Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis

Felix Burkhardt and Walter F. Sendlmeier

Technical University of Berlin, Germany.

e-mail: felixbur@kgw.tu-berlin.de

ABSTRACT

This paper explores the perceptual relevance of acoustical correlates of emotional speech by means of speech synthesis. Besides, the research aims at the development of »emotion-rules« which enable an optimized speech synthesis system to generate emotional speech. Two investigations using this synthesizer are described: 1) the systematic variation of selected acoustical features to gain a preliminary impression regarding the importance of certain acoustical features for emotional expression, and 2) the specific manipulation of a stimulus spoken under emotionally neutral condition to investigate further the effect of certain features and the overall ability of the synthesizer to generate recognizable emotional expression. It is shown that this approach is indeed capable of generating emotional speech that is recognized almost as well as utterances realized by actors.

1. INTRODUCTION

Although emotional speech has been investigated for many years, there are still open questions regarding the primary acoustical correlates of certain emotional speaker states. Especially with regard to the development of high-quality text-to-speech systems a better understanding of the expression of emotional speech is desirable.

At the Institute for Communication Science of the Technical University of Berlin a speech database of emotional speech was recorded. This database comprises 10 sentences spoken by 10 actors (5 male and 5 female) who simulated 7 emotional states (neutral, anger, joy, fear, sadness, disgust and boredom) for each of the 10 sentences. The utterances were analyzed with respect to prosodic and segmental features [5, 12]. Some of the results from these analyses as well as results described in the literature (see e.g. [10, 1]) are investigated regarding the perceptual relevance by means of speech synthesis.

The purpose of this investigation is twofold:

- How important are certain features as carriers of information of emotional speaker states?
- How successful can modern TTS-systems simulate emotional speech?

To answer these questions, emotionally neutral sentences are copied by a parametric speech synthesizer. These stimuli are then presented in perception experiments. Mainly three parametric synthesis concepts have been developed in the past: articulatory, formant- and LPC-synthesis. For the problem at hand it is necessary to match phonetic features with acoustic characteristics. At the same time high speech quality is to be achieved. Formant-synthesis is used because its basic approach is closer to physical modeling than LPC-synthesis, yet it has a better synthesis quality than articulatory synthesis. As speech synthesis system the KLSYN88 synthesizer [7] was used because its source-code is freely available. Experiments introducing the simulation of emotional speech with formant-synthesis have already shown promising results (see e.g. [9, 11, 3, 2]), although some of these studies [11, 2] suffered from the limitations introduced by using a commercial synthesizer without full control of the formant parameters. Similar investigations using an LPC-synthesizer [13] have shown to be less successful, lacking the capabilities of a phonetically motivated synthesis model (especially for the voice source).

2. EMOSYN: A SPEECH SYNTHESIZER OPTIMIZED FOR EMOTIONAL EXPRESSION

In order to obtain a prototype of a speech synthesizer optimized for the generation of emotional speech and to overcome the limitations of using a commercial product without access to the source-code, a new synthesizer was developed and named »emoSyn« (acronym for emotion-synthesizer).

2.1. Overview

The input-format of the system is a list of phonemes with assigned prosody descriptors similar to the MBROLA-format [4], but extended by syllable and stress markers. A program to generate this extension automatically is provided by the system. The output are the parameters for the KLSYN88 synthesizer. For the experiments, the implementation of Sensymetrics Corp. was used.

The data-concept is hierarchical: An utterance is a set of syllables containing a set of phonemes. Apart from its prosodic characteristics, certain features can be assigned to each phoneme describing voice quality or articulatory settings. The emotional expression is generated by the application of adjustment rules. These rules describe enhancements or reductions concerning duration, F_0-contours, intensity, voice quality, articulatory features and vowel precision. They can be applied to specific phoneme categories, syllable-types or the whole utterance. There are three syllable-types: phrase-stressed, word-stressed and unstressed. Further information about emoSyn as well as sound-examples can be found on the internet [14].

