ACOUSTICAL ANALYSIS OF SPECTRAL AND TEMPORAL CHANGES IN EMOTIONAL SPEECH

Miriam Kienast, Walter F. Sendlmeier

Technical University Berlin, Institute of Communication Science, Germany
kienast@kgw.tu-berlin.de

ABSTRACT

In the present study, the vocal expressions of the emotions anger, happiness, fear, boredom and sadness are acoustically analyzed in relation to neutral speech. The emotional speech material produced by actors is investigated especially with regard to spectral and segmental changes which are caused by different articulatory behavior accompanying emotional arousal. The findings are interpreted in relation to temporal variations.

1. INTRODUCTION

Many studies of the past 50 years concerned with the vocal expression of emotion dealt with the investigation of prosodic and durational parameters as well as intensity. Recently parameters describing laryngal processes on voice quality were taken into account more intensively [1, 2]. The impact of supralaryngal configurations on vocal quality, such as different articulatory behavior and the accuracy of articulation were not of primary interest. It seems reasonable to take these parameters into more detailed consideration, because articulation is closely related to two other parameters, speech rate and muscular tension [3] influenced by different arousal of the autonomous nervous system. And it is a widely known fact that several emotions differ with respect to these parameters [2, 4].

The few studies which examined changes of articulation in emotional speech aimed at the investigation of this parameter predominantly by examining formants. E. g. Goldbeck et al. [5] or Van Bezooijen [6] compared the real formant position with an ideal one to measure the deviation of the tongue from the aspired position. The method of measuring precision of articulation indexed by formant precision was taken up in this study. By analyzing formant values it was intended to draw conclusions about more global articulatory settings preferred under different emotions.

Furthermore, the degree of segmental changes was examined by analyzing the extent of segmental reductions on the basis of Kohler's concept of reduction forms in German [3]. This concept describes the deletion and assimilation of segments as a result of various principles, namely reduction of effort, listener orientation, gestural reorganization, cognitive constraints, and speaking styles.

Beside this examination of the segmental reduction and the measure of vowel quality the spectral energy distribution in voiceless fricatives was examined. Here the question was addressed, how emotional arousal effects the opening degree and the articulatory effort during consonant articulation.

2. MATERIAL

The database used in this examination consists of ten utterances which were produced by five male and five female German actors enacting the emotions (hot) anger, happiness, fear (panic), sadness (sorrow), boredom, disgust and a neutral version. The sentences were taken from everyday communication and could be interpreted in all emotional contexts without semantic inconsistency. The occurence of articulatory reduction is influenced by the articulatory context of each segment. Thus the phonotactic construction of the sentences provides the possibility of all reduction forms listed by Kohler [3]. Furthermore, they contain all German vowels. The recordings were carried out in separate sessions in an anechoic chamber using a Sennheiser MKH 40 P 48 microphone and a Tascam DA P1 portable DAT-recorder. Because of the differences between read speech and spontaneous speech with respect to articulatory reduction as noted by Kohler [3], emphasis was placed on the actors producing the speech material as spontaneously as possible.

The emotional content of the speech material and its naturalness were evaluated in an automated forced-choice perception experiment by 20 naive listeners. The recorded sentences were played to each listener in a separate session using the same listening procedure via headphones. For each listener a different order of the presented utterances was selected randomly to prevent possible training effects on the recognition rate of single sentences. Only those utterances which were judged as natural and for which the emotion was recognized by at least 80% of the listeners were used for further analysis.

For the work presented in this paper a smaller set of the whole database was selected comprising three sentences of six speakers (three male, three female) and the emotions fear, happiness, boredom, sadness, anger and the neutral state.

Statistic analysis was performed by using a oneway ANOVA and two-tailed t-tests.

