phoniatrica (Basel)*, 42:118–124, 1990.

[9] I. Karlsson. *Analysis and Synthesis of Different Voices with Emphasis on Female Speech*. PhD thesis, Royal Institute of Technology, KTH, Stockholm, 1992.

[10] J. Liljencrants. The OVE III speech synthesizer. *IEEE Trans AU-16*, no 1:137–140, 1968.

[11] S. E. Linville. *Vocal Aging*. Singular Thomson Learning, San Diego, 2001.

[12] S. Narayanan and A. Alwan. *Text to Speech Synthesis: New Paradigms and Advances*. Prentice Hall PTR, IMSC Press Multimedia Series, 2004.

[13] D. Öhlin and R. Carlson. Data-driven formant synthesis. In *Procedings of Fonetik 2004*, pages 160–163. Dept. of Linguistics, Stockholm University, 2004.

[14] S. A. Xue and D. Deliyski. Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications. *Educational Gerontology*, 21:159–168, 2001.