2.2. Description of Modifiable Parameters

The following describes the modifiable features of emoSyn that were used in the perception experiments. In general, these features are parameterized by a rate given in percent.

mean pitch: Mean pitch-height can be modified directly by shifting all pitch-values by a specified amount.

pitch range: Pitch range is modified by expanding or compressing all pitch values around a reference value which is determined by the mean pitch-height of the last syllable.

pitch variation: This is the application of the pitch range algorithm on syllables. The reference value in this case is the syllable's mean pitch.

pitch contour (phrase): The pitch contour of the whole phrase can be designed as a rising, falling or straight contour. The contours are parameterized by a gradient (in semitones/sec). As a special variation for happy speech, the »wave model« can be used where the main-stressed syllables are raised and the in-between syllables are lowered. It's parameterized by the max. amount of raising and lowering and connected with a smoothing of the pitch contour.

pitch contour (syllables): A rising, falling or level contour can be assigned to each syllable-type. Additionally, the last syllable can be handled separately.

F_0-flutter: As a feature similar to jitter, Klatt suggested the parameter F_0-flutter [7]. It is applied to all vowels.

intensity (syllables): The intensity for each syllable-type can be modified by an amount in dB.

speech rate: The speech rate can be modified for the whole phrase, sound categories or syllable-types separately by changing the duration of the phonemes. A change in speech rate has a stronger effect on stressed vowels than on unstressed ones and a stronger effect on vowels than on consonants.

phonation type: The phonation type can either be a modal, falsetto, breathy, creaky, or tense voice (see Laver's terminology [8]). The KLSYN-88-voice-source parameters as well as formant-bandwidths and spectral notches are modified accordingly. Creaky voice can be assigned either to all voiced phonemes of the utterance or only to the first half of vowels after an unvoiced/voiced transition. It is implemented in a more harsh version (with short open-glottis phase) or a somewhat breathy one, which is suitable for emotions with low arousal. Falsetto voice is implemented by a pitch-shift and by introducing an irregularity to the pitch.

vowel precision: The vowel precision is realized by formant-target undershoot or overshoot in specified syllable-types. The first two formants are shifted towards or away from the neutral position.

lip-spreading: This feature is implemented by raising the frequencies of the first two formants by a given rate.

3. EXPERIMENT 1: SYSTEMATIC VARIATION OF ACOUSTICAL FEATURES

To confirm results from earlier studies as well as the perceptual relevance of rarely investigated features (i.e. voice-quality or articulatory precision [5, 6, 1]), a listening experiment was set up comprising five features (mean pitch, pitch range, speech rate, phonation type and vowel precision).

3.1. Generating the Stimuli

The five features were varied in the following way:

- mean pitch (3 steps): original, 50 % lifted and 30 % lowered
- pitch range (3 steps): original, 50 % narrower and 50 % broader (be aware that a pitch variation by 50 % does not necessarily result in a 50 % larger range)
- speech rate (4 steps): original, 20 % faster, 20 % slower, stressed syllables 20 % slower with all other syllables 20 % faster (further called mixed model)
- phonation type (5 characteristics): modal voice, creaky voice, falsetto voice, breathy voice and tense voice
- vowel precision (3 characteristics): original, target overshoot (phrase-stressed syllables 80 %, word-stressed 30 %), target undershoot (unstressed syllables 50 %, stressed syllables 20 %).

The stimuli were generated by emoSyn, modifying an emotionally neutral version of the German sentence: »*An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.« (At the weekends I always drove home and visited Agnes)*. This utterance was copy-synthesized from a phrase of the above mentioned emotional speech database. It is understood to be semantically neutral in the sense that it can be convincingly expressed with a variety of emotions.

3.2. The Perception Experiment

As the systematic variation of all five features would have resulted in a set of more than 2000 stimuli, the feature set was split into three main groups and each combined with the others, resulting in three tests. The first group comprises intonatory features (mean pitch and pitch range). Phonation types are a group of their own, and the last group, segmental features, consisted of speech rate and vowel precision. Three tests resulted:

- Test 1: intonation features and phonation types (45 stimuli)
- Test 2: phonation types and segmental features (60 stimuli)
- Test 3: intonation and segmental features (108 stimuli)

All tests were performed at a computer terminal using headphones in a quiet surrounding. The stimuli were presented to each listener in a different random order. The first two training stimuli were not included in the analysis. Thirty native Germanspeaking listeners participated in each test (15 female and 15 male). The average age was 30 years. Some were expert listeners, but most were naïve.

