OBJECTIVE VOICE PARAMETERS TO CHARACTERIZE THE EMOTIONAL CONTENT IN SPEECH

Gudrun Klasmeyer, Walter F. Sendlmeier
TU Berlin, Institute of Communication Science, Germany
klasmeyer@kgw.tu-berlin.de

ABSTRACT

In the present study, the hypothesis is tested that voice parameters derived from clinical measurement of pathological voices (Jitter, Shimmer, Harmonics-To-Noise-Ratio) and the glottal pulse shape could serve as useful features to characterize specific emotional contents in spoken utterances.

DATABASE

The emotionally loaded speech material was produced by three students of acting (2 male, 1 female). It was DAT-recorded in separate sessions in an anechoic room using a B&K measuring microphone. The utterances are 10 short sentences frequently used in everyday communication which could appear in all emotional contexts without semantic contradictions. Each utterance was spoken several times in a neutral voice and several times with each of the following emotions: happiness, sadness, anger, fear, boredom and disgust. The most appropriate realisation was selected by the authors and the respective actor for a listening test.

LISTENING TEST

For each actor, the selected 70 sentences were randomized. Every sentence was repeated three times in short intervals followed by a 30 second pause. A 1 kHz tone announced the next triplet of sentences. The series of stimuli were acoustically presented via headphones to 20 naive listeners in separate sessions to evaluate the emotional content within 8 categories: neutral, happiness, sadness, anger, fear, boredom, disgust and not recognizable emotional content. Only those sentences recognized by at least 80% of all listeners were used for the analysis.

PARAMETERS

The sound of emotional speech differs from that of neutral speech, which is partly due to differing articulator movements and partly due to differeing glottis behaviour. In clinical measurement of pathological voices, sustained signals of open vowels are used for analysis. In fluent speech however, the role of articulator movements and intonation has to be considered. The glottal pulse signal can be derived from the acoustic speech signal by inverse filtering. The pulse shape contains information about glottis movement. In this context also voicing irregularities (Jitter [1] and Shimmer [1]) are discussed. Another parameter investigated in this study is the Harmonics-To-Noise-Ratio [2]. A phoneme based set of Energy Distribution Parameters is developed to differentiate between specific emotional contents in the frequency domain.

I. GLOTTAL PULSE SHAPE

Theoretically the process of glottal closure resembles an impulse. In the following closed-glottis-interval, the acoustic speech signal can be interpreted as impulse response of the vocal tract, because the subglottal volume is decoupled from the upper tract during this time. The filter coefficients for inverse filtering are calculated during the closed-glottis-interval. The actual point of glottal closure can be determined by inverse filtering with filter coefficients derived from a time interval (3-4 times) longer than one period duration of the acoustic signal [3]. In the present study, 18th order covariance LPC and rectangular data windows are used. The filter coefficients for inverse filtering are calculated during the closed-glottis-interval of the middle period of each realisation of the German phoneme /a/ in the emotional speech database. Due to different period durations, the length of the data window had to be adapted with regard to reasonable spectral shaping of the inverse filter. 100 ms of the speech signal are filtered. Only the middle period within which the LPC coefficients were calculated is examined further. In emotional speech the glottal cycles vary considerably from normal speech. So the inverse filtered signal should be interpreted as glottal pulse signal with great care, because some of the premises for this theory may be violated. For example, pulses filtered from sad speech hardly show any obvious closed-glottis-interval and (or because) the closure is not abrupt. But still the shape of the inverse filtered signal does represent important characteristics of the glottal cycle. In general it is more difficult to determine the exact point where the opening begins. The relative duration of the closing phase can be measured more reliably. To parameterize its shape the filtered signal is amplitude-normalized. The duration of the closed-glottis-interval (T1), the opening-phase (T2) and the closing-phase (T3) are measured as fractions of the full period. (See figure 1. for explanation.)

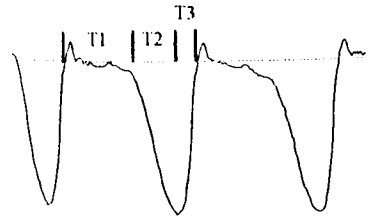

Figure 1. Parametrization of the inverse filtered signal

Within the speech database a relation between specific glottal pulse shapes and specific emotions could be proved. The general correlations are not speaker-dependent. For example in angry speech, pulses show very abrupt closure and remarkably long closed-glottis-intervalls. In contrast sad utterances show hardly any closed-glottis-intervals, and in anxious speech the relative duration of glottal closing is nearly similar to that of glottal opening. Table 1. gives the precise data.

