7 Conclusion

Automatic detection of emotions has been evaluated using spectral and pitch features, all modeled by GMMs on the frame level. Two corpora were used: telephone services and meetings. Results show that frame level GMMs are useful for emotion classification.

The two MFCC methods show similar performance, and MFCC-low outperforms pitch features. A reason may be that MFCC-low gives a more stable pitch measure. Also, it may be due to its ability to capture voice source characteristics, see Syrdal (1996), where the level difference between the first and the second harmonic is shown to distinguish between phonations, which in turn may vary across emotions.

The diverse results of the two corpora are not surprising considering their discrepancies.

A possible way to improve performance for the VP corpus would be to perform emotion detection on the dialogue level rather than the utterance level, and also take the lexical content into account. This would mimic the behavior of the human labeler.

Above we have indicated the difficulty to compare emotion recognition results. However, it seems that our results are at least on par with those in Blouin & Maffiolo (2005).

Acknowledgements

This work was performed within CHIL, Computers in the Human Interaction Loop, an EU 6th Framework IP (506909). We thank Voice Provider for providing speech material.

References

Batliner, A., J. Buckow, R. Huber, V. Warnke, E. Nöth & H. Niemann, 1999. Prosodic Feature Evaluation: Brute Force or Well Designed? *Proc. 14th ICPhS*, 2315-2318.

Batliner, A., K. Fischer, R. Huber, J. Spilkera & E. Nöth, 2003. How to find trouble in communication. *Speech Communication 40*, 117-143.

Blouin, C. & V. Maffiolo, 2005. A study on the automatic detection and characterization of emotion in a voice service context. *Proc. Interspeech*, Lisbon, 469-472.

Chul, M.L. & S. Narayanan, 2005. Toward Detecting Emotions in Spoken Dialogs. *IEEE, Transactions on Speech and Audio Processing 13*(2), 293-303.

Dellaert, F., T.S. Polzin & A. Waibel, 1996. Recognizing emotion in speech. *Proc. ICSLP*, Philadelphia, 3:1970-1973.

Gauvin, J-L. & C.H. Lee, 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. SAP 2*, 291-298.

Hermansky, H. & N. Morgan, 1994. RASTA processing of speech. *IEEE Trans. SAP 4*, 578-589.

Langlais, P., 1995. *Traitement de la prosodie en reconnaissance automatique de la parole.* PhD-thesis, University of Avignon.

Laskowski, K. & S. Burger, 2006. Annotation and Analysis of Emotionally Relevant Behavior in the ISL Meeting Corpus. *LREC*, Genoa.

Oudeyer, P., 2002. Novel Useful Features and Algorithms for the Recognition of Emotions in. Human Speech. *Proc. of the 1st Int. Conf. on Speech Prosody.*

Reynolds, D., T. Quatieri & R. Dunn, 2000. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing 10*, 19-41.

Ross, M., H. Shafer, A. Cohen, R. Freudberg & H. Manley, 1974. Average magnitude difference function pitch extraction. *IEEE Trans. ASSP-22*, 353-362.

Scherer, K.R., 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication 40*, 227-256.

Syrdal, A.K., 1996. Acoustic variability in spontaneous conversational speech of American English talkers. *Proc. ICSLP*, Philadelphia.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 105–108

Data-driven Formant Synthesis of Speaker Age

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

Abstract

This paper briefly describes the development of a research tool for analysis of speaker age using data-driven formant synthesis. A prototype system was developed to automatically extract 23 acoustic parameters from the Swedish word 'själen' ['ɧɛːlən] (the soul) spoken by four differently aged female speakers of the same dialect and family, and to generate synthetic copies. Functions for parameter adjustment as well as audio-visual comparison of the natural and synthesised words using waveforms and spectrograms were added to improve the synthesised words. Age-weighted linear parameter interpolation was then used to synthesise a target age anywhere between the ages of 2 source speakers. After an initial evaluation, the system was further improved and extended. A second evaluation indicated that speaker age may be successfully synthesised using data-driven formant synthesis and weighted linear interpolation.

