CASCADED EMOTION CLASSIFICATION VIA PSYCHOLOGICAL EMOTION DIMENSIONS USING A LARGE SET OF VOICE QUALITY PARAMETERS

Marko Lugger and Bin Yang

Chair of System Theory and Signal Processing, University of Stuttgart, Germany
Marko.Lugger@Lss.uni-stuttgart.de

ABSTRACT

In this paper we improve the speaker independent emotion classification of a well known German database consisting of 6 basic emotions: sadness, boredom, neutral, anxiety, happiness, and anger. We achieve this by adding a large set of voice quality parameters to the standard prosodic features. In addition, our observation that the optimal feature set strongly depends on the emotions to be classified, leads to a 3-stage cascaded classification motivated by the psychological model of emotion dimensions. After activation recognition in the first stage, we classify the potency and evaluation dimension in the second and third stage, respectively. Compared to the 2-stage approach in [1], the average classification rate is improved by 14% to 88.8%.

Index Terms— Emotion recognition, Feature extraction, Cascaded classification, Psychological emotion dimensions

1. INTRODUCTION

There are many different applications of classifying paralinguistic properties of speech, e.g. gender, age, voice quality. The most known application is the detection of emotions from the recorded speech signal. Up to now, a good speaker independent recognition could only be achieved by using very large feature sets in combination with very complex classifiers [2], [3]. In [2], an average recognition rate of 86.7% was achieved for seven emotions by using 4000 features and support vector machine as classifier.

In this paper, we achieve comparable results by using only 333 features and a naive Bayesian classifier for the same German database. We improve the speaker independent emotion recognition by two measures. First, we propose an extended voice quality parameter set. It is an extension of the parameter set reported in [4]. We use spectral gradients which were first introduced by Stevens and Hanson [5]. Second, we use a psychological motivated cascaded classification strategy. It solves a 6-class problem in 3 stages consisting of 5 binary classifications. For each binary classification, the optimal feature set is selected separately.

The paper is organized as follows: First, the features are presented and our new voice quality parameter set is introduced in section 2. In section 3, its performance for emotion recognition is compared with the well known mel frequency cepstral coefficients. Moreover, results of the classification using a 2-stage respectively 3-stage cascaded classification are presented. Finally, some conclusions are drawn.

2. FEATURE GROUPS

In the field of emotion recognition, mainly suprasegmental prosodic features are used. Sometimes segmental spectral parameters as mel frequency ceptral coefficients (MFCC) are added. In our approach, the common prosodic features are combined with a large set of so called voice quality parameters (VQP). Their performance in speaker independent emotion classification is compared with that of MFCC parameters and their contribution in addition to the standard prosodic features is studied.

2.1. Prosodic features

There are three main classes of prosodic features: pitch, energy, and duration. Two more classes that do not belong directly to prosody are articulation (formants and bandwidths) and zero crossing rate. The features are obtained by measuring statistical values of corresponding extracted contours. Mean, median, minimum, maximum, range, and variance are the most used measurements. All together we extracted 201 prosodic features from the speech signal.

2.2. Mel frequency cepstral coefficients

The cepstrum of a signal is the inverse Fourier transform of the logarithm of the Fourier transform. In comparison to the standard cepstrum, MFCC uses frequency bands which are positioned logarithmically based on the mel scale motivated by the human auditory system. MFCC is a standard spectral parameter set in automatic speech recognition. According to [6], its performance is, however, not satisfying for emotion recognition. So we want to know whether VQP is a better alternative than MFCC for emotion recognition. For this comparison, we use in our study the mean value as well as the 2nd to the 5th central moments of 13 MFCC. The total number of MFCC features is thus 65. The implementation we use was first published in [7].

2.3. Voice quality parameters

In contrast to other spectral features, the voice quality parameters describe the properties of the glottal excitation. Phonation is one aspect besides articulation and prosody in generating emotional coloured speech. By inverse filtering, the influence of the vocal tract is compensated to a great extent. The feature set we use is a parameterization of the voice quality in the frequency domain by spectral gradients.