	T1	T2	T1+T2	T3
Neutral	0.15	0.65	0.80	0.20
Happiness	0.33	0.43	0.76	0.24
Sadness	0.0	0.72	0.72	0.28
Anger	0.47	0.40	0.87	0.13
Fear	0.26	0.40	0.66	0.34
Boredom	0.25	0.48	0.73	0.28

Table 1. Average durations of closed-glottis-interval (T1), opening-phase (T2) and closing-phase (T3) as fraction of the full period

It is clear that narrow pulses with short opening and closing phases show less high frequency damping of their harmonics. Discussing the perceptible influence on the speech signal the absolute amplitude and period duration has to be taken into consideration, too. The pulse shapes measured in this study do not correlate with period duration. It can rather be assumed that pulse shapes correlate with loudness. This could not be tested, however, because the microphone signals for different specific emotions were recorded at the same level.

II. VOICING IRREGULARITIES

In clinical voice measurement, the patient produces isolated vowels with flat F0 contours, whereas in fluent speech there is a permanent rise and fall due to the intonation pattern. This implies that the parameter Jitter has to be measured taking into account such meaningful variation of F0. This is done by calculating a polynom approximation of the F0 contour and subtracting this approximation from the measured values. The difference values are used to calculate the absolute jitter, which has to be interpreted taking into consideration the absolute period duration [4].

In the emotional speech database remarkably high Jitter values were found in all anxious utterances produced by the female actress and in most anxious utterances produced by one male actor. This male actor also

laryngealized most vowels in sad utterances. The other male actor showed no voicing irregularities.
Shimmer was not found in the speech database used for this study.

III. HARMONICS-TO-NOISE-RATIO

In neutral speech the energy of harmonics is damped in higher frequencies, so there is very little energy above 4 kHz in vowels, whereas in fricatives there is very little energy below this frequency. In emotional speech, high frequency noise can be present in vowels which has its origin in abnormal articulation. For example with fear, the face can be stiff, and teeth are pressed together. This produces fricative noise in vowels, which does not derive from the voice source, but can be measured with the same parameter used in clinical measurement of pathological voices to detect noise produced by imperfect closure of the glottis. Especially in bored speech the articulation is imprecise, and voiceless fricatives in VCV clusters tend to be voiced. This effect can also be detected with the Harmonics-To-Noise-Ratio.

IV. SPECTRAL ENERGY DISTRIBUTION

Discussing the Harmonics-To-Noise-Ratio some phenomena were explained, by which energy is shifted to different frequency regions. Also specific glottal pulse shapes correlate with specific spectral damping of harmonic energy. From these considerations a set of Energy-Distribution-Parameters is developed on a phoneme basis to differentiate between specific emotional contents in spoken utterances. A spot check revealed that the introduction of 4 frequency bands leads to meaningful parameters for the characterization of different emotional contents in spoken utterances on a phoneme basis. The acoustic speech signal is lowpass filtered with an adaptable filter cutting all energy above F0. This is the very-low-frequency-band (VL). A second lowpass filter with constant 1.5 kHz cut off frequency is used to produce a low-frequency-band (L) signal. The middle-frequency (M) region between 1.5 and 4 kHz is filtered from the acoustic signal with a bandpass. The high-frequency (H) region between 4 and 8 kHz is filtered from the speech signal with a highpass. The energy distribution is compared within different emotionally loaded realisations of the same phoneme by measuring the energy within single bands as fraction of the total energy of that phoneme. As an example, the results for the vowel /a/ for one male speaker are presented in table 2. Significant frequency shifts for specific emotions are also apparent in fricatives. It is clear that the energy distribution within the frequency bands is speaker dependent; for example, female speakers have higher formants than male speakers. But the general tendency of energy shifts correlating with specific emotions is not speaker dependent. Ratios of different frequency bands can be calculated to discriminate specific emotions even stronger, but for a general survey the fractions of total energy are more illustrative.

