INTERSPEECH 2009 BRIGHTON

On the relevance of high-level features for speaker independent emotion recognition of spontaneous speech

Marko Lugger and Bin Yang

Chair of System Theory and Signal Processing, Universität Stuttgart, Germany

Abstract

In this paper we study the relevance of so called high-level speech features for the application of speaker independent emotion recognition. After we give a brief definition of high-level features, we discuss for which standard feature groups high-level features are conceivable. Two groups of high-level features are proposed within this paper: a feature set for the parametrization of phonation called voice quality parameters and a second feature set deduced from music theory called harmony features. Harmony features give information about the frequency interval and chord content of the pitch data of a spoken utterance. Finally, we study the gain in classification rate by combining the proposed high-level features with the standard low-level features. We show that both high-level feature sets improve the speaker independent classification performance for spontaneous emotional speech.

Index Terms: emotion recognition, high-level features, harmony features, voice quality parameters

1. Introduction

In the area of emotion recognition usually low-level features of basic acoustic characteristics are used. Among them, prosodic and spectral parameters are the most popular. In contrast to other areas of acoustic signal processing, e.g. music information retrieval (MIR), the idea of using high-level speech features is not very common in the community of speech signal processing. There are mainly two reasons that are responsible for that fact. On the one hand, for historical reasons, spectral speech features are still state of the art in many speech processing application, e.g. automatic speech recognition. On the other hand, there are no apparently well performing and universally applicable high-level features for the different speech applications.

In contrast to low-level features that can be extracted by a straight forward method, high-level features are always based on a certain theory or model. These models can be adopted from other disciplines and should be verified by expert knowledge. Mathematic formulas deduced from this expert knowledge lift the features onto a higher-level and distinguish them from standard low-level features. That requires interdisciplinary research in cooperation with for example, linguists, psychologists, medical scientists, or even musicians.

For the application of emotion recognition, we see the potential for high-level features mainly in three speech feature groups: intonation, spectral information of voiced speech, and speech rhythm. For these feature groups low-level features do not satisfactory describe the complexity of the information contained in the speech data. For both of the first two groups a feature set is proposed in this paper. Rhythm features had to be built upon the low-level despriptors of both energy and duration but are not presented here.

The paper is organized as follows. Section 2 summarizes the basic acoustic low-level speech features. Then two approaches for the extraction of high-level features called voice quality parameters and harmony features are introduced in section 3 and 4, respectively. In section 5 the proposed high-level and low-level features are compared by three classification experiments using the FAU AIBO database [1].

2. Standard speech features

In the area of emotion recognition, mainly two important feature groups can be distinguished. These are spectral features and the well known prosodic features. Statistical functionals are applied to mathematically describe the shape of the contours of both feature groups. By doing this we obtain the so called low-level features.

2.1. Prosodic features

There are three main subgroups of prosodic features: intonation, intensity, and duration. The intonation of a spoken utterance can be approximated by the corresponding pitch contour. This is a vector containing the fundamental frequency for every speech segment. The intensity of the uttered words is covered by the signal energy. In our implementation, duration is parametrized by the speed of talking and both the length and number of pauses within an utterance.

2.2. Spectral features

The whole spectral content of a spoken utterance is contained in spectral parameters. As state of the art in automatic speech recognition, mel frequency cepstral coefficients (MFCC) are the most popular spectral features. Usually 13 MFCC parameters and the first and second derivatives are used. Another feature set corresponding to the spectral feature group is the zero crossing rate (ZCR). It simply counts the zero crossings of the time signal within a defined time frame.

2.3. Feature contour and statistics

By concatenating the feature values from the single speech segments, a so called feature contour in obtained. Low-level features are extracted by measuring statistical values of the corresponding feature contours. Mean, median, minimum, maximum, range, interquartile range, and variance are the most often used functionals. Thus low-level features are describing both the mean level and the variability of the feature contour. Sometimes, higher order statistics as skewness or kurtosis are also extracted. All together our low-level feature set contains 290 features. The numbers of low-level features belonging to the different feature subgroups are summarized in Table 1.

features	energy	pitch	duration	spectral	overall
low-level	75	55	16	144	290
high-level	-	7	-	99	106