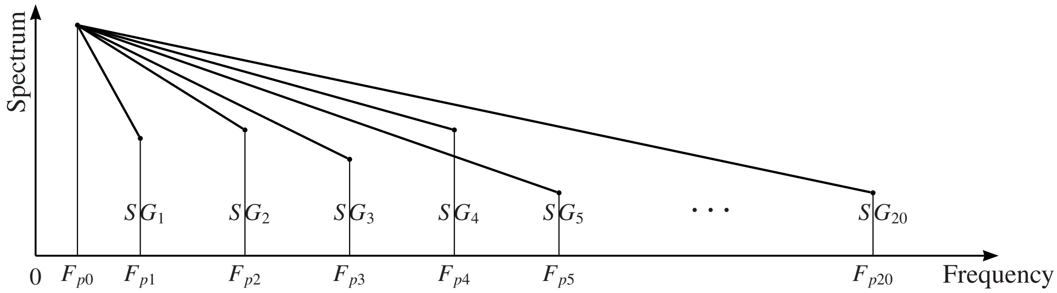

Fig. 1. Spectral gradients at fixed frequencies

2.3.1. Measurement of speech features

First, we estimate some well known speech features from windowed, voiced segments of the speech signal, see Table 1. We perform the voiced-unvoiced decision and the pitch estimation F_0 according to the RAPT algorithm [8] that looks for peaks in the normalized cross correlation function. We further extract higher harmonics F_{pk} next to fixed frequencies at multiples of 200 Hz. All together 21 harmonics are used covering 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 [9].

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

Table 1. Speech features for estimation of spectral gradients

2.3.2. Compensation of the vocal tract

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 $\tilde{H}_k$.

2.3.3. Estimation of the voice quality parameters

Up to now only 4 spectral gradients were used to characterize the glottal source signal. In order to better parameterize the glottal excitation signal the parameter set is extended to 20 gradients. Fig. 1 illustrates the definition of the spectral gradients.

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

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

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

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

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

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

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

The last three voice quality parameters describe the voicing, the harmonicity, and the periodicity of the signal, see [10]. In total, we obtain a set of 67 voice quality features.

3. CLASSIFICATION

In this section, the results of two classification studies are presented. First we show the relevance of our new voice quality parameter set by comparing it to MFCC. Then, two different multi-stage classification strategies are compared using all features including the voice quality parameters.

We classify six emotions: anger (ag), happiness (hp), sadness (sd), boredom (bd), anxiety (ax), and neutral (nt). We use short acted utterances (approximately between two and five seconds) from the Berlin emotional database [11]. For this speaker independent classification, a ”leaving-one-speaker-out” cross validation is used. There are 690 utterances. A naive Bayesian classifier for all classifications is used. That means, the class-conditional densities are modelled as unimodal Gaussians. By using the Gaussian mixture model with a variable number of Gaussians, we could not observe remarkable improvements for this database.

To find the optimal feature set, we use the sequential floating forward selection algorithm (SFFS) in all classifications. It is an iterative method to find the best subset of features [12]. For all classifications, the best 25 features are selected.

3.1. Voice quality parameters vs. MFCC

Fig. 2 compares the average classification rates of six emotions by using prosodic features only and by combining them with MFCC and/or VQP. In each of the four cases, a varying (up to 25) number of the best features are selected by SFFS. A flat 1-stage classification is used. As we see, the classification rate by using additional MFCC is higher than using only prosodic features. But the classification rate when combining prosodic features with SQP outperforms that when combining with MFCC. In comparison to prosodic features only, a gain of at least 3% is achieved. Adding both VQP and MFCC to prosodic features brings no noticeable improvement, so that the two upper curves are almost on top of each other.

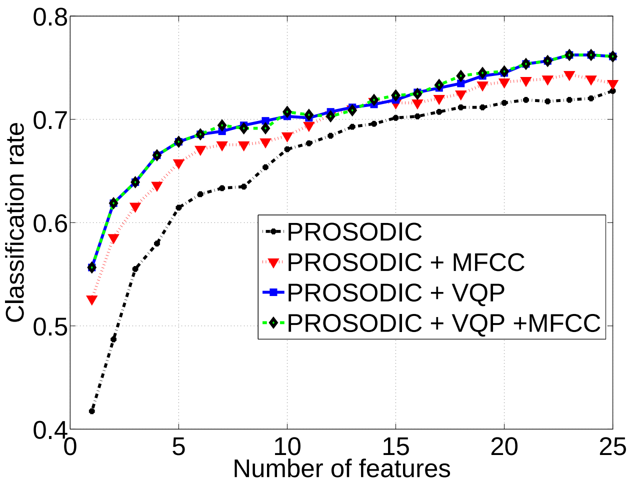

Fig. 2. Comparison between VQP and MFCC

3.2. Psychological motivated cascaded classification

Our former investigation in [13] showed that the optimal feature set strongly depends on the emotions to be separated. This means, using one global feature set for the discrimination of all emotions is clearly suboptimal. This conclusion motivates a cascaded classification strategy, consisting of different classification stages. The stages chosen in our approach are motivated by the 3 emotion dimensions of the psychological model shown in Fig. 3. Psychological research in this area states that we can locate different emotions in a two- or three-dimensional space [14].