	VL	L	M	H
Neutral	0.036	0.964	0.024	0.013
Happiness	0.069	0.879	0.045	0.017
Sadness	0.364	0.966	0.007	0.031
Anger	0.016	0.850	0.081	0.024
Fear	0.224	0.843	0.084	0.060
Boredom	0.170	0.988	0.006	0.007
Disgust	0.150	0.683	0.178	0.084

Table 2. Average distribution of energy within frequency bands as fraction of total energy in the vowel /a/

In the vowel /a/ spoken in a neutral voice most energy is below 1.5 kHz. In most emotionally loaded utterances some energy is shifted to the middle and even to the high frequency region. Only in sad and bored speech there is less energy in middle and high frequencies. A remarkable difference between fear and anger is that in angry voice there is little energy in the very-low-frequency band, though F0 is often high in angry utterances, whereas in utterances spoken with fear the appearance of energy in high frequencies is combined with high values in the very-low-frequency band. In bored and sad utterances there is also much energy in the very-low-frequency-band, even though F0 is usually very low in these utterances. The results of the energy-distribution measurement confirm very well what could be expected from the analysis of the glottal pulse shapes.
Examining the energy distribution on a phoneme basis leads to many interesting results. There is also emotion specific information in the shift of energy in vowel-fricative transitions and in the ratio of total vowel-to-fricative energy. The emotion specific differences in the distribution of energy are visualized in figure 2. The narrowband spectrogram of the preemphasized signal is calculated and amplitude-normalized. All values below a fixed threshold are presented in white, all values above this threshold are printed in black. In these 'binary spectrograms' different energy distributions in emotionally different realisations of the same utterance are obvious.

SUMMARY

It could be shown that parameters derived from clinical measurement of pathological voices and the glottal pulse shape are useful to characterize specific emotions in spoken utterances. Discrimination of these specific emotional contents is possible from the examination of spectral energy distributions on a phoneme basis. The results correlate very well with predictions from the examination of glottal pulse shapes.

REFERENCES

[1] Orlikoff R.F., Baken R.J., *Clinical Speech and Voice Measurement*, Sing.Publ.Group, 1993
[2] Deliyski D.D., *Acoustic Model and Evaluation of Pathological Voice Production*, Kay Elemetrics Corp., Dept. of Development and Research, 1993

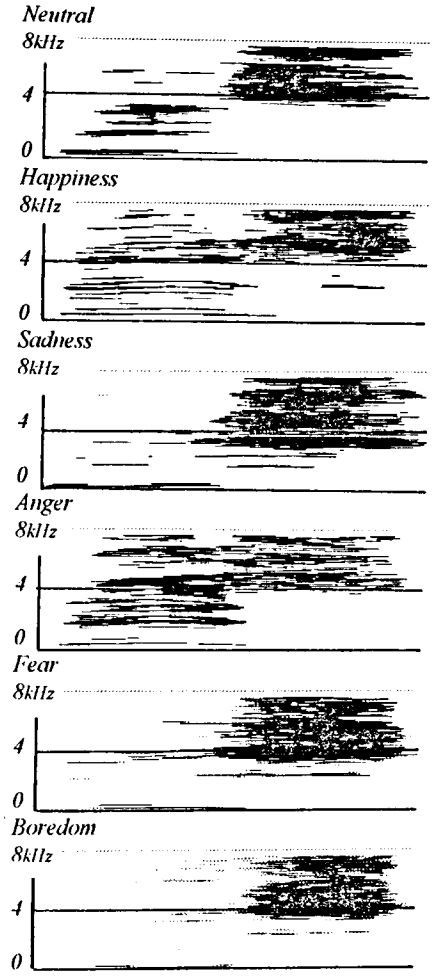

Figure 2. 'Binary Spectrograms' of diffrent emotionally loaded realisations of the phonemes /i/, /s/

[3] Wong D., Markel J.D., Gray A.H., *Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform*, IEEE Transactions ASSP Vol. 27(4), 1979
[4] Rosken W., Klasmeyer G., *Erfassung von F0-Irregularitäten in gesprochener Sprache als messbarer Parameter zur Beschreibung von Stimmqualitäten*, Fortschritte der Akustik, DAGA '95