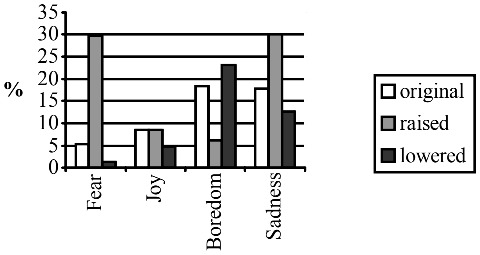


Figure 1: Average judgments for mean pitch modification.

The listeners were asked to assign one emotion to each stimulus in the set (original denotation in parenthesis): neutral, fear (*Angst*), anger (*Wut/Ärger*), joy (*Freude*), sadness (*Trauer*) or boredom (*Langeweile*). This set of emotions is the same as in the emotional database. Disgust was omitted, because it was not recognized well enough in the database.

3.3. Statistical Analysis and Interpretation of the Results

The results were analyzed by a series of univariate multifactorial ANOVAs with complete repetition of measurement using the statistical software SPSS. In the following interpretation, only results are discussed that yielded a significance level under 5 % (in fact most were lower than 1 %). The mean recognition rates are visualized in the figures 1-5. Only significant effects are regarded..

Fear: Utterances were judged as fearful when they had a high pitch, a broad range, falsetto voice and a fast speech rate or a speech rate according to the mixed model. Especially striking is the result of high pitch: the combination of raised pitch and falsetto voice, resulting in a pitch shift of 121 %, is perceived by 66 % of the judges as fearful. The significant effect of the mixed speech rate model is interesting, too; this effect could not be predicted from results in other studies.

Joy: The recognition of joy yielded the least obvious results in all three series of tests. Joy was also the emotion with the lowest recognition rates. It seems that important features (like intonation patterns) of joyful speech were not taken into account in this experiment. Nonetheless, a broader pitch range and a faster speech rate as well as modal or tense phonation are more often judged as joyful than the other characteristics. A lowered pitch sounds less joyful. Striking is the noticeable effect of vowel precision: A precise articulation enhances a joyful impression and an imprecise one reduces it.

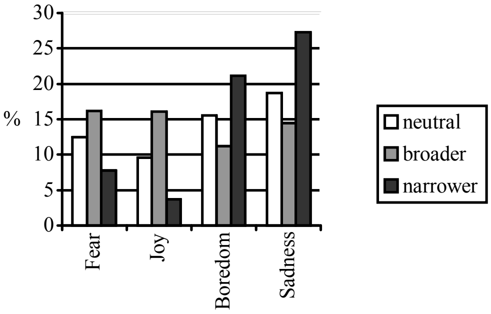


Figure 2: Average judgments for pitch range modification.

Boredom: A lowered mean pitch and a narrow pitch range as well as a breathy or creaky voice results significantly often in an assessment of the stimuli as bored. Furthermore, a slow speech rate and an imprecise vowel articulation enhances a bored expression. For all these features the reverse modifications weaken the assessment of the stimuli as bored.

Sadness: The expression of sadness is revealed by a narrow range and a slow speech rate as well as a breathy articulation. Surprisingly, a raised pitch contour and falsetto voice also enhance a sad impression. This is not consistent with findings in earlier literature, where sadness is described as an emotion with low arousal. We argue that the denotation »sadness« (or at least the German translation *Trauer*) is not specific enough and should be split into two categories: »crying despair« with high arousal and »quiet sorrow« with low arousal (in analogy to the often mentioned (e.g. [1]) differentiation between hot and cold anger).

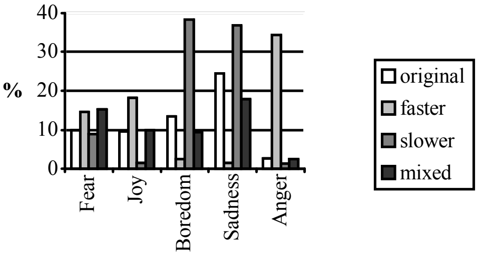


Figure 3: Average judgments for speech rate modification

Anger: For anger there are few, but obvious results. A faster speech rate and tense phonation is judged by the majority as angry. The combination of these features reached a recognition rate of 64 %. In combination with a fast speech rate, a raised mean pitch leads to a lower identification rate of anger. It seems that this combination of features tend to express cold anger, a category not explicitly defined in this experiment.

