BMBF

Verbmobil
Verbundvorhaben

Synthesizing Prosody: a prominence-based approach

Barbara Heuft
Thomas Portele

IKP Universität Bonn

√m Report 176
September 1996

September 1996

Barbara Heuft
Thomas Portele

Institut für Kommunikationsforschung und Phonetik
Universität Bonn
Poppelsdorfer Allee 47
53115 Bonn

Tel.: (0228) 7356 - 80
Fax: (0228) 7356 - 39
e-mail: {tpo}@ikp.uni-bonn.de

Gehört zum Antragsabschnitt:
4.3

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 D 08 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

EMOTIONS IN TIME DOMAIN SYNTHESIS

Barbara Heuft, Thomas Portele**, Monika Rauth***

*now: Lernout & Hauspie Speech Products, Ieper, Belgium, email: barbara.heuft@lhs.be
**Institut für Kommunikationsforschung und Phonetik, Universität Bonn, Germany

ABSTRACT

A preliminary test exploring 4 emotions showed that conveying emotions by time domain synthesis may be possible. Therefore, a more sophisticated test was carried out in order to determine the influence of the prosodic parameters in the perception of a speaker's emotional state. Six different emotional states were investigated. The stimuli of the second test were used in three different testing procedures: as natural speech, resynthesized and reduced to a sawtooth signal. The recognition rates were lower than in the preliminary test, although the differences between the recognition rates of natural and synthetic speech were comparable for both tests. The outcome of the sawtooth test showed that the amount of information about a speaker's emotional state transported by F_0, energy and overall duration is rather small. However, we could determine relations between the acoustic prosodic parameters and the emotional content of speech.

1. MOTIVATION

This study explores the possibility of simulating emotions in time domain speech synthesis. In earlier studies dealing with the acoustic-phonetic correlates of emotions (see e.g. Klasmeyer, 1995), voice quality-phenomena such as jitter or different modes of exitation have been found to be important factors .

These phenomena cannot easily be controlled in time domain speech synthesis. However, it would be useful to be able to simulate emotions in order to make the synthesis sound more lively.

The factors that can easily be manipulated in time domain speech synthesis are the prosodic parameters duration, fundamental frequency and energy. So the question about emotions in time domain synthesis can be reformulated as follows: How much information about the speaker's emotional state is conveyed by these three prosodic parameters?

1. PRELIMINARY EXPERIMENT

1.1 Natural Speech

In a preliminary experiment, three emotionally neutral German sentences were chosen. The sentences were <Am Wochenende soll es Schnee geben> (*There will be snow this weekend*); <nein> (*no*) and <Morgen wird alles anders> (*Tomorrow everthing will be different*). They were uttered by three speakers in a neutral style, and simulating three different emotions: Joy, fear and anger. The recordings were done with a movable microphone held by the speaker in order to allow the subjects to gesticulate. The 36 stimuli were played to 8 subjects. They recognized the intended emotions in 82% of cases (chance level:25%; Chi square test: for all subjects $p<0.05$). Angry and neutral speech were recognized most reliably (see **Figure 1**). The speaker and the sentence with the lowest identification rates were excluded for the following experiment.

1.2 Resynthesis

The 16 remaining utterances were resynthesized by a time domain synthesis system (Portele et al., 1994) with the same prosodic features as the original utterances, using two different unit inventories for one male and one female. Durations and energy values were measured by hand; the pitch was determined automatically. (One difficulty were numerous overmodulations caused by the recording conditions; the pitch marks could not be set correctly so that a transfer of pitch contours was not always possible with the desired quality).

The stimuli were played to 9 subjects. As expected, the classification was worse than for the natural speech: 55% correct (chance level: 25%; Chi square test: for 8 subjects $p< 0.05$). The emotions most often classified correctly were fear and neutral speech (see **Figure 1**).

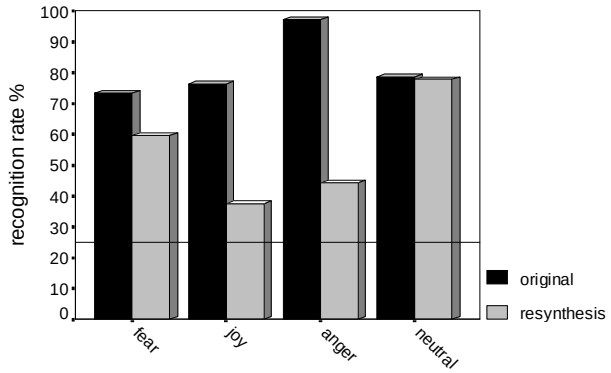

Figure 1: Recognition rates for the pilot test. The horizontal line indicates the chance level at 25%.