Table 1: Number of low-level and high-level features

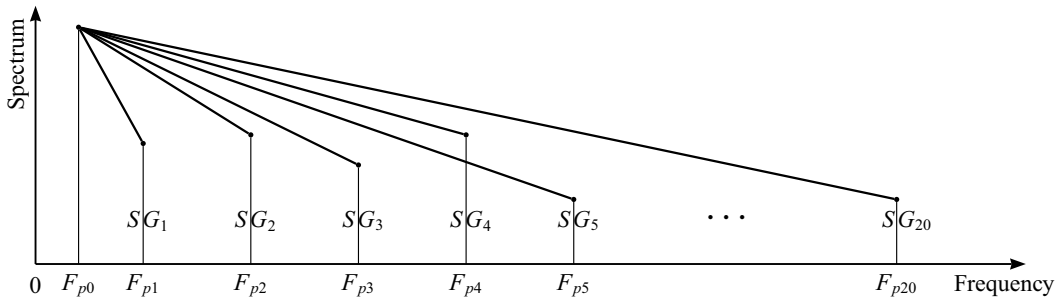

Figure 1: Spectral gradients at fixed frequencies

3. Voice quality features

Voice quality describes the phonation type of speech like modal, breathy, rough, or creaky voice. Besides the standard prosodic aspects, voice quality is an important factor in conveying emotional information. Hence, one idea of generating spectral high-level features is based on the well known source filter model of speech production. It originates from the expert knowledge that the non-interacting processes of phonation and articulation can be separated by a method called inverse filtering. After that, independent parameters for both the phonation (VQP) and the articulation (ART) can be extracted. While the parameterization of the articulation process by formants is a state of the art method, the parameterization of the glottal source activity for emotion recognition is a less explored field. We propose a method in the frequency domain to calculate gradients of the glottal excitation spectrum.

3.1. Measurement of basic speech features

First, we estimate some well known basic speech features from windowed, voiced segments of the speech signal, see Table 2. We perform the voiced-unvoiced decision and the pitch estimate F_0 according to the RAPT algorithm [2] that looks for peaks in the normalized cross correlation function. To measure the spectral gradients, we also use higher harmonics. In order to get a fixed number of 20 gradients, we extract the harmonics F_{pk} next to fixed frequencies at multiples of 200 Hz. So all together 21 harmonics are used, which cover the relevant frequency range for voice quality up to 4000 Hz. The frequencies and bandwidths of the first four formants are estimated by an LPC analysis [3].

feature	meaning
$F_{p0} = F_0$	pitch
$F_{p0}, \ldots, F_{p20}$	frequency of harmonics
$H_0, \ldots, H_{20}$	amplitude at $F_{p0}, \ldots, F_{p20}$ [dB]
F_1, F_2, F_3, F_4	formant frequencies
B_1, B_2, B_3, B_4	formant bandwidths

Table 2: Speech features for voice quality parameter estimation

3.2. Compensation of the vocal tract influence

Since the voice quality parameters shall only depend on the excitation and not on the articulation process, the influence of the vocal tract has to be compensated. This is done by subtracting terms which represent the vocal tract influence from the amplitudes of each harmonic H_k as described in [4]. The amplitudes of the compensated harmonics are $\tilde{H}_k$.

3.3. Estimation of the voice quality parameters

Former approaches used only 4 amplitude quotients to characterize the glottal source signal [5]. In order to better parameterize the glottal excitation signal for emotion recognition this parameter set is extended to 20 gradients. Figure 1 illustrates the definition of the spectral gradients (SG).

$$\mathrm{SG_k} = \frac{\tilde{H}_0 - \tilde{H}_k}{F_{pk} - F_{p0}} \qquad (k = 1, \ldots, 20) \tag{1}$$

In addition to the 20 gradients normalized to the linear frequency difference $\Delta f_k = F_{pk} - F_{p0}$, the same amplitude differences $\Delta\tilde{H}_0 - \tilde{H}_k$ are also normalized to frequency differences in both octave and bark scale. Octave is a logarithmic scale

$$\mathrm{octave}\,(k) = \log_2 \frac{F_{pk}}{F_{p0}} \tag{2}$$

and the bark scale is based on the human auditory system:

$$\mathrm{bark}\,(\Delta f_k) = 13 \tan^{-1}(0.00076 \cdot \Delta f_k) + 3.5 \tan^{-1}\left(\left(\frac{\Delta f_k}{7500}\right)^2\right) \tag{3}$$

In addition, the four formant bandwidths B_n normalized to the corresponding formant frequencies F_n are calculated.

$$\mathrm{IC_n} = \frac{B_n}{F_n} \qquad (n = 1, \ldots, 4) \tag{4}$$

Another three more voice quality parameters describe the voicing, the harmonicity, and the periodicity of the signal, see [6]. In total, we obtain a set of 67 voice quality features. Together with formant frequencies and their bandwidths we estimate 99 spectral high-level features.

