GENERATING EMOTIONAL SPEECH WITH A CONCATENATIVE SYNTHESIZER

Erhard Rank *Hannes Pirker*

Austrian Research Institute for
Artificial Intelligence (ÖFAI)
Schottengasse 3, A–1010 Vienna
Email: erhard|hannes@ai.univie.ac.at

ABSTRACT

We describe the attempt to synthesize emotional speech with a concatenative speech synthesizer using a parameter space covering not only f0, duration and amplitude, but also voice quality parameters, spectral energy distribution, harmonics-to-noise ratio, and articulatory precision. The application of these extended parameter set offers the possibility to combine the high segmental quality of concatenative synthesis with a wider range of control settings needed for the synthesis of natural affected speech.

1 INTRODUCTION

The quality of synthesized speech is usually measured in terms of intelligibility and naturalness. State of the art synthesizers are well intelligible under regular conditions. Therefore, development has concentrated on the improvement of naturalness. An obvious way to follow is the incorporation of correct and functionally adequate prosody. Another maybe less obvious but very interesting aim is the generation of non-neutral affect.

For the English language a system for the reproduction of affected speech by a synthesizer is presented in [1]. Physiological influences that contribute to a source modeling of emotional speech were investigated as well as acoustic correlates that can be directly measured and used in the synthesizer. The implementation of an 'affect editor' coupled to the DECtalk speech synthesizer resulted in correct recognitions of six different emotions significantly above chance level.

There is an obvious influence of affect on the prosodic parameters f0, segmental duration, and signal energy. However, extracting these parameters from emotional speech of a human speaker and using them to control a concatenative synthesizer ('prosody transplantation') showed that these cues do not suffice for identification of the intended emotion by human listeners [4, 5].

More parameters than f0, segmental duration, and signal energy were identified and used for the 'affect editor' in [1]: e.g., breathiness, brilliance, or precision of articulation. Specific analysises of acoustic parameters other than the prosodic parameters in German speech with emotional content [2, 3] discovered influences on the glottal pulse shape, harmonics-to-noise ratio, and the mean spectral energy distribution and correlations with some kind of voicing irregularities.

In this study we want to test the hypothesis that incorporation of these additional parameters—in particular spectral energy distribution, harmonics-to-noise ratio (breathiness), voice quality parameters (brilliance, creaky voice), as well as articulatory precision—can contribute to a more definite representation of affect in synthesized speech. To that end we derive a parameter set for our demisyllable based concatenative synthesizer in order to synthesize speech with the four emotions *anger, sadness, fear,* and *disgust* based on the analyses in [1] and [2], present the results of a listening test, and conclude with considerations on the requirements for improved emotional speech synthesis.

2 ACOUSTIC CORRELATES FOR EMOTION

Since 'emotion' in speech is directly coupled to the speaker's mental state a production model (i.e., modeling of the 'speaker') would seem the appropriate way to generate emotional speech. The speakers behaviour for a certain emotional state would then directly result in the desired affect in the synthesized speech: for example, a higher rate of breathing pauses when the speaker is nervous or increased noise due to turbulent air flow when the teeth are pressed together in fear.

But as speech synthesizers usually do not process physical states of humans[1] but concatenate recorded segments according to a phonetic input specification (probably derived from a textual representation), the processing in order to achieve a certain affect has to take place in the regime of the acoustic segments and the synthesizer parameters.

As mentioned before, emotional content is connected with the prosodic parameters f0, segmental duration, and energy. For the

[1]Remarkable exceptions can be found in [6], chapter III, or (for a singers voice) in [7]; of course all Klatt-type synthesizers also use the physical state of the vocal tract as parameter.

parameter f0 a further dissection into at least two measures—namely 'mean f0' and 'f0 range'—is generally applied, but also terms like amount of 'final lowering' or 'accent shape' are used [1]. A valuable cue for the characterization of anxious speech is the amount 'f0 jitter' which describes the variation of f0 from one pitch period to another.

Segmental duration has to be divided into duration of phonemes and pause duration. Pausing, and particularly the presence or absence of hesitation pauses are of great importance for the characterization of affected speech.

Energy is measured as mean energy, or as energy at a reference point (e.g., inside a certain vowel). A generalization of the energy measure is used for the analysis in [2]: the *spectral energy distribution* for different emotions is classified through energy values in four different frequency bands. It is also stated that a shift of the mean energy towards higher frequencies for some affects may be due to a higher amplitude of fricatives and additional friction noise. This additional friction noise can be classified by the *harmonics-to-noise ratio* for voiced phonemes.