3. METHODS AND PARAMETERS

3.1. Assimilations and Segment Deletions

The recorded speech material was digitized (16 kHz, 16 bit). For all selected sentences two label files were created using spectrographic and oscillographic representations as well as auditory analysis. The first label file represents the segmental structure and comprises a narrow transcription of the utterances. The second label file includes the syllable structure of the sentence and a categorization of each syllable into unstressed, word stressed and phrase stressed syllable. The label files were produced

manually by a phonetician, who was uninfluenced by the rest of the investigation.
In order to specify the degree of accuracy of articulation the following segmental parameters were measured.
On the basis of the narrow transcription from the first label file the number of the following reduction forms - described in more detail by Kohler [3] - were analyzed:

- progressive and regressive assimilations with respect to place of articulation and manner of articulation
- progressive and regressive assimilations with respect to nasality and voicing
- changes of opening degree

The omissions of segments were measured according to the 'Lautminderungsquotient' (LMQ) developed by Hildebrandt [7]. The LMQ acts as a comparison between the number of the factually produced segments with the number of segments which should have been produced for standard pronunciation. A disadvantage of this concept is that there is no non-ambiguous definition of standard pronunciation in German. So Hildebrandt's original calculation was modified into an equation that provides the possibility to compare the number of the produced segments in neutral speech (n_{neut}) with the number of segments which was produced in the respective emotional speaking styles (n_{emot}):

$$LMQ = 10 - \frac{10 \cdot n_{emot}}{n_{neut}}$$

Thus, positive values represent segment deletions, while negative values indicate segment insertions compared with neutral speech.

3.2. Vowel Formant Analysis

Acoustic vowel reduction is a well-established phenomenon. Vowels become more centralized when the speaking rate is faster, the speaking style is more informal, or when the vowels are part of unstressed syllables [8, 9]. In order to investigate whether different emotional arousals have an effect on the spectral distinction between vowels, formant values of the first and second formant were determined. Measures were taken approximately at the vowel midpoint of the vowels a, e, i, o, u, E and (IPA-symbols) automatically by LPC-analysis using an esps formant-routine (window type: Hamming, lpc-order 14, window size: 25 ms). The extracted formant values were verified manually and measurement errors identifiable by a very large bandwidth (over 1000 Hz) or extrem values (under 200 Hz, over 3000 Hz) were eliminated. Formant analysis was performed separately for male and female speakers due to sex specific differences concerning formant positions.

3.3. Spectral Balance

As a parameter that serves for the description of acoustic consonant reduction the spectral balance (SB) was calculated according to the following equation [10]:

$$SB = \frac{\sum f_i \cdot E_i}{\sum E_i}$$

(f_i is the frequency in Hz and E_i the spectral power as a function of the frequency.)

In a sense, the spectral balance is the average frequency weighted by the acoustic energy. For turbulent noise, the spectral balance is closely related to the size of the constriction area: the smaller the constriction, the higher the spectral balance. So the spectral balance was taken as a measurement of the degree of the constriction in voiceless fricatives, which is again determined by the articulatory effort. The spectral balance was calculated from separate long term average spectra (FFT) of the fricatives ɸ, σ, Σ, X and ξ (IPA-symbols). It is important to separate the different places of articulation, because the spectral balance is inversely related to the size of the cavity in front of the noise source. Furthermore, only full voiceless fricatives were taken into account. Voiced parts of fricatives, e.g. the beginning of the fricative after a vowel, were eliminated from measurement. Otherwise voice characteristics would have influenced the results and could not be distinguished from articulatory effects.