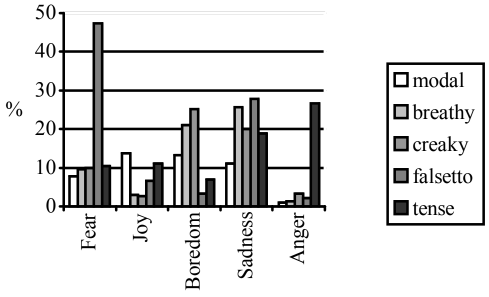


Figure 4: Average judgments for phonation type modification

3.4. Summary

The results strengthen earlier findings regarding the significance of the voice quality on the assessment of emotional speech. As the parameterization of natural glottis signals is extremely tedious, the use of analysis-by-synthesis methods are helpful to gain information about the role of phonation types with respect to emotional expression. Where the results are ambiguous, further optimization seems possible and worth while.

4. EXPERIMENT 2: FURTHER OPTIMIZATION OF SPECIFIC FEATURES

The restricted feature set and the limited number of emotions of the first experiment did not allow for a more specific exploration of the acoustical correlates. Also the ability of the synthesis system to generate emotional output could not be verified for all emotions. Therefore, a second experiment was run. To allow for the exploration of a greater feature set and at the same time to distinguish between emotions that differ mainly in the extent of arousal, a set of specific prototypes was designed. This approach differs fundamentally from the first one: While the stimuli were systematically varied and then classified by the judges in experiment one, now for each emotion a prototype is generated and varied slightly.

4.1. The Prototypes and their Variations

The three basic emotions anger, joy and sadness are split into pairs, each differing by the extent of arousal. This leads to a set of 8 emotions: hot anger, cold anger, joy, happiness, crying despair, quiet sorrow, fear and boredom. Again stimuli were generated by manipulating the neutral sentence used in the first experiment. The modifications are now discussed in detail. The rate of the applied changes is noted in parenthesis.

Hot/Cold anger (5 versions each) Both emotions are characterized by a tense voice (50 %) and faster speech rate (30 %) as compared to the neutral version. For hot anger the pitch was raised (50 %), and for cold anger it was lowered (20 %). For both of these prototypes, versions were generated by changing the following features: the pitch range was broadened (200 % for hot anger, 100 % for cold anger), the stressed syllables were given an extra pitch shift upwards (50 %, only hot anger), all stressed syllables were associated with a descending pitch contour (30 ST/sec), the articulatory precision was changed (30 % overshoot for stressed syllables and 20 % undershoot for unstressed ones) and all stressed syllables were intensified (9 dB).

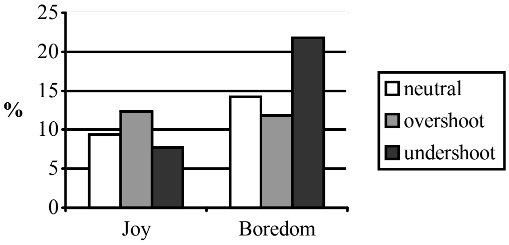


Figure 5: Average judgments for vowel precision modification

Joy/Happiness (5 versions each): The fundamental prototype of joy is very similar to hot anger: It is characterized by a faster speech rate (30 %), a raised pitch (50 %) and a broader pitch range (100 %). For happiness the speech rate is slower (20 %) and the pitch is not raised. For both emotions, versions were generated that differ with respect to lip-spreading (10 %), stressed syllables with rising pitch contours (50 ST/sec), formant target overshoot (30 %) and the wave pitch contour model (100 % max. raise, 20 % max. lowering). For happiness additionally a version with breathy phonation was generated.

Crying Despair (5 versions): The basic version is modeled with a slower speech rate (20 %), a raised pitch (100 %), a narrowed pitch range (20 %) and narrowed variability (20 %). Further modifications comprise descending inflections (20 ST/sec), F_0-flutter (FL=200), breathy and falsetto phonation (each 50 %).