1.3 Preliminary Conclusions

Our hypothesis regarding the low recognition rates was that the poor recording conditions had influenced the pitch transfer and thus hindered recognition. This was supported by the fact that most subjects had difficulties recognizing joy, because this emotion was marked by a enhanced pitch range. Still, the results suggested that it should be possible to convey emotions in time domain synthesis without difficulty.

2. FOLLOW-UP TEST

In the second test, the target sentences were embedded in short texts with emotional content in order to make it more easy for the speakers to simulate natural sounding emotions. It was not explained to the speakers which emotion they should express but they were asked to

read the paragraphs in an appropriate style.

The sentences were <nein> (*no*), <Um Gottes Willen> (*For God's sake*) and <Ich verstehe das nicht> (*I don't understand it*). The English equivalents of the last two sentences have succesfully been used by Williams & Stevens (1972). The recordings were done in an anechoic room. Headsets were used in order to allow the speakers to gesticulate *and* keep the microphone distance constant.

This time, six different emotional states were investigated: Fear, joy, anger, neutral, disgust and sadness. Five speakers were recorded. From these recordings, three speakers (2m 1f) were chosen by the authors in an informal evaluation.

2.1 Natural Speech

Again, the natural utterances were presented to 9 subjects in order to find out the utterances in which the emotions could be recognized best. The stimuli were presented via headphones; each stimulus was played twice.

The recognition rates were lower than for the first experiment. One reason for this is of course, that the subject could choose between 6 possibilities whereas in the first experiment, only 4 possibilities had been given. Further, the speakers had been speaking without exaggerating too much, maybe because they were not explicitly told the aim of the recordings. This produced (at least to our impression) a very natural sounding of the emotional speech but on the other hand, it meant that the emotions were more difficult to recognize.

As **Figure 2** shows, anger and neutral speech as well as fear were regognized well, whereas the recognition rates for disgust and sadness lay only slightly above the chance level. Possibly, the range of emotions chosen was too wide for the experiment.

2.2 Resynthesis

The two sentences <Nein> and <Um Gottes Willen> had the highest recognition scores and were chosen for the following experiments. One male speaker was also excluded. The same procedures as for the first test were used to transfer duration and energy values. The pitch contours were parametrized (see Heuft et al., 1995) in order to avoid problems with pitch detection errors. The resulting synthetic speech was of a much better quality than in the first test.

Nevertheless, the results, shown in **Figure 2**, show very low recognition rates. This is, at least in part, due to the fact that the natural speech stimuli had already obtained bad scores.

Again, joy was obviously the most difficult emotion to recognize. Only fear and neutral speech could be identified reliably. If the results are normalized for the number of options, the differences between the recognition rates for natural and resynthesized speech are almost identical (see **Figure 3**). The normalization was done following the formula X = R - (F / m-1); with R: number of correct answers; F: number of false answers; m: number of options.

2.3 Sawtooth signals

Sawtooth signals of the stimuli were generated from the pitch marks of the natural speech stimuli. This way, only the prosodic information (i.e. F_0, energy and the overall duration) was left. Sonntag (1996) has shown that subjects are able to recognize prosodic structures such as accentuation and phrasing with high consistency from such stimuli.

These sawtooth stimuli were presented to 10 subjects in a similar procedure as in the previous experiments. This time, the recognition rates were even worse than for the resynthesized stimuli. Most subjects claimed that it was impossible to recognize any emotion. The higher regonition rate for neutral speech is due to the fact that most subjects chose the option "neutral" much more often than the other options (see **Figure 2**).

It can be seen from this experiment that the problem of resynthesizing emotions does not lie in the synthesis as such, but in the fact that emotions are not always prosodically marked, or at least not marked enough to be easily recognizable.

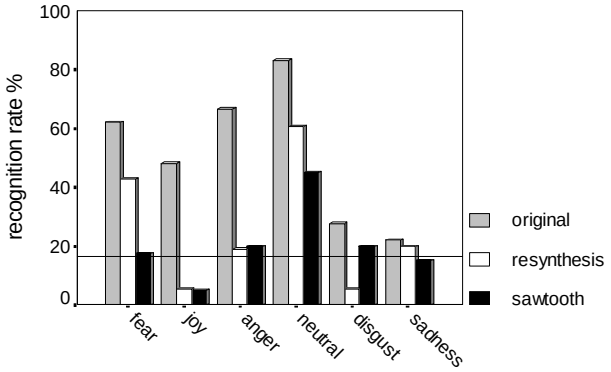

Figure 2: Recognition of the intended emotions in the second experiment for original speech, resynthesized speech and sawtooth signals. The chance level (16,6 %) is indicated by the horizontal line.