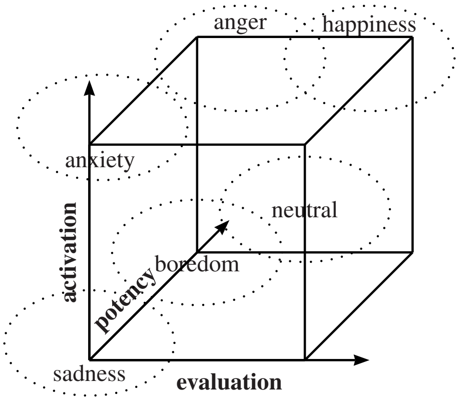

Fig. 3. Psychological emotion dimension model

3.2.1. 2-stage classification

The observation in [1] that prosodic features are very useful in discriminating different levels of activation and voice quality features perform better in discriminating other emotion dimensions leads to the 2-stage classification shown in Fig. 4.

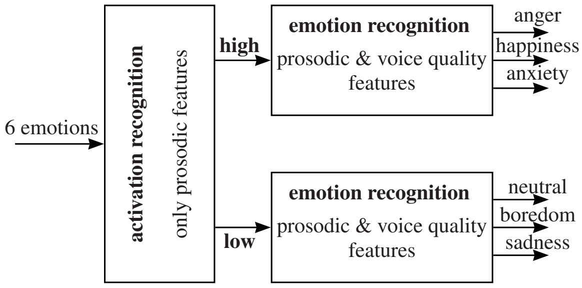

Fig. 4. 2-stage approach of emotion recognition

Important is the fact that all three subclassifications in Fig. 4 are trained separately by using different feature sets. Even the two feature sets used for emotion recognition with a high and a low activation level are different. In each case, the best 25 features were selected using SFFS. With this strategy and by using our new set of voice quality parameters, we achieved an overall classification rate of 83.5%. This is an improvement of 9% compared to [1]. The main contributions to this improvement are the new enlarged voice quality feature set as well as a larger number of features used. The confusion matrix is depicted in Table 2.

emotions	happy	bored	neutral	sad	angry	anxious
happy	**67.3%**	0.0%	0.0%	0.0%	21.5%	11.2%
bored	0.0 %	**91.0%**	2.7%	6.3%	0.0%	0.0%
neutral	0.0%	14.6%	**78.7%**	6.8%	0.0%	2.9%
sad	0.0%	9.1%	1.7%	**86.7%**	0.8%	1.7%
angry	5.1%	0.0%	0.0%	0.0%	**86.1%**	8.8%
anxious	5.3%	0.0%	0.9%	0.9%	3.5%	**89.4%**

Table 2. 2-stage cascaded classification result

3.2.2. 3-stage classification

Motivated by the psychological emotion model, we found out that one can further optimize the classification performance by using only binary subclassifications. That means we perform 5 classifications in 3 stages as shown in Fig. 5.

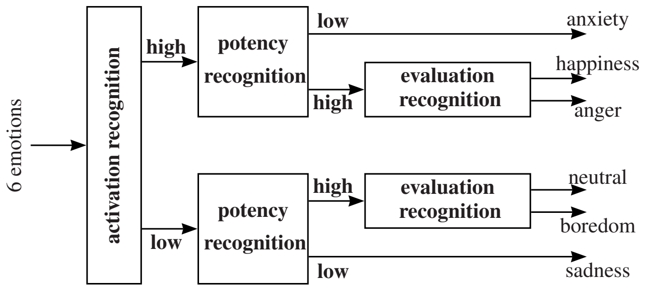

Fig. 5. 3-stage approach of emotion recognition

Every frame corresponds to one classification whose best 25 features are optimized by SFFS separately. In the first stage, we classify two different activation levels. One class including anger, happiness, and anxiety has a high activation level while the second class including neutral, boredom, and sadness has a low activation level. For this activation discrimination, we achieve a very good classification rate of 98.8% on average. Table 3 shows the confusion matrix.