Quiet Sorrow (5 versions): The fundamental version is characterized by an even slower speech rate (40 %), a lowered pitch (20 %), narrower pitch range and narrower variability (both 20 %). The versions differ with respect to descending pitch contours on stressed syllables (30 ST/sec), F_0-flutter (FL=300), and breathy voice (50 %).

Fear (4 versions): The previous experiment has shown that a fearful expression can be reached by lifting the pitch contour by more than 200 %. As the result sounds somewhat unnatural, now more subtle modifications are made. The prototype is characterized, like hot anger and joy, by a raised pitch (150 %, in combination with falsetto voice 50 %), a faster speech rate (30 %) and a broadened range (20 %). The versions differ by the application of straight pitch contours to stressed syllables, a rising pitch contour on the last syllable (30 ST/sec), F_0-flutter (FL=300) and falsetto voice (100 %).

Boredom (5 versions): The basic version for boredom is very similar to sorrow and has a slower speech rate (20 %), an additional lengthening of stressed syllables (40 %), a lowered pitch (20 %), a reduced pitch range (50 %) and reduced variability (20 %). The last modifications provide an almost flat pitch contour. Variations were made regarding formant target undershoot (20 % stressed syllables, 50 % unstressed ones), breathy phonation, and the application of creakiness to the voice onsets.

4.2 The Perception Experiment

The stimuli were presented to 42 listeners (23 female, 19 male) using the same procedure as in the first experiment. The larger number of judges was motivated by the observation that the variance of the judgments was extremely high. The listeners were asked to assign the stimuli to either neutral, hot anger (*Wut, Zorn*), cold anger (*Ärger, Gernervt*), happiness (*Wohlbefinden, Zufriedenheit*), joy (*Freude*), crying despair (*weinerliche Trauer*), quiet sorrow (*stille Trauer*), fear (*Angst*) or boredom (*Langeweile*). In order to provide the judges with a reference stimulus and to test the fundamental quality of the speech synthesizer, the emotionally neutral utterance was presented four times randomly during the test. Four training stimuli were played in advance, resulting in a number of 47 stimuli.

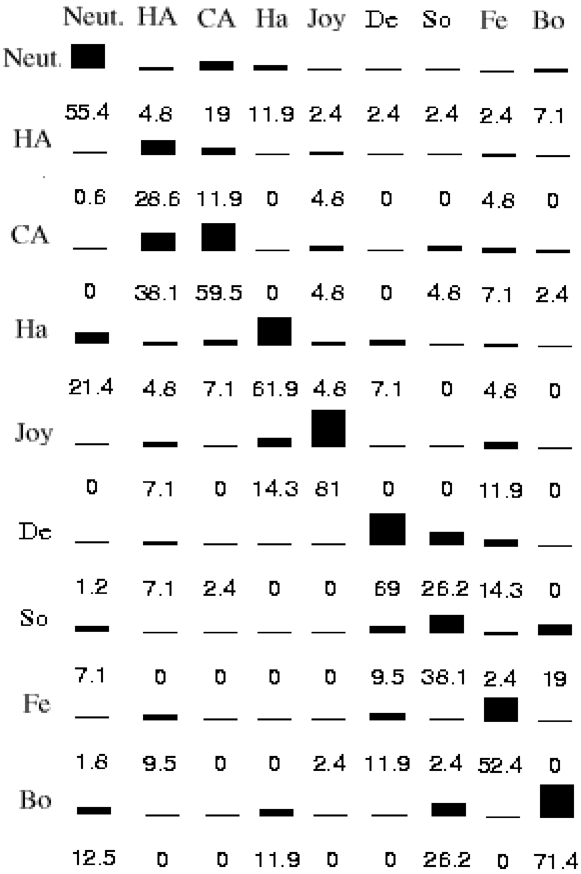
4.3 Analysis and Discussion of the Results

To analyze the results, again for each emotion an ANOVA was computed by SPSS. Only significant results (on the 5 % level) will be reported. Additionally, the results from confusion matrices were taken into account (see Figure 6). Only confusions above chance-level (11 %) are discussed.