4. Harmony features

Starting from infancy, we are influenced by music. That leads to the fact that even children without any musical education are able to distinguish between melodies having positive or negative modality [7]. The key for either positive or negative impression on the listener is mainly the harmonic structure of music. This raises the question whether people also adapt these structures in their own speech prosody. In case they do, especially the speech melody called intonation may be influenced. Thus, we try to detect and quantify these basic harmonic structures in the pitch data of spoken utterances by the parameters dissonance, tension, and modality. These terms were also used by [8] upon approximating the pitch histogram by mixtures of Gaussians.

Our proposed method is based on the second- and third-order autocorrelation of the circular pitch histogram. The basic idea is to scan the pitch data regarding the intensity of intervals respectively the triads: major, minor, diminished, and augmented. Based on the detected intervals and triads, the parameters dissonance, tension, and modality are estimated.

4.1. Circular pitch histogram

The pitch contour $F_{0,\mathrm{Hz}}$ of voiced segments of a spoken utterance is extracted by normalized crosscorrelation followed by dynamic programming (RAPT algorithm). It is then transformed to the logarithmic semitone scale, where one octave (frequency ratio of 2) contains 12 semitones (ST).

$$F_{0,\mathrm{ST}} = 12 \log_2 \left\{ \frac{F_{0,\mathrm{Hz}}}{F_{\mathrm{ref}}} \right\} \quad (5)$$

According to the music theory, the perception of a 2-tone interval or a 3-tone chord should be invariant with respect to the modification of tone frequencies by powers of 2 (octaves). Hence, we map all pitch values $F_{0,\mathrm{ST}}$ to one octave by applying a modulo-12 operation.

$$F_{0,\mathrm{ST,mod}} = \mathrm{mod}_{12} \{F_{0,\mathrm{ST}}\} \quad (6)$$

The circular pitch histogram is calculated by quantizing all the pitch values $F_{0,\mathrm{ST,mod}}$ to L bins per ST, resulting in a histogram $h(n)$ $(0 \le n \le M-1)$ with $M = 12\,L$ bins, see Figure 2.

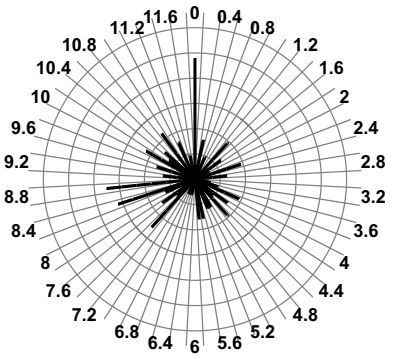

Figure 2: Circular pitch histogram with $L = 5$ bins per semitone

4.2. Higher-order autocorrelation of pitch histogram

To obtain the interval and chord content, autocorrelations of the circular pitch histogram $h(n)$ are calculated. The value of the autocorrelation at a certain lag corresponds to the intensity with which a certain tone combination is present in the pitch histogram $h(n)$. For detecting 2-tone intervals, we determine the second-order circular autocorrelation:

$$r_{hh}(k) = \frac{1}{M} \sum_{n=0}^{M-1} h(n)\, h\left(\mathrm{mod}_M(n+k)\right) \quad (0 \le k \le M-1) \quad (7)$$

Similarly, we compute the third-order circular autocorrelation to measure the intensity of a 3-tone combination :

$$r_{hhh}(k,l) = \frac{1}{M} \sum_{n=0}^{M-1} h(n)\, h\left(\mathrm{mod}_M(n+k)\right)\, h\left(\mathrm{mod}_M(n+l)\right) \quad (8)$$

4.3. Harmony parameters

Dissonance DIS is a parameter describing the valence of the sound perception of a tone pair. It can be computed from interval content $r_{hh}(k)$ by the inner product:

$$DIS = \frac{1}{M} \sum_{k=0}^{M-1} w(k)\, r_{hh}(k). \quad (9)$$

The weights $w(k)$ contain the dissonance values for different intervals motivated by the music theory [9].