activation	high	low
high	**99.1%**	0.9%
low	1.9%	**98.1%**

Table 3. Classification of 2 activation levels

In the second stage, we classify two potency levels in each activation class. That means, all patterns that were classified to high activation in the first stage are classified to one class containing happiness and anger or to a second class only containing anxiety. Similarly, all patterns that were classified to low activation in the first stage are classified to one class containing neutral and boredom or to sadness. Table 4 shows the classification results for the second stage.

	high activation		low activation	
potency	high	low (ax)	high	low (sa)
high	**98.8%**	1.2%	**97.6%**	2.4%
low	11.7%	**88.3%**	11.2%	**88.8%**

Table 4. Classification of 2 potency levels

In the third stage, we distinguish between the emotions that differ only in the evaluation dimension: happiness vs. anger as well as neutral vs. boredom. The confusion matrix for the third stage is shown in Table 5.

	high activation		low activation	
evaluation	high (hp)	low (ag)	high (nt)	low (bd)
high	**84.6 %**	15.4%	**91.8%**	8.2%
low	5.9%	**94.1 %**	2.8%	**97.2%**

Table 5. Classification of 2 evaluation levels

This 3-stage strategy leads to the overall confusion matrix shown in Table 6. It corresponds to an overall recognition rate of 88.8%. This is an additional improvement of more than 5% in comparison to the two-stage approach in Table 2. In particular, the recognition rate of happiness is improved by nearly 15%.

emotions	**happy**	**bored**	**neutral**	**sad**	**angry**	**anxious**
happy	**82.2%**	0.0%	0.0%	0.0%	15.0%	2.8%
bored	0.0 %	**94.6%**	2.7%	2.7%	0.0%	0.0%
neutral	1.9%	7.8%	**87.4%**	1.9%	0.0%	1.0%
sad	0.0%	8.3%	2.5%	**86.7%**	0.0%	2.5%
angry	5.9%	0.0%	0.0%	0.0%	**94.1%**	0.0%
anxious	5.3%	0.9%	0.0%	0.9%	6.2%	**86.7%**

Table 6. 3-stage cascaded classification result

4. CONCLUSION

In this paper we proposed a large set of 67 voice quality parameters. They outperform the well known mel frequency cepstral coefficients in speaker independent emotion recognition. We showed that they improve the well known prosodic features using a 1-stage classification by about 3%. Combining the new feature set with the 2-stage classification from [1] led to a recognition rate of 83.5%. Motivated by the psychological emotion dimension model, we used our new feature set in a 3-stage cascaded classification approach. For every subclassification, we applied the SFFS algorithm to find the best 25 features. With this approach, we could further increase the overall recognition rate to 88.8%. Thus, the total improvement in comparison to [1] is 14%.

5. REFERENCES

[1] Marko Lugger and Bin Yang, “The relevance of voice quality features in speaker independent emotion recognition,” *ICASSP, Hawaii, USA*, 2007.

[2] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, “Emotion recognition in the noise applying large acoustic feature sets,” *Speech Prosody, Dresden*, 2006.

[3] Chul Min Lee and S. Narayanan, “Toward detecting emotions in spoken dialogs,” *Transaction on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

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

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

[6] T. Nwe, S. Foo, and L. De Silva, “Speech emotion recognition using hidden Markov models,” *Speech communication*, vol. 41, pp. 603–623, 2003.

[7] Orsak et. al., “Collaborative SP education using the internet and matlab,” *IEEE Signal processing magazine*, vol. 12, no. 6, pp. 23–32, 1995.

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

[9] David Talkin, “Speech formant trajectory estimation using dynamic programming with modulated transition costs,” *Technical Report, Bell Labs.*, 1987.

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

[11] Burkhardt et. al., “A database of German emotional speech,” *Proceedings of Interspeech*, 2005.

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

[13] Marko Lugger and Bin Yang, “An incremental analysis of different feature groups in speaker independent emotion recognition,” *ICPhS, Saarbrücken*, 2007.

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

[15] Diane J. Litman and Kate Forbes-Riley, “Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors,” *Speech communication*, vol. 48, no. 5, pp. 559–590, 2006.