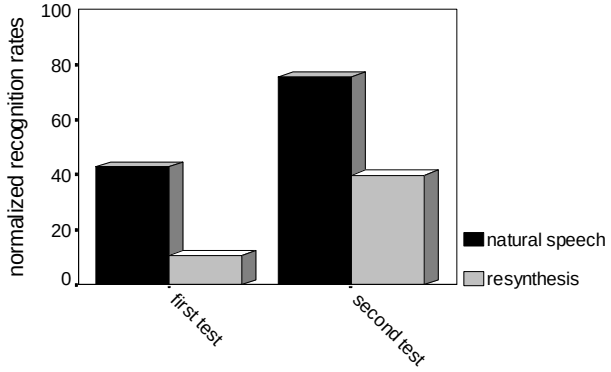

Figure 3: Recognition rates normalized for the number of options.

3. INFLUENCE OF THE PROSODIC PARAMETERS

Even if the results of the previously described experminents were not as promising as we had expected, we had a closer look at the prosodic parameters that characterize the different emotions. Of course, it only makes sense to analyse the prosodic features depending on the recognition scores of the stimuli. For the analysis, only the stimuli with a reconition significantly above the chance level were chosen. We analyzed mean F_0, F_0-range and overall duration. Because of the limited number of utterances, no significances can be given. Thus, everything that is said about the prosodic characterization of emotional speech should be understood as being no more than a tendency.

3.1 Fundamental frequency

Figure 4 shows the results for F_0 range. First, it becomes clear that the male speaker seemed to make use of this parameter much more than the female. He produced a big F_0 range for all emotions exept fear and neutral speech. Exept for sorrow, this agrees with earlier findings (e.g. Fonagy & Magdics, 1963; Fairbanks & Prosnovost, 1939). The female speaker showed a quite narrow range for all emotions, the biggest values were found for anger and neutral speech. For both speakers, we found a rather high mean fundamental frequency for fear and anger and lower values for all other emotions. However, these differences were not very marked.

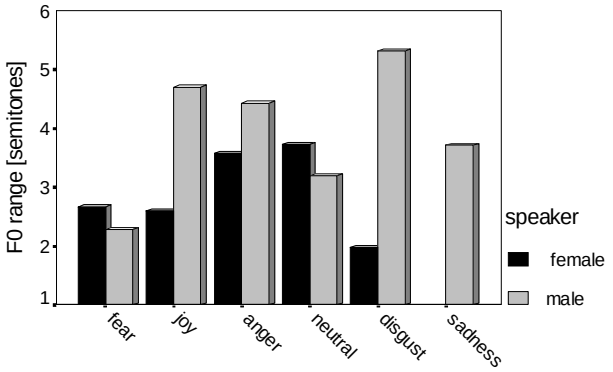

Figure 4: Mean F_0 range for the different emotions. Only stimuli with recognition rates above chance level were analyzed.

3.2 Duration

The overall duration of the sentences was measured to determine the speech tempo. The results are shown in **Figure 5**. Again, it is the male speaker who is more in line with the results from earlier studies (see Murray & Arnott, 1995, 1993 for an overview). The longest durations are found for anger and disgust, average durations for neutral speech and fear; short durations for joy and sadness. The short durations as well as the large F_0 range contradict other findings (e.g. Murray & Arnott, 1995; Williams & Stevens, 1972) for sadness.

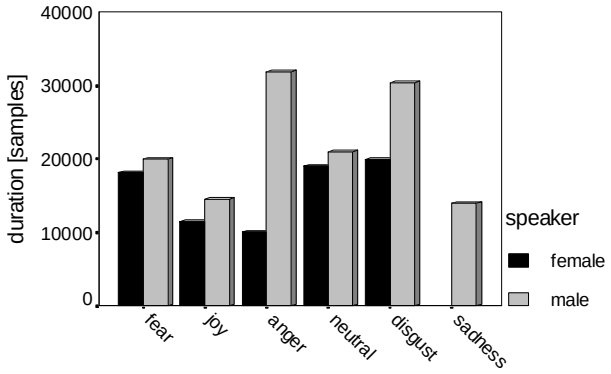

Figure 5: Mean overall duration of the sentences depending on the simulated emotion. Only stimuli with recognition rates above chance level were analyzed.