4. RESULTS

4.1. Assimilations and Segment Deletions

The results obtained by analyzing the number of assimilations and by evaluating the LMQ, revealed the following:
Compared with the neutral speaking style the deletion of segments was more frequent in utterances expressing fear and sadness. Figure 1 illustrates that in anxious utterances there were 7 % of the segments more deleted than in the neutral version. In sad utterances there were 6.6 % more deleted and in sentences expressing boredom 5.1 %. These differences are all statistically highly significant ($\alpha \leq 0.01$).
Analyzing the number of assimilations led to similar findings (see figure 2). The number of assimilations in the utterances expressing fear, boredom and sadness was higher than in the

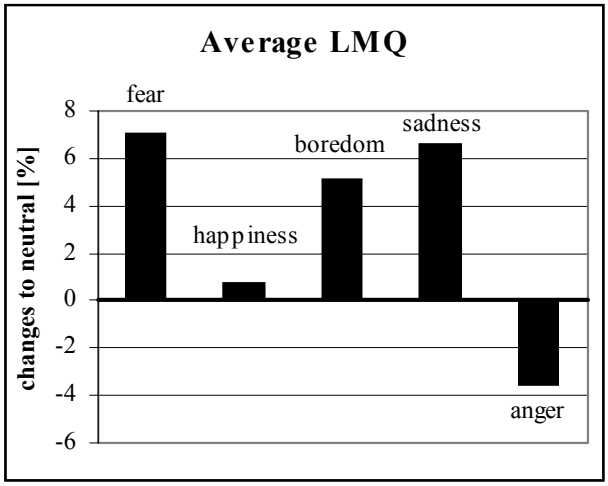

Figure 1. Average LMQ

neutral version and all other emotions. The results for sadness and fear are statistically significant ($\alpha \leq 0.05$), but not for boredom.

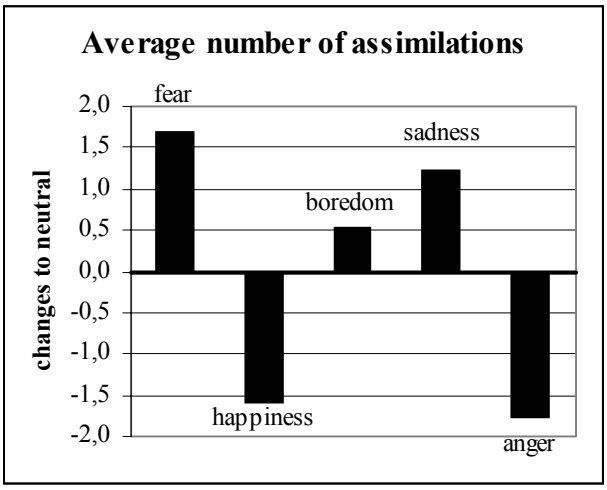

Figure 2. Average number of assimilations

In contrast, sentences representing anger showed a more accurate pronunciation. Deletion of segments (LMQ) was observed less frequently than in all other emotions including neutral speech (figure 1). This effect is statistically highly significant ($\alpha \leq 0.01$). Furthermore, the number of assimilations was significantly lower than in neutral speech ($\alpha \leq 0.05$) as can be seen in figure 2.

The results of the utterances expressing happiness showed a slightly higher LMQ-value than the neutral version, but do not differ significantly. Concerning the number of assimilations the results indicated a lower value for sentences with happy content (figure 2). This effect is statistically significant ($\alpha \leq 0.05$).

4.2. Vowel Formant Analysis

With regard to vowel formant analysis it was observed that sentences expressing fear, sadness or boredom are characterized by a formant shift towards a centralized position in all different vowels. Overall, the vowel chart became smaller. The different vowels were less distinct than in neutral speech and the tense vowels moved in the direction of the corresponding lax categories. For practical reasons not all results can be shown in this paper. So the figures (3 to 7) illustrate the findings only for either female or male speakers.

There were statistically significant differences between formant values in neutral and sad speech for the vowels a, e (only the second formant), u ($\alpha \leq 0.01$) as well as ɛ and o ($\alpha \leq 0.05$). Differences with respect to i and ɔ (only the first formant) were only significant for female speakers ($\alpha \leq 0.05$). Significant differences could also be found between fear and neutral for the vowels a and e ($\alpha \leq 0.01$) as well as i and o ($\alpha \leq 0.05$). Regarding the formant values of ɔ there were significant differences only for female speakers ($\alpha \leq 0.05$). Furthermore, there were significant differences between formant values in neutral speech and utterances expressing boredom for the vowels a and e ($\alpha \leq 0.05$). For male speakers there were also significant differences for o and the first formant of i ($\alpha \leq 0.05$).