1 Introduction

In speech synthesis applications like spoken dialogue systems and voice prostheses, the need for voice variation in terms of age, emotion and other speaker-specific qualities is growing. 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 and results.

Research has shown that acoustic cues to speaker age can be found in almost every phonetic dimension, i.e. in F_0, duration, intensity, resonance, and voice quality (Hollien, 1987; Jacques & Rastatter, 1990; Linville, 2001; Xue & Deliyski., 2001). However, the relative importance of the different cues has still not been fully explored. One reason for this may be the lack of an adequate analysis tool in which a large number of potential age parameters can be varied systematically and studied in detail.

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 (Narayanan & Alwan, 2004). Lately, formant synthesis has made a comeback in speech research, e.g. in data-driven and hybrid synthesis with improved naturalness (Carlson et al., 2002; Öhlin & Carlson, 2004).

2 Material

Four female non-smoking native Swedish speakers of the same family and dialect were selected to represent different ages, and recorded twice over a period of 3 years: *Speaker:1*: girl (aged 6 and 9), *Speaker 2*: mother (aged 36 and 39), *Speaker 3*: grandmother (aged 66 and 69), and *Speaker 4*: great grandmother (aged 91 and 94). The isolated word 'själen' ['ɧɛːlən] (the soul), was selected as a first test word, and the recordings were segmented into phonemes, resampled to 16 kHz, and normalized for intensity.

3 Method and procedure

The prototype system was developed in several steps (see Figure 1). First, a Praat (Boersma & Weenink, 2005) script extracted 23 acoustic parameters every 10 ms. These were then used as input to the formant synthesiser GLOVE, which is an extension of OVE III (Liljencrants, 1968) with an expanded LF voice source model (Fant et al., 1985). GLOVE was used by kind permission of CTT, KTH. For a more detailed description, see Carlson et al. (1991).

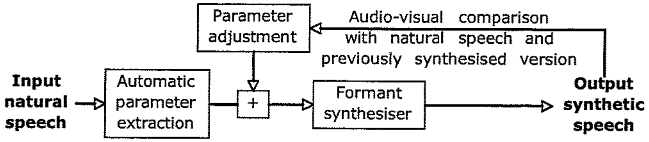
Parameter adjustment
Audio-visual comparison with natural speech and previously synthesised version
Input natural speech
Automatic parameter extraction
+
Formant synthesiser
Output synthetic speech

Figure 1. Schematic overview of the prototype system.

Next, the parameters were adjusted to generate more natural-sounding synthesis. 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. If an adjustment improved the synthesis, it was added to the adjustment rules. Formants, amplitudes and voice source parameters (except F_0) caused the most serious problems, which were first solved using fixed values, then by parameter smoothing.

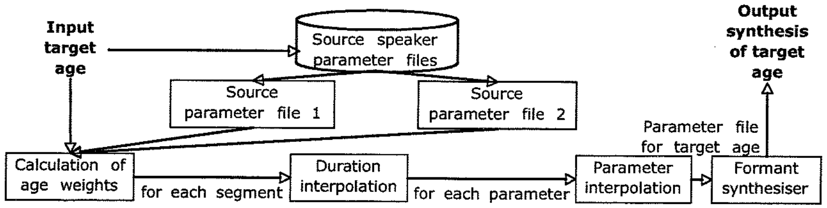

Figure 2. Schematic overview of the age interpolation method.

An attempt to synthesise speaker age was carried out using the system. The basic idea was to use the synthetic versions of the words to generate new words of other ages by age-weighted linear interpolation between two source parameter files. A 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 synthesised using GLOVE, and displayed (waveform and spectrogram) in Praat along with the two input synthetic words for comparison. A schematic overview of the procedure is shown in Figure 2.