The basic triads major, minor, diminished, and augmented correspond to samples of $r_{hhh}(k,l)$ at specific lag values according to Table 3. These autocorrelation values can be used as a measure for the intensity of the basic chords contained in the pitch data.

Tension TEN parameterizes the phenomenon that triads containing neighbouring intervals of equivalent size are perceived as unresolved, and thus have negative valence. Among the basic chords, diminished and augmented triads are unresolved. So, the overall value for tension can be computed as the sum of the diminished part DIM and the augmented part AUG: $TEN = DIM + AUG$. For resolved triads, one can further distinguish between a major- or minor-like modality.

Modality MOD can be computed as the quotient of the major part MAJ and the minor part MIN: $MOD = \frac{MAJ}{MIN}$. All together we extract 7 pitch related high-level features called harmony parameters.

3-tone chord		k/L	l/L
major chord	MAJ	4 ST	7 ST
minor chord	MIN	3 ST	7 ST
diminished chord	DIM	3 ST	6 ST
augmented chord	AUG	4 ST	8 ST

Table 3: Basic triads and the corresponding lag values of the third-order autocorrelation function

5. Database and classification experiments

In order to test the relevance of the proposed high-level features beyond the low-level features, we perform three classification experiments. First we compare all low-level features with the combined feature set of low-level and high-level features. In two further experiments, we only use those low-level feature groups for comparison that are related to the high-level features. Thus, features of the subgroups energy and duration are ignored. Second, we study the relevance of spectral high-level features by comparing it with MFCC features. Finally, we study the performance of the harmony features in addition to standard low-level pitch features. The FAU AIBO database is used for this purpose. It contains the emotion-related states anger, emphatic, neutral, positive, and rest. All relevant information about this database can be found in [1].

In this paper, we performed a classification on turn level, because in this case the labels for the test set are available and we can indicate detailed classification results. The feature set selected by SFFS is optimized on the test set. An overview of all the features is given in Table 1. For all the classifications a GMM classifier with a variable number of Gaussians is used.

5.1. Feature selection

To select the best features out of the whole feature set, we use the sequential floating forward selection algorithm (SFFS). It is an iterative method to find a subset of features that is near the optimal one. It was first proposed in [10]. In each iteration, a new feature is added to the subset of selected features and afterwards the conditionally least significant features are excluded. This process is repeated until the final dimension is obtained. As selection criterion the unweighted classification rate is used.

5.2. Comparing low-level and high-level features

The performance of all low-level features is compared to that of both low-level and high-level features. Figure 3 shows the average classification rate of all five emotion-related classes when

using low-level features only and additional high-level features. In both cases, feature sets with an increasing number of features (up to 25) are selected by SFFS. As we can see, by using the combined feature set the unweighted average classification rate is up to 2% higher than by using low-level features only. So there is a principle gain in using high-level features. In the following two experiments we check where the gain is exactly coming from.

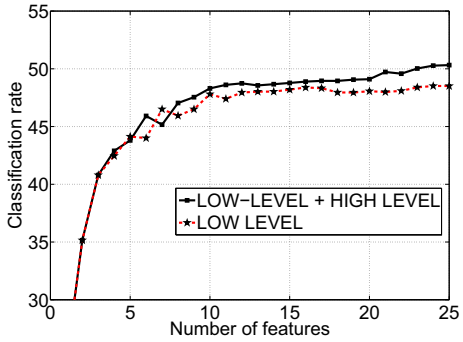

Figure 3: Comparison between all low-level and high-level features, 5 emotions classification

5.3. Comparing MFCC and VQP

Now the performance of spectral high-level features is compared to that of MFCC as standard low-level feature set. Figure 4 shows the average classification rate of the four emotion-related classes: anger, emphatic, neutral, and positive. Due to very high inhomogeneity, the rest class is excluded in this experiment. As we see, the classification rate by combining MFCC and spectral high-level features outperforms that of using only MFCC by far. In comparison to MFCC only, a gain of up to 10% is achieved. Clearly, the voice quality (VQP) and formant features (ART) supplement the MFCC for emotion recognition.