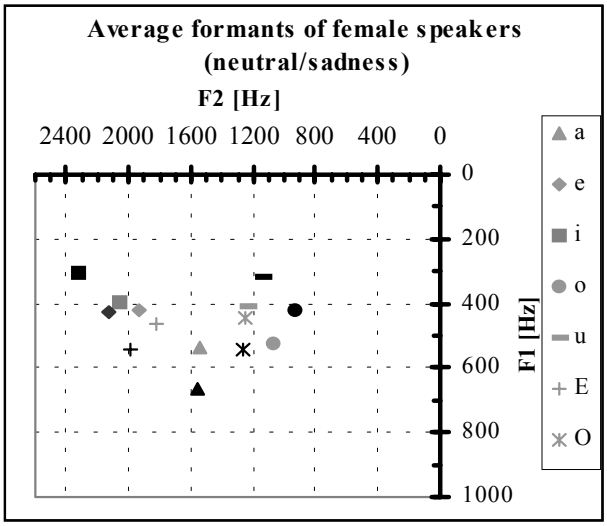

Figure 3. Average formant values for sadness (grey symbols) compared with neutral (black symbols) of female speakers.

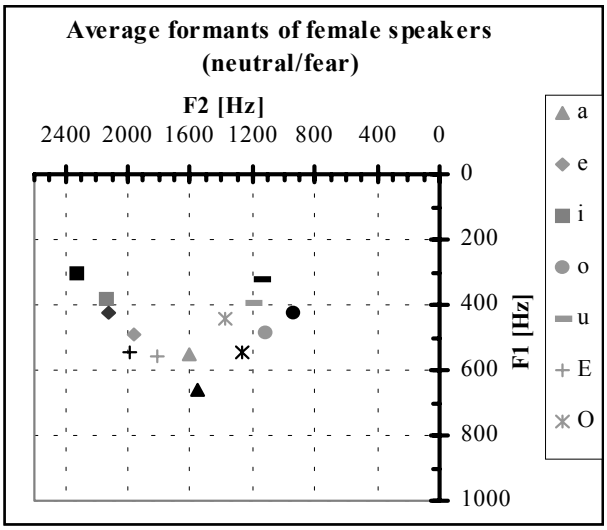

Figure 4. Average formant values for fear (grey symbols) compared with neutral (black symbols) of female speakers.

The results obtained by analyzing the formant values in the sentences with angry emotion indicate the opposite to the findings for sadness, fear and boredom. In angry sentences a displacement of the vowels in a more extreme part of the vowel chart was observed in the speech samples of both male and female speakers. Overall, the vowel chart became larger and the different vowels became more distinct (as could be seen in figure 6). Statistically significant effects were found for the vowels a ($\alpha \leq 0.01$), e (only the second formant) and ɛ ($\alpha \leq 0,05$). The vowel i showed a significant difference only for the male speakers ($\alpha \leq 0.05$) and the vowel u only for the female speakers ($\alpha \leq 0.05$).