4 Results

To evaluate the system's performance, two perception tests were carried out to estimate direct age and naturalness (on a 7-point scale, where 1 is very unnatural and 7 is very natural). Stimuli in the first evaluation consisted of natural and synthetic versions of the 6, 36, 66 and 91 year old speakers. The second evaluation was carried out at a later stage when the 9, 39, 69 and 94 year olds had been included, and when parameter smoothing and pre-emphasis filtering (to avoid muffled quality) had improved the synthesis. 31 students participated in the first evaluation test, also including interpolations for 8 decades (10 to 80 years), while 21 students took part in the second, which also comprised interpolations for 7 decades (10 to 70 years).

4.1 First evaluation

In the first evaluation, the correlation curves between chronological age (CA, or simulated "CA" for the synthetic words) and perceived age (PA) displayed some similarity for the natural and synthetic words, though the synthetic ones were judged older in most cases, as seen in Figure 3. The interpolations were mostly judged as much older than both the natural and synthetic words. As for naturalness, the natural words were always judged more natural than the synthetic ones. Both the natural and synthetic 6 year old versions were judged least natural.

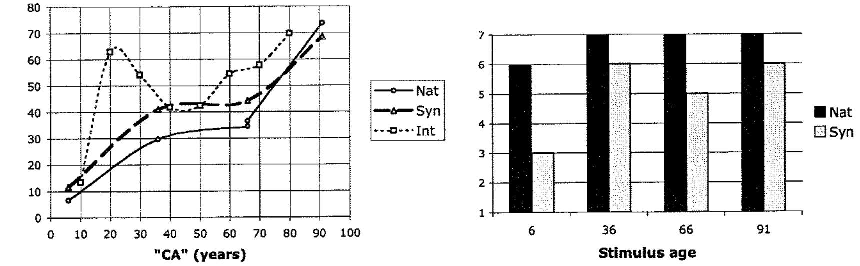

Figure 3. Correlation between PA and CA for natural, synthetic and interpolated words (left), and median perceived naturalness for natural and synthetic words in the first evaluation.

4.2 Second evaluation

Figure 4 shows that not only the correlation curves for the natural and synthetic words, but also for the interpolations did improve in similarity in the second evaluation compared to the first one. However, the natural and synthetic versions of the 39, 66 and 69 year olds were quite underestimated. All natural words were judged as more natural than the synthetic ones, and all synthetic words except the 6 and 94 year old achieved a median naturalness value of 6.

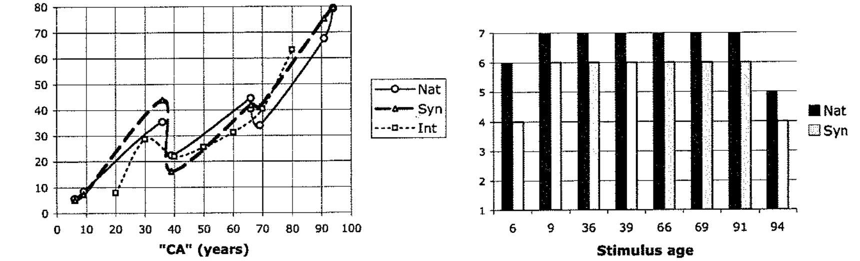

Figure 4. Correlation between PA and CA for natural, synthetic and interpolated words (left), and median perceived naturalness for natural and synthetic words in the second evaluation.

5 Discussion and future work

The synthetic words obtained a reasonable resemblance with the natural words in most cases, and the similarity in age was improved in the second evaluation. The interpolated versions were often judged as older than the intended age in the first evaluation, but in the second evaluation they had become more similar in age to the natural and synthetic versions, indicating that speaker age may be synthesised using data-driven formant synthesis. Still, some of the age estimations were quite unexpected. For instance, the 39, 66 and 69 year olds were judged as much younger than their CA. This may be explained by that these voices were atypical for their age.