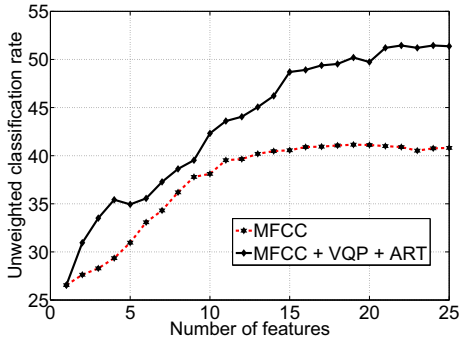

Figure 4: Comparison between spectral low-level and high-level features, 4 emotions classification

5.4. Comparing standard pitch and harmony parameters

For the group of pitch related high-level features the gain is marginal by performing the 4 class experiment above. However, the harmony features provide an additional gain in comparison to standard low-level pitch features when only classifying the emotions anger and positive. Figure 5 shows the unweighted average classification rate for this binary classification. In this experiment, we observe a gain of approximately 2% by adding harmony features to low-level pitch features. This is consistent to the idea that the harmony features can help in discriminating the evaluation dimension, but not the other dimensions of the psychological emotion dimension model [11]. Note, that this improvement is very interesting as it is by nature very difficult to distinguish the evaluation dimension.

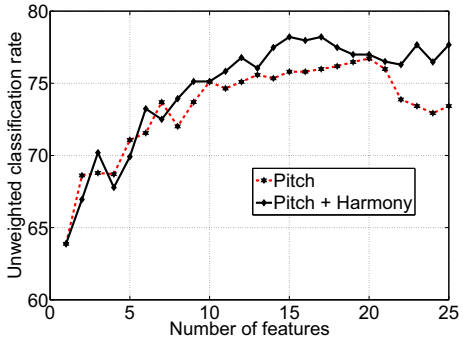

Figure 5: Comparison between pitch-related low-level and high-level features, anger vs. positive classification

6. Conclusions

In this paper we studied the relevance of high-level speech features for speaker independent emotion recognition of spontaneous speech. We proposed two groups of high-level features, a feature set based on the source-filter model of speech production called voice quality parameters and a feature set based on the harmony structure of music. We showed that using high-level features can improve the recognition performance of emotions even for spontaneous speech. Most of the gain is made up by the voice quality parameters for the emotions anger, emphatic, neutral, and positive. In addition, the harmony features lead to a further gain, for the discrimination between the classes anger and positive.

7. References

[1] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," *Interspeech, ISCA, Brighton, UK*, 2009.

[2] D. Talkin, W. Kleijn, and K. Paliwal, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis, Elsevier*, pp. 495–518, 1995.

[3] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *Technical Report, Bell Labs.*, 1987.

[4] M. Lugger, B. Yang, and W. Wokurek, "Robust estimation of voice quality parameters under real world disturbances," *IEEE ICASSP*, 2006.

[5] K. Stevens and H. Hanson, "Classification of glottal vibration from acoustic measurements," *Vocal Fold Physiology*, pp. 147–170, 1994.

[6] M. Lugger and B. Yang, "Classification of different speaking groups by means of voice quality parameters," *ITG-Sprach-Kommunikation*, 2006.

[7] L. Roberts, "Consonant judgements of musical chords by musicians and untrained listeners," *Acustica*, pp. 163–171, 1986.

[8] D. Cook and T. Fujidawa, "The psychophysics of harmony perception: harmony is a three-tone phenomenon," *Empirical musicology review*, vol. 1, no. 2, pp. 106–126, 2006.

[9] R. Plomp and J. M. Levelt, "Tonal consonance and critical bandwidth," *Journal of the Acoustical Society of America*, vol. 38, pp. 548–560, 1965.

[10] P. Pudil, F. Ferri, N. J., and J. Kittler, "Floating search method for feature selection with nonmonotonic criterion functions," *Pattern Recognition*, vol. 2, pp. 279–283, 1994.

[11] H. Schlosberg, "Three dimensions of emotions," *Psychological Review*, vol. 61, pp. 81–88, 1954.