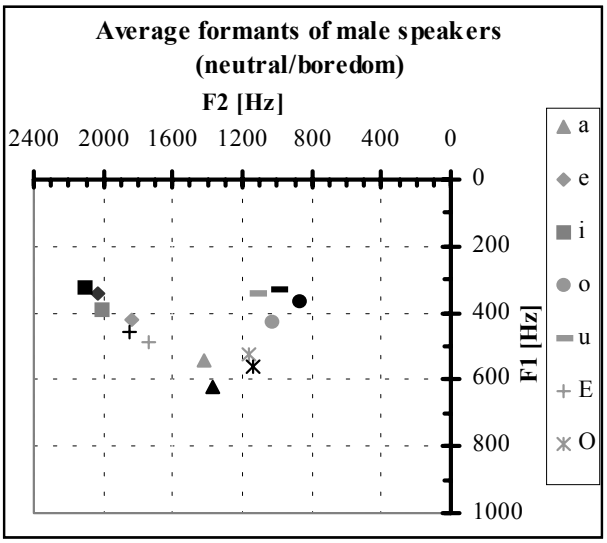

Figure 5. Average formant values for boredom (grey symbols) compared with neutral (black symbols) of male speakers.

For happy sentences the results reveal a general rise of both formants, whereas the first formant of most vowels seems to be more affected than the second one (figure 7).
There were significant differences between the formant values in sentences with happy content and the neutral version with respect to the following vowels: a ($\alpha \leq 0.01$), e, i, ɛ and ɔ (only for the second formant) ($\alpha \leq 0.05$). For the vowels u and o significant differences were only found for male speakers ($\alpha \leq 0.05$).

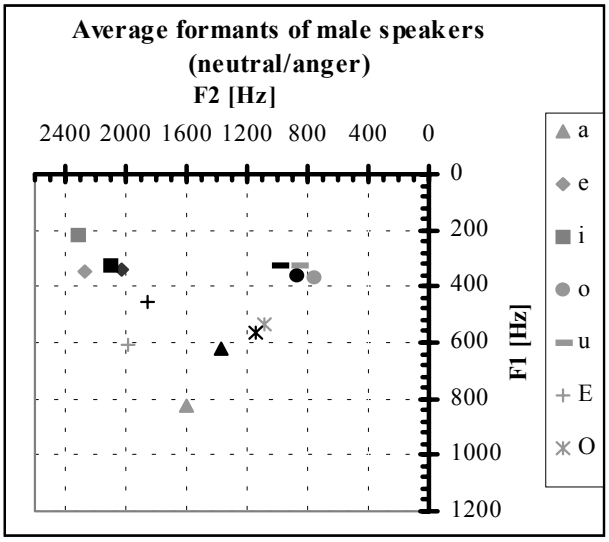

Figure 6. Average formant values for anger (grey symbols) compared with neutral (black symbols) of male speakers.

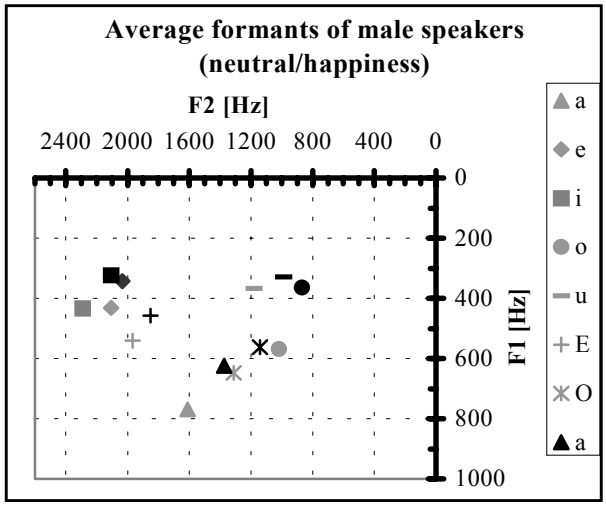

Figure 7. Average formant values for happiness (grey symbols) compared with neutral (black symbols) of male speakers.

4.3. Spectral Balance

The results obtained by calculation of the spectral balance reveal that the investigated emotions can be splitted into two different groups. The first group consists of the emotions fear, happiness and anger. The voiceless fricatives in utterances expressing these emotions show an increased spectral balance compared to neutral speech (see figures 8 to 12). This rise of energy in the higher frequencies indicate a closer articulation in the fricative production in angry, happy and anxious speaking style.
In the second group containing the emotions boredom and sadness the values of the spectral balance decrease in comparison to neutral speech. This is an indication of less constriction in production of voiceless fricatives in speech with sad or bored emotions.