The harmonics-to-noise ratio is one of the *voice quality parameters*, also subsuming, e.g., glottal pulse shape, or creaky voice. Glottal pulse shape determines the spectrum of the glottal excitation signal: a shorter pulse yields more energy in the high frequency range than a longer pulse. The influence of affect on the shape of the glottal pulse has been investigated in [2]. Creaky voice is due to 'missing' glottal pulses.

Another correlate for affect is *articulatory precision*: this term describes the changes in quality of vowels—e.g., whether an /E/ is reduced to a schwa /@/—and the reduction of unvoiced consonants to their voiced counterparts [1]. Generally, articulatory precision is proportional to speaker arousal, articulatory precision of vowels is related to the vocal tract properties.

3 SYNTHESIZER PARAMETERS

The synthesizer used is a concatenative synthesizer based on a demisyllable inventory recorded by a male speaker of standard Austrian German. This synthesizer was designed for the use in a concept-to-speech system [8] and uses a combination of phoneme and tone specifications as input. Phonemes are stated in SAMPA (speech assessment methods phonetic alphabet) notation, tones according to the G-ToBI (German Tone and Break Indices) notation.

The demisyllable inventory is stored in the form of LPC (linear predictive coding) coefficients for a lattice filter and residual, the LPC analysis is performed pitch synchroneous. Prosody modification is achieved by the use of simple residual excited linear prediction (SRELP) synthesis [9], which allows to overlay a pitch contour for voiced regions at will and duration modification for each phoneme. The details of the prosody modification algorithms are laid out in [10].

The concept of tone specifications determining the f0 contour allows a simple realization of the parameters f0 base and f0 range. f0 base shifts the reference line and f0 range determines the dynamic of the f0 excursions for a specified tone. The parameter f0 jitter is realized by a random variation in the length of the pitch periods with an amplitude according to the parameters value. This random variation is controled by a white noise signal filtered by an one pole lowpass filter. The tone specifications itself are identical for all synthesized emotions, so the same qualitative shape of the f0 contour is used for all four emotions for one sentence.

To modify segmental duration of phonemes in affected speech a length correction factor can be specified independently for vowels, voiced consonants and unvoiced consonants. Another length correction factor can be specified for pause lengths, in combination with an interval for random variation of the pause length.

Energy can be varied in two ways. First, an amplitude correction factor can be specified for vowels, voiced consonants and unvoiced consonants. So, to some extend the desired mean spectral energy distribution can be achieved by tuning the amplitude proportions of these phoneme groups. Second, the speech signal is filtered by a custom filter for each of the emotions under consideration. These filters were designed after the results in [2], table 2.

To vary the harmonics-to-noise ratio additional bandlimited noise (1–5 kHz) is added to the speech signal. Creaky voice is simulated by using sporadic residual pulses with with very low amplitude.

To incorporate articulatory precision the transitions between vowels and schwas, resp. between voiced and unvoiced consonants can be achieved by specifying a suitable phoneme as synthesizer input instead of the original one. Also an automatic transition by the synthesizer like described in [1]—probably with specified transition probability—is thinkable (but not implemented). But for vowels the hard transition by choosing another inventory phoneme can be avoided thanks to the inventory format as LPC coefficients and residual. The LPC coefficients in our synthesizer are reflection (parcor) coefficients for a lattice LPC filter. They can be easily transformed to LAR (log area ratio) coefficients (c.f. [10]). In the LPC source-filter model, LARs are the logarithmic ratios of cross-section areas at a fixed spacing along the vocal tract. If the LAR coefficients are multiplied by a factor $f > 1$ the cross-section of the model vocal tract gets larger where it is large and smaller where it is small. Otherwise, if the LAR coefficients are multiplied by a factor $f < 1$ the vocal tract tends towards the shape of a uniform tube. In a limited range this multiplication of LAR coefficients can be used to change the quality of vowels towards the schwa ($f < 1$) or to a more enunciated version of the very vowel.