One very important point in this study is that synthesis of age by linear interpolation is indeed a crude simplification of the human aging process, which is far from linear. Moreover, while some parameters may change considerably during a certain period of aging (i.e. F_0 and formant frequencies during puberty), others remain constant. 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 also between a number of individual characteristics, even when the speakers are closely related.

Future work involves (1) improved parameter extraction for formants, (2) better interpolation algorithms, and (3) expansion of the system to handle more speakers (of both sexes), as well as a larger and more varied speech material. Further research with a larger material is needed to identify and rank the most important age-related parameters. If further developed, the prototype system may well be used in future studies for analysis, modelling and synthesis of speaker age and other speaker-specific qualities, including dialect and attitude. The phonetic knowledge gained from such experiments may then be used in future speech synthesis applications to generate more natural-sounding synthetic speech.

References

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

Carlson, R., B. Granström & I. Karlsson, 1991. Experiments with voice modelling in speech synthesis. *Speech Communication 10*, 481–489.

Carlson, R., T. Sigvardson & A. Sjölander, 2002. Data-driven formant synthesis. *Proceedings of Fonetik 2002, TMH-QPSR*, 121–124.

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

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

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

Liljencrants, J., 1968. The OVE III speech synthesizer. *IEEE Trans AU-16*(1), 137–140.

Linville, S.E., 2001. *Vocal Aging*. San Diego: Singular Thomson Learning.

Narayanan, S. & A. Alwan (eds.), 2004. *Text to Speech Synthesis: New Paradigms and Advances*. Prentice Hall PTR, IMSC Press Multimedia Series.

Öhlin, D. & R. Carlson, 2004. Data-driven formant synthesis. *Proceedings of Fonetik 2004*, Dept. of Linguistics, Stockholm University, 160–163.

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

How do we Speak to Foreigners? – Phonetic Analyses of Speech Communication between L1 and L2 Speakers of Norwegian

Rein Ove Sikveland

Department of Language and Communication Studies,
The Norwegian University of Science and Technology (NTNU), Trondheim
rein.ove.sikveland@hf.ntnu.no

Abstract

The major goal of this study was to investigate which phonetic strategies we may actually use when speaking to L2 speakers of our mother tongue (L1). The results showed that speech rate in general was slower and that the vowel formants were closer to target values, in L2 directed speech compared to L1 directed speech in Norwegian. These properties of L2 directed speech correspond to previous findings for clear speech (e.g. Picheny et al., 1986; Krause & Braida, 2004). The results also suggest that level of experience may influence L2 directed speech; teachers of Norwegian as a second language slowed down the speech rate more than the non-teachers did, in L2 directed speech compared to L1 directed speech.

1 Introduction

When speaking to foreigners in our mother tongue (L1) it might be natural to speak clearer than normal to make ourselves understood, which implies the use of certain phonetic strategies when speaking to these second language learners (L2 speakers).

Previous findings by Picheny et al. (1986) and Krause & Braida (2004) have shown that clear speech can be characterized by a decrease in speech rate, more pauses, relatively more energy in the frequency region of 1-3 kHz, less phonological reductions (e.g. less burst eliminations), vowel formants closer to target values, longer VOT and a greater F0 span, compared to conversational speech. What characterizes L2 directed speech has not been subject to any previous investigations, but one might assume that strategies in L2 directed speech correspond to the findings for clear speech. This has been investigated in the present studies, and the results for speech rate and vowel formants will be presented.

2 Method

To be able to compare speech in L1 and L2 contexts directly, the experiment was carried out by recording native speakers of Norwegian 1) in dialogue with L2 speakers, and 2) in dialogue with other L1 speakers. The dialogue setting was based on a keyword manuscript, to facilitate natural speech, and at the same time be able to compare phonetic parameters in identical words and phonological contexts.

2.1 Subjects

Six native speakers of Norwegian (with eastern Norwegian dialect background) participated as informants. Three of them were teachers in Norwegian as a second language, called P