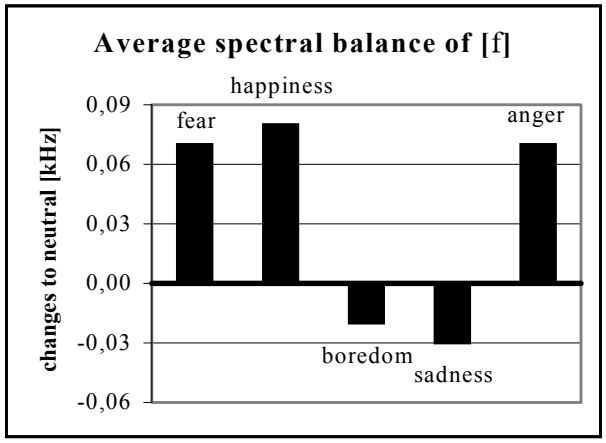

Figure 8. Average spectral balance for the fricative [f] related to neutral speech.

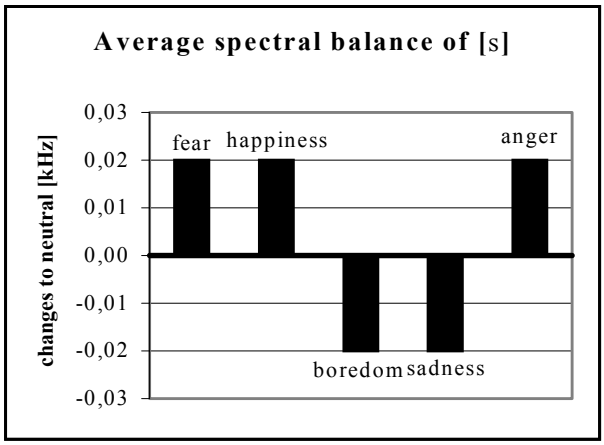

Figure 9. Average spectral balance for the fricative [s] related to neutral speech.

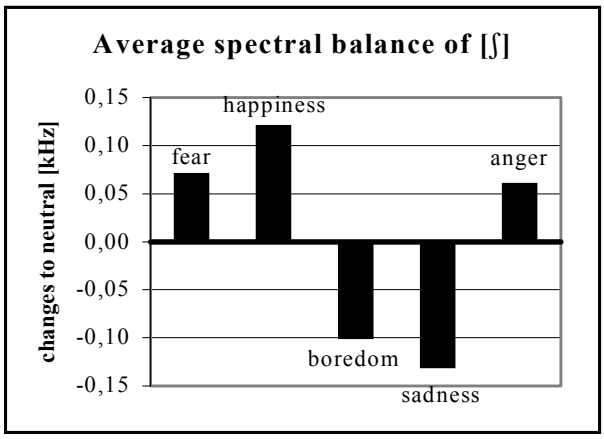

Figure 10. Average spectral balance for the fricative [ʃ] related to neutral speech

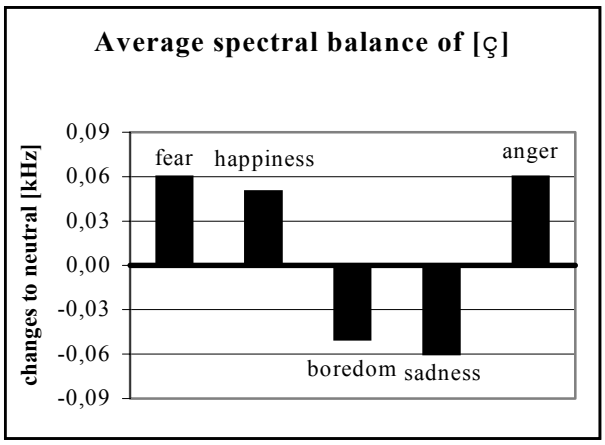

Figure 11. Average spectral balance for the fricative [ç] related to neutral speech.