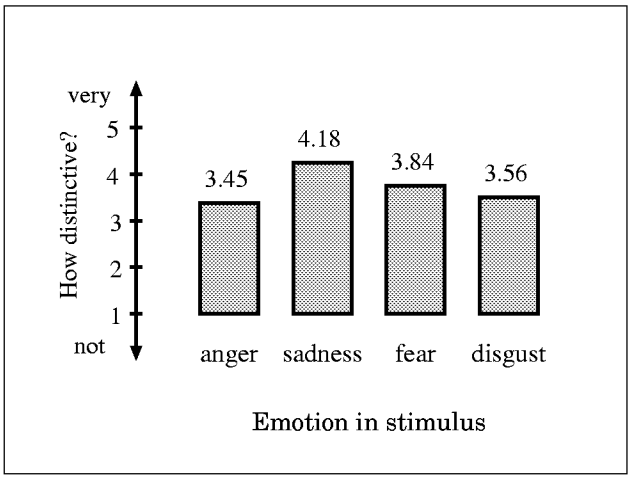

Figure 1: 'How well can the presented emotions be distinguished?' Mean value of the rated distinction for each synthetic emotional stimulus (compared to stimuli of each of the other emotions).

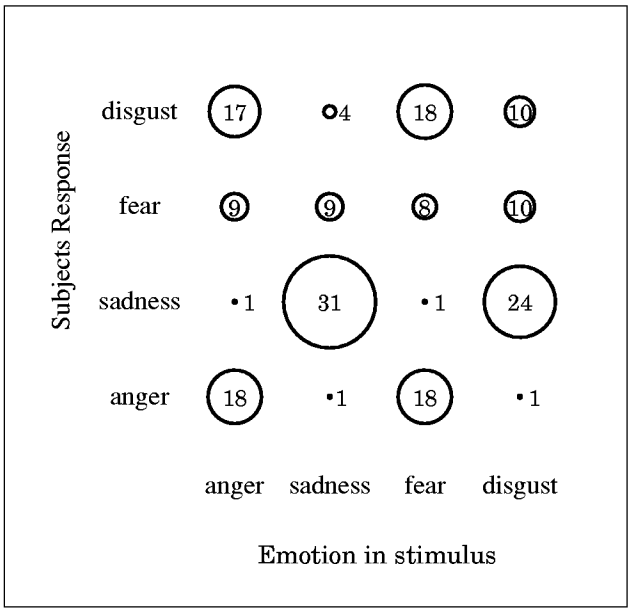

Figure 2: Number of classifications as a particular perceived emotion depending on the stimulus emotion.

4 PERCEPTUAL EVALUATION

A perception test was performed in order to judge the recognizability of the desired affect. After listening to a series of examples of affected synthetic speech the subjects had to rate how good two different synthetic emotions could be distinguished on a scale from 1 to 5 in the first part of the test. In the second part of the test they had to classify a stimulus as one of the four emotions *anger, sadness, fear,* and *disgust*. Moreover, the subjects were to rate (again on a scale from 1 to 5) how appropriate the stimulus was for the particular emotion.

Five different sentences with either of the four emotions were used as stimuli. In the first part two differently affected versions of one sentence (30 stimuli) and in the second part one affected sentence (20 stimuli) were presented. The test was carried out automatically on a workstation, stimuli were presented over headphones, and the test program presented the stimuli in random order. The subjects were able to repeat the stimuli at will before rating them.

In figure 1 the mean value of the rating how good two different synthetic emotions could be distinguished is shown. The mean is calculated over all possible pairs of one emotion and one of the other emotions. A high value ('very distinctive') is reached for sadness and also the other ratings are above the average rating of 3. This suggests that the parameter space is of sufficient dimensionality to describe four different emotions.

However, the results of the second part of the test show that either the parameter setting or the effect modeling are not satisfactory: only sadness was recognized clearly, anger was correctly classified in the majority, but synthetic fear and disgust were rated as one of the other emotions more often than as for the intended emotion. A plot with the number of perceived emotion depending on the stimulus emotion is shown in figure 2.

The rating of how appropriate a stimulus was perceived is depicted in figure 3. The shaded boxes show the ratings for stimuli classified as the intended emotion whereas the white boxes show the ratings for false classified stimuli.

5 CONCLUSION AND OUTLOOK

In this study we tried to apply the parameters of an affect generator for the English language [1] and the analysis of German emotional speech [2] to our concatenative synthesizer for standard Austrian German. We wanted to exploit the possibilities of the LPC model used in the synthesizer and use more than the prosodic parameters f0, segmental duration, and signal energy to achieve improved emotional speech synthesis. To that means the structure of the synthesizer was modified, e.g., to include an additional noise source or to allow for modifications of the LAR coefficients to change articulatory precision.