3.3 Listeners expectations

It makes sense to look at the dependencies between recognition rate and the realization of the prosodic parameters. In this way, we may determine which features the listeners probably would have *expected*. Therefore, we calculated for each intended emotion the correlation coefficients between the recognition rates and the parameters.

Table 1 gives a survey of the results. Fear was expected to be marked by a small F_0-range and a short duration. These results are consistent with the speakers' production and with results of other experiments (e.g. Fonagy & Magdics, 1963). This is probably the reason for the fact that fear was recognized relatively well in the experiment with resynthesized speech. For joy we can find short duration as the only acoustic parameter causing better recognition. Previous studies characterize joy as having a shorter overall duration (e.g.Murray & Arnott, 1995; Williams & Stevens, 1972), but still, joy was hardly recognized when only the prosodic information was left. All authors give a larger pitch range as an important acoustic correlate for joy. Maybe our speakers had used something other than the prosodic features to characterize this emotion. The recognition rates for anger had negative correlations with both mean F_0 and F_0 range. This is neither consistent with the speakers production nor with the findings of e.g. Carlson et al.(1992), where a high pitch and a wide pitch range was a clear sign of anger. There was no correlation between recognition of anger and duration. Disgust was expected to be marked by a larger pitch range. It was produced that way by the male speaker. Sadness seemed to be expected to have a low fundamental frequency (which is commonly assumed) and a short duration (which is *not* commonly assumed, but which is what we found in the speakers' actual realizations).

Of course, these results can be biased by the actual realizations: If e.g. an emotion was classified correctly using other cues than prosodic cues than the prosodic features, the results of the correlation analysis might be misleading.

As the male speaker seemed to use the parameters more often in the expected way, it is not surprising that his utterances were more often classified correctly than those of the female speaker (61% vs 43% for the natural speech stimuli.

	mean F_0	F_0 range	duration
fear	- 1	**-5**	**-9**
joy	1	-1	**-8**
anger	**-8**	**-7**	1
neutral	**-6**	**-6**	**5**
digust	-1	**6**	1
sadness	**-6**	-2	**-7**

Table 1: Correlations between the recognition rates and the realization of the prosodic parameters.

CONCLUSIONS

Although we did not find the clear results we had expected, we can draw some conclusions for the generation of emotional speech in time domain speech synthesis. If we want listeners to perceive emotions in synthetic speech, we obviously cannot simply copy the prosodic features of natural utterances, because in natural utterances, the prosodic features are supported by other features such as voice quality. Therefore, in a further study, one should employ a different strategy, i.e. systematically vary the prosodic features according to our and other results and see which combination of parameters will most clearly evoke impressions of the different emotions.

REFERENCES

1. Carlson, R.; Granström, B.; Nord, Lennart (1992): Experiments with emotive speech - acted utterances and synthesized replicas. *Proc. ICSLP'92*, pp. 671-674

2. Fónagy, I.; Magdics, K. (1963): Emotional patterns in intonation and music. *Z. Phonet. Sprachwiss. und Kommunikationsforsch*.**16**, pp. 293-326

3. Fairbanks, G.; Prosnovost (1939): An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monograph* **6**, 87-104

4. Heuft, B.; Portele, T.; Höfer, F.; Krämer, J.; Meyer, H.; Rauth, M.; Sonntag, G. (1995): Parametric description of F_0-contours in a prosodic database. *Proc. ICPhS'95*, Stockholm, pp.378-381

5. Klasmeyer, G. (1995): Objective voice parameters to characterize the emotional content in speech. *Proc. ICPhS'95*, Stockholm, pp. 182-185

6. Murray, I.R.; Arnott, J.L. (1995): Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication* **16**, pp. 359-368

7. Murray, I.R.; Arnott, J.L. (1993): Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J. acoust. Soc. Am*. **93**, pp. 1097-1108

8. Portele, T.; Höfer, F.; Hess, W. (1994): Structure and representation of a unit inventory for German speech synthesis. *Proc. ICSLP'94*, Yokohama:1759-1762

9. Sonntag, G. (1996): Klassifikation syntaktischer Strukturen aufgrund rein prosodischer Information. To appear: *Fortschritte der Akustik - DAGA'96*

10. Williams, C. E. ; Stevens, K.N (1972): Emotions and Speech: Some Acoustical correlates. *J. Acous. Soc. Am*. **52**, pp. 1238-1250