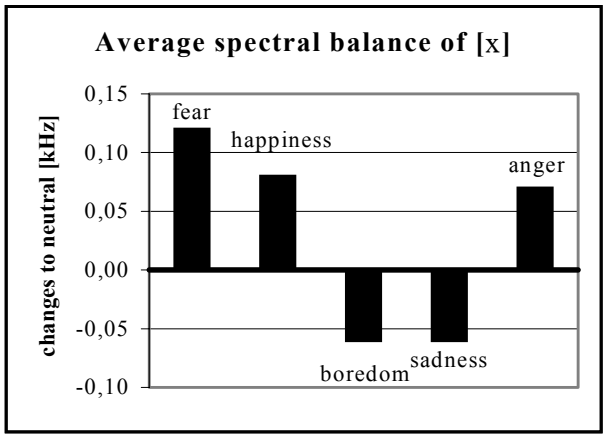

Figure 12. Average spectral balance for the fricative [x] related to neutral speech.

For all examined fricatives there were statistically significant differences between anger, happiness and fear on the one hand and sadness and boredom on the other ($\alpha \leq 0.05$). The difference between fear and neutral was only significant for the fricatives ɸ and x ($\alpha \leq 0.05$). For anger there were no significant effects for the fricative ʃ, and in sentences representing happy speech for the fricatives s and ç.

5. SUMMERY AND DISCUSSION

In the present study the vocal expression of the emotions fear, sadness, boredom, anger and happiness show various specific characteristics concerning vowel quality, segmental reduction and energy distribution in voiceless fricatives.

The utterances expressing anger showed the highest accuracy of articulation compared with the other emotions and the neutral version. This results from findings on segmental reduction and vowel formant analysis. These outcomes correspond with other studies [4, 5, 6] which report more distinct opening and closing movements and a higher vowel quality for anger.

The higher accuracy in vowel production may be explained by the fact that vowels in sentences with angry emotion are prolonged [11] compared with neutral speech. The vowel overshoot obtained in angry utterances can also be interpreted in the light of findings on prosodic features by Paeschke et al. [12]. Paeschke et al. found an increase in the number of heavily stressed syllables and a decrease of unstressed syllables in angry sentences. And vowels in stressed syllables as well as long vowels are reported to be articulated more precisely [8, 9]. The results concerning the spectral balance in voiceless fricatives indicate a smaller constriction area for sentences with angry content. This can be seen as evidence for an increase of articulatory effort in angry emotion.

Furthermore, the results suggest that vowel undershoot and segmental reduction is most frequent in the utterances expressing fear, boredom and sadness. This seems reasonable for fear, as it is produced with the fastest speech rate [11] which is one of the most important parameters for the degree of reduction [3]. An additional reason for this findings in utterances expressing fear is that vowels are shorter in anxious sentences and only a few

heavily stressed syllables can be found [11, 12]. The analysis of the spectral balance in sentences expressing fear revealed that the constriction area is smaller compared with neutral speech. A possible explanation for this result may be that consonant production, especially the articulation of voiceless fricatives, is more accurate in contrast to the less precise production of voels. This corresponds to the finding that consonants, especially fricatives are less shortened in anxious utterances than vowels [11]. Another explanation could be that in anxious speaking style there is in general a close articulatory setting. But this should be verified by acoustical analysis of other consonants.

For sadness and boredom the reduction of segments and formants is remarkable, because here the speech rate is slower than in the other emotions including the neutral version [11]. An explanation for the high degree of articulatory simplification in sad and bored utterances may be a low muscular tension as a result of the low arousal level associated with sadness and boredom. This thesis is supported by the results concerning the energy distribution in voiceless fricatives. The reduced constriction could be seen as indicator for less articulatory effort in speech with sad or bored emotion. The results concerning vowel reduction in sad utterances correspond to findings of other studies on formant precision [4, 5, 6].