Hot anger: The different versions were significantly identified as hot anger as opposed to the neutral stimulus. The most successful version yields a recognition rate of 28.6 %, but is confused with cold anger by 38.1 % of the judges. It seems that an angry quality is achieved primarily on the basis of a faster speech rate and tense phonation (as predicted by the first experiment), but a high arousal could not be simulated by the raised pitch. Perhaps the limited bandwidth of the synthesizer (5 kHz max freq.) did not allow to induce the harsh sound of hot anger. Another explanation may be that the system is unable to stress syllables that were not stressed before. Hot anger is reportedly known as an emotion with many stressed syllables (e.g. [5]). Three of the versions were also confused with fear, this can be explained by the fact that hot anger and fear are similar with respect to high arousal and negative valence.

Cold anger: All versions of cold anger were highly significantly identified as the intended emotion. The version with formant target overshoot achieves the highest recognition rate (59.5 %). If confused, cold anger was primarily interpreted as neutral or hot anger.

Joy: The results for joy reveal a large effect for the intonation wave model. The version introducing this model is recognized by 81 %. It remains unclear whether this is primarily achieved by the smoothing of the pitch contour or the raising and lowering of the syllables. Another result is the effect of the lip-spreading feature: The stimulus without this feature is not recognized. All versions except the one including the wave model are poorly recognized and frequently confused with anger and less often with fear or despair. This is explainable by a similarity in intensity (see also [1]).



	Neut.	HA	CA	Ha	Joy	De	So	Fe	Bo
Neut.	55.4	4.8	19	11.9	2.4	2.4	2.4	2.4	7.1
HA	0.6	28.6	11.9	0	4.8	0	0	4.8	0
CA	0	38.1	59.5	0	4.8	0	4.8	7.1	2.4
Ha	21.4	4.8	7.1	61.9	4.8	7.1	0	4.8	0
Joy	0	7.1	0	14.3	81	0	0	11.9	0
De	1.2	7.1	2.4	0	0	69	26.2	14.3	0
So	7.1	0	0	0	0	9.5	38.1	2.4	19
Fe	1.8	9.5	0	0	2.4	11.9	2.4	52.4	0
Bo	12.5	0	0	11.9	0	0	26.2	0	71.4

Figure 6: Confusion plot of the best recognized stimuli**,** rows=judged, columns=intended.(HA=hot anger, CA=cold anger, Ha=happiness, De=crying despair, So=quiet sorrow, Fe=fear, Bo=boredom)

Happiness: The favorable effect of the wave model is also valid for the stimuli with intended happiness. The best version yields a recognition rate of 61.9 %. If confused, it was categorized as joy (14.3 %) or neutral (11.9 %). As with joy, only stimuli with lip-spreading characteristic are recognized. All versions except the one with the wave model are furthermore confused with sorrow (14 %) and boredom (16 %), accounting for a similarity in intensity. Only the ones with lip-spreading characteristic and rising pitch patterns are confused with joy.

Crying despair: All versions achieve high recognition rates. The ones with F_0-flutter or falsetto phonation are recognized best (69 %). As both are characterized by a high pitch and F_0 irregularities, they sound very similar. The descending pitch contours do not enhance a sad expression. The best version of despair is confused with fear only slightly above chance level (11.9 %). This confusion is explainable by a similarity in arousal, valence and potency. Interestingly, the version with breathy phonation is confused with quiet sorrow by 28.6 % of the judges.

Quiet sorrow: All versions are recognized, but often confused with boredom, as boredom is often confused with quiet sorrow. The version without modifications concerning voice quality features is recognized by 62 % of the judges as bored. Only the versions including F_0-flutter are confused with despair. Obviously pitch irregularity is a feature that is relevant for discriminating between sadness and boredom, whereas descending pitch patterns seem not to be relevant.

Fear: The fearful stimuli are also recognized without exception. With respect to a difference between them there is only a tendency that the one with falsetto voice is recognized better (52.4 %) than the version with straight stressed syllable contours and a final pitch rise, which is often confused with despair (31 %). The others are also confused with despair (average 22 %), which is explainable as noted above.