The results of an perception test show that the choice of parameters to simulate emotions was not ideal. However, the increased parameter space seems to suffice to represent four or more different emotions. A shortcome of the test setting may be the use of identical tone specifications, and thus qualitatively the same f0 contour for all emotions[2].

[2]This was also criticized by the test subjects.

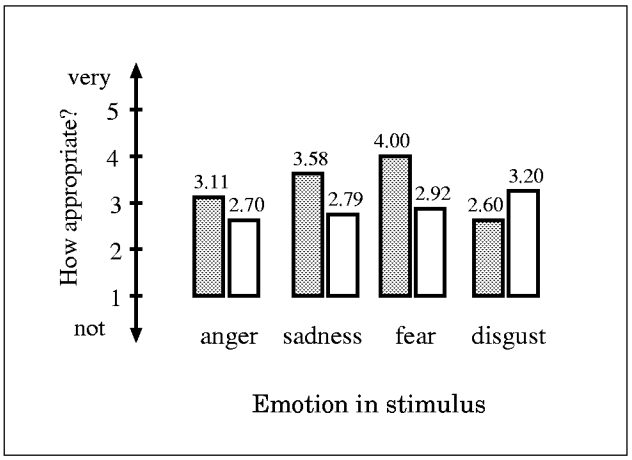

Figure 3: 'How appropriate is the stimulus for the perceived emotion?' Mean ratings for correctly classified (shaded boxes) and for false classified (white boxes) stimuli.

Since this study has been performed without an analysis step of its own, part of the results may be due to the insufficient coverage of the acoustical correlates for emotion in the Austrian variant of German by the analysises in the referenced sources. For further development a language specific analysis is requisite.

An intention that could not be realized yet was the direct modification of the glottal pulse. Since the LPC residual does not resemble the shape of the glottal pulse when a minimum energy error criterion is used in the analysis, manipulations of the opening and closing phase and the closed glottis interval (as proposed in [2]) are not possible with the current version of our synthesizer. LPC analyses methods suited better for speech signals could be a remedy [11].

Acknowledgement

Financial support for ÖFAI and for this specific project is provided by the Austrian Federal Ministry of Science and Transport.

REFERENCES

1. J. Cahn, "Generating Expression in Synthesized Speech," Master's thesis, MIT, 1989. Published as Technical Report, MIT Media Laboratory, 1990.

2. G. Klasmeyer and W. F. Sendlmeier, "Objective Voice Parameters to Characterize the Emotional Content in Speech," in *Proceedings of ICPhS'95*, (Stockholm, Sweden), 1995.

3. G. Klasmeyer, "The Perceptual Importance of Selected Voice Quality Parameters," in *Proceedings of ICASSP'97*, (Munich, Germany), 1997.

4. B. Heuft, T. Portele, and M. Rauth, "Emotions in Time Domain Synthesis," in *Proceedings of ICSLP'96*, (Philadelphia, PA), 1997.

5. M. Edington, "Investigating the Limitations of Concatenative Synthesis," in *Proceedings of Eurospeech'97*, (Rhodes, Greece), 1997.

6. J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, eds., *Progress in Speech Synthesis*. New York: Springer, 1997.

7. P. R. Cook, "SPASM, a Real-Time Vocal Tract Physical Model Controller; and Singer, the Companion Software Synthesis System," in *Computer Music Journal*, vol. 17, no. 1, 1993.

8. K. Alter, E. Buchberger, J. Matiasek, G. Niklfeld, and H. Trost, "VIECTOS – The Vienna Concept to Speech System," in *Natural Language Processing and Speech Technology - Results of the 3rd KONVENS Conference*, (Bielefeld), pp. 166–170, October 1996.

9. M. Macchi, M. J. Altom, D. Kahn, S. Singhal, and M. Spiegel, "Intelligibility as a Function of Speech Coding Method for Template-Based Speech Synthesis," in *Proceedings of Eurospeech'93*, (Berlin, Germany), pp. 893–896, 1993.

10. E. Rank and H. Pirker, "VIECTOS—Speech synthesizer, Technical Overview," tech. rep. TR-98-13, Austrian Research Institute for Artificial Intelligence, Vienna, 1998.

11. R. Ansari, D. Kahn, and M. J. Macchi, "Pitch Modification of Speech Using a Low-Sensitivity Inverse Filter Approach," in *IEEE Signal Processing Letters*, vol. 5, no. 3, 1998.