In sentences with happy emotion the analysis of the formant frequencies reveal an increase of both the first and second formants. This result can be explained by a general shortening of the vocal tract. The shortening could be caused by lip spreading occuring during smiling as noted by Tartter [13]. A supplemental factor for this effect can also be seen in a rising of the larynx which is often accompanied by a high fundamental frequency. A setting of a raised larynx in happy speech is probable, because happy speech is reported to have generally a great increase in mean F_0 [2, 4] which is also the case for the sentences used [12]. The result that the first formant was more affected than the second can be interpreted by a greater opening degree during mandible movements in happy speech.

Concerning segmental reduction in happy utterances, a consistent conclusion can hardly be drawn. The accuracy of articulation in utterances expressing happiness seems to be like that in neutral speech, respectively a little bit more precise. Furthermore, in line with the higher arousal reported for happiness [2, 4] the production of voiceless fricatives was found to be more narrow (higher spectral balance), i. e. with higher articulatory effort.

Since the results of this study are based on a small sample size, caution should be taken when making any generalized conclusions. Therefore, the described analysis will be expanded to the rest of the database. Beside more detailed acoustical and segmental analysis future work on this issue may take into consideration a more direct gathering of articulatory data in emotional speech, e.g. by articulographic or palatographic measurements.

ACKNOWLEDGEMENT

This work was partly supported by the DFG (German research foundation, Se 462/3-1)

REFERENCES

1. Klasmeyer, G. & Sendlmeier, W. F. (2000): *Voice and emotional states*. In: Kent, R. D. & Ball, M. J. (eds.): Voice quality measurement. San Diego: Singular, p. 339-357.
2. Banse, R. & Scherer, K. R. (1996): Acoustic profiles in vocal emotion expression. *Journal of Personality and Social psychology*, Vol. 70, No. 3, p. 614-636
3. Kohler, K. (1990*): Segmental reduction in connected speech in German: phonological facts and phonetic explanations*. In: Hardcastle, W. J. & Marchal, A. (eds.): Speech production and speech modelling. Dordrecht: Kluwer, p. 69-92
4. Murray, I. R. & Arnott, J.L. (1993): Towards a simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *JASA* 93 (2), p. 1097-1108
5. Goldbeck, T., Tolkmitt, F. & Scherer, K. R. (1988): *Experimental studies on vocal content communication*. In: K. R. Scherer (ed.): Facets of emotion. Recent Research. Hillsdale: Lawrence Erlbaum, p. 119-137
6. Van Bezooijen, R. (1984): *The characteristics and recognizability of vocal expressions of emotions.* Dordrecht: Foris
7. Hildebrandt, B. (1963): Die arithmetische Bestimmung der durativen Funktion. Eine neue Methode der Lautdauerbewertung. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 14, p. 328-336
8. Lindblom, B. (1963): Spectrographic study of vowel reduction. *JASA* 35, p. 1773-1781
9. Moon, S.-J. & Lindblom, B. (1994): Interaction between duration, context, and speaking style in English stressed vowels. *JASA* 96, p. 40-55
10. Van Son, R. J. J. H. & Pols, L. C. W. (1999): An acoustic description of consonant reduction. *Speech communication* 28, p. 125-140
11. Kienast, M., Paeschke, A. & Sendlmeier, W. F. (1999): Articulatory reduction in emotional speech. *Proceedings Eurospeech 99*, Budapest, Vol. 1, p. 117-120
12. Paeschke, A., Kienast, M. & Sendlmeier, W. F. (1999): F0-contours in emotional speech. *Proceedings ICPhS 99*, San Francisco, Vol. 2, p. 929-932
13. Tartter, V. C. (1980): Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception and Psychophysics*, Vol. 27 (1), p. 24-27