Boredom: Like most of the other emotions, the bored stimuli are recognized well above chance level. The version with vowel target undershoot, but without modifications regarding the phonation is recognized best (71.4 %) and in fact better than the ones that include voice-source modifications. These are as often confused with quiet sorrow as recognized (average 40.1 %). It seems that the feature phonation type is not suited to discriminate between bored and sad speech. Quiet sorrow is the only emotion with which boredom is confused.

5. CONCLUSIONS

As the purpose of this investigation was twofold, the results will be discussed under two aspects:

- Most of the results from earlier work could be confirmed. The relevance of previously less regarded features like vowel precision, phonation types or intonation patterns could be shown. The high complexity of the parameters led to an optimization in the intended recognition of the simulated emotions. In order to optimize further, some limitations of the system have to be overcome (e.g. the fact that the number of stressed syllables could not be increased) and the subtle interaction of some features have to be analyzed in greater detail.
- The developed synthesis system has shown to be capable of generating recognizable emotional expression. This is even true for emotions that differ only gradually. If the emotions which have been split into two categories are put together again, all of them reach recognition rates that are comparable to those achieved for natural speech samples. Although the used synthesis system did not show to be biased for a specific emotional expression, the speech quality could be improved by exceeding the 5 kHz cut-off-frequency.

It should be noted that some easily confused emotion-pairs like crying despair/fear or quiet sorrow/boredom were not clearly distinguished from each other. We feel that the attention of further synthesis experiments should account more for the differentiation between emotions similar in various aspects as opposed to the distinction between the usual four basic emotions.

6. ACKNOWLEDGMENTS

This research was funded by the Deutsche Forschungs Gemeinschaft, grant No. Se 462/3-1. We further thank all the patient participants of the perception experiments.

7. REFERENCES

1. Banse, R. and Scherer, K. R., Acoustic Profiles in Vocal Emotion Expression, *Journal of Personality and Social Psychology*, 70(3):614-636, 1996
2. Cahn, J. The Affect Editor, *Journal of the American Voice I/O Society*, 8:1-19, 1990
3. Carlson, R. Granström, G. and Nord, L. Experiments with Emotive Speech, Acted Utterances and Synthesized Replicas, *Speech Comm.*, 2:347-355, 1992
4. Dutoit, T., Pagel, V., Pierret, N., Bataille, F. and Van der Vreken, O. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes, *Proc ICSLP*, 3:1393-1396, Philadelphia, 1996
5. Kienast, M. and Paeschke, A. and Sendlmeier, W. F. Articulatory Reduction in Emotional Speech, *Proc Eurospeech*, Budapest, 1:117-120, 1999
6. Klasmeyer, G. and Sendlmeier, W. F. Voice and Emotional States, Ed. Kent, R. D. and Ball, M. J. *Voice Quality Measurement*, Singular Publishing Group, 339-359, 1999
7. Klatt, D. H. and Klatt, L. C. Analysis, Synthesis and Perception of Voice Quality Variations among Female and Male Talkers, *JASA*, 87 (2):820-856, 1990
8. Laver, J. *The Phonetic Description of Voice Quality*. Cambridge University Press, 1980
9. Montero, J. M., Guitèrrez-Arriola, J., Colàs, J., Maciàs, J., Enrìquez, E. and Pardo, J. M. Development of an Emotional Speech Synthesizer in Spanish, *Proc Eurospeech*, Budapest, 5:2099-2102, 1999
10. Murray, I. R. and Arnott, J. L. Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion, *JASA*, 93(2):1097-1108, 1993
11. Murray, I. R. and Arnott, J. L. Implementation and Testing of a System for Producing Emotion-by-rule in Synthetic Speech, *Speech Comm.*, 20:85-91, 1995
12. Paeschke, A., Kienast, M. and Sendlmeier, W. F. F_0-Contours in Emotional Speech, *Proc ICPhS*, San Francisco, 2:929-931, 1999
13. Rank, E. and Pirker, H. Generating Emotional Speech with a Concatenative Synthesizer, *Proc ICSLP*, Sidney, 975-978, 1998
14. http://www.kgw.tu-berlin.de/~felixbur/emoSyn.html