IEEE copyright notice

THE RELEVANCE OF VOICE QUALITY FEATURES IN SPEAKER INDEPENDENT EMOTION RECOGNITION

Marko Lugger and Bin Yang

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

ABSTRACT

This paper investigates the classification of different emotional states using prosodic and voice quality information. We want to exploit the usage of different phonation types within the production of emotions. Therefore, as features we use prosodic features, voice quality parameters, and different combinations of both types. We study how prosodic and voice quality features overlap or complement each other in the application of emotion recognition. The classification is speaker independent and uses a reduced subset of 8 features and a Bayesian classifier.

Index Terms— Speech analysis, Feature extraction, Pattern classification

1. INTRODUCTION

There are many approaches to classify paralinguistic properties of speech in the literature. The most known application is the detection of emotions from the recorded speech signal. Various attempts show quite good results in the case of speaker dependent classification [1], [2], [3], [4]. But most of them fail in speaker independent emotion recognition. By using very large feature sets, some approaches achieve pretty good results even in speaker independent classification [5]. Speaker independent means that the speaker of the classified utterances is not included in the training database. He is unknown for the classifier and the deduced learning rules in the training phase.

Our goal is to improve the classification performance of the speaker independent emotion recognition by incorporating a new feature type. We try to achieve that by combining suprasegmental prosodic features with segmental spectral voice quality parameters. Those are features extracted from the glottal source signal, which describe the phonation type that is used during the production of voiced parts of the uttered speech. In addition, we want to study whether the information contained in prosodic and voice quality features is supplementing or overlapping.

The paper is organized as follows: First the psychological approach of emotion dimension is explained. Then, relevant acoustic features for the different dimensions are introduced in section 3. In section 4, the results for the classification of six emotions with different strategies for the combination of feature sets are presented. Finally, some conclusions are drawn.

This work was supported by a grant from the Ministry of Science, Research, and the Art of Baden-Wuerttemberg, Germany

2. DIMENSIONAL EMOTION APPROACH

Psychological research in the area of emotion production says that we can locate different emotions in a two- or three-dimensional space [6]. The most often used dimensions are activation, potency, and evaluation. As we will see below, most of the features used in acoustical emotion recognition, mainly prosodic features, describe the activation dimension. This is why emotions which do not obviously differ in the activation dimension can not be well separated in emotion recognition. They are, for example, anger, happiness, and anxiety with a high activation or neutral, boredom, and sadness with a low activation. So our task is to find acoustic features that describe more the other dimensions, e.g. the evaluation to distinguish between positive and negative emotions. Fig. 1 shows three different emotional dimensions and six basic emotions.

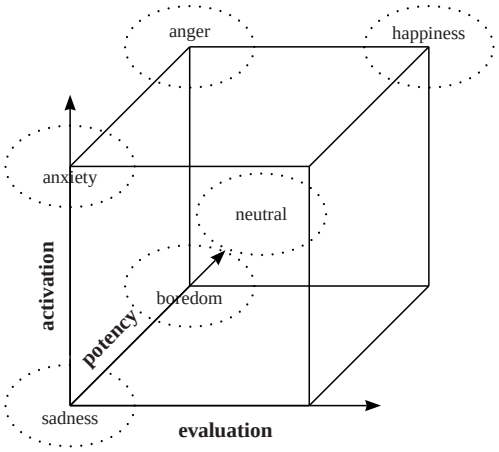

Fig. 1. Three-dimensional emotional space and 6 basic emotions

3. FEATURES

In the field of emotion recognition mainly suprasegmental prosodic features are used. Sometimes segmental spectral parameters as mel frequency ceptral coefficients (MFCC) are added. But according to [7], MFCC features achieve poor results and log-frequency power coefficients (LFPC) are better suited for emotion recognition. In our approach, the common prosodic features are combined with the so called voice quality parameters.

3.1. Prosodic features

There are three main classes of prosodic features: pitch, energy, and duration. A fourth class that does not belong directly to prosody is articulation (formants and bandwidths). 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 over 200 prosodic features from the speech signal.

3.2. Voice quality parameters

In contrast to other spectral features, the voice quality parameters (VQP) describe the properties of the glottal source. By inverse filtering, the influence of the vocal tract is compensated to a great content. Parameter values which describe the kind of phonation type are used. Phonation is one aspect besides articulation and prosody in generating emotional coloured speech. The feature set we use is a parameterization of the voice quality in the frequency domain by spectral gradients. The definition and the robustness of VQP are reported in [8]. All together there are 8 voice quality parameters. As we can see later in section 4, the VQP parameters have an obvious contribution to the discrimination of different emotions beyond the prosodic features.

3.3. Feature selection

There are two main reasons for reducing the number of features from the original set. First, the number of training patterns had to be enormous if we want to use all features. Second, the training and classification would take a long time when using the whole feature set. So the original number of over 200 prosodic and eight voice quality features is reduced by using an iterative selection algorithm. The final number of features used is eight. We used the sequential floating forward selection algorithm (SFFS). It is an iterative method to find the best subset of features. It was first proposed in [9]. In each iteration, a new feature is added to the subset of selected features and afterwards the conditionally least significant feature is excluded.

4. CLASSIFICATION

In this section, the results of different classification strategies are presented and the relationship between prosodic and voice quality features is discussed. We try to classify six emotions: anger, happiness, sadness, boredom, anxiety, and neutral. We used short acted utterances (approximately between two and five seconds) from the Berlin emotional database [10]. For this speaker independent classification, a "leaving-one-speaker-out" cross validation was used. There are 694 utterances, that means over 100 patterns per emotion.

A Bayesian classifier for all classifications is used. The class-conditional densities are modeled as unimodal Gaussians. By using the Gaussian mixture model (GMM) with a variable number of Gaussians, we could not observe significant changes in the classification rate, because mostly only one Gaussian per emotion was decided.

We compared the classification rates of using prosodic features only, voice quality parameters only, and combinations of both feature types. Especially the gain of adding voice quality information to the prosodic information is investigated. In the following, different strategies to exploit the information contained in both feature types are presented.

4.1. Classification with prosodic features

First of all we classify with prosodic features. Over 200 features were reduced to eight by using SFFS. The confusion matrix is shown in Table 1. The overall recognition rate with 66.7% is quite good, but mainly the discrimination between anger and happiness is bad. Happiness is least classified with a recognition rate of only 48.2%. Furthermore, the confusions between angry and anxious and between neutral with bored are noticeable. As we know from Fig. 1, these emotions do

Emotion	happy	bored	neutral	sad	angry	anxious
happy	**48.2%**	0.9%	5.5%	0.9%	*34.5%*	10.0%
bored	1.8%	**68.2%**	18.2%	8.2%	1.8%	1.8%
neutral	5.8%	13.6%	**62.1%**	10.7%	0.0%	7.8%
sad	0.9%	5.0%	9.2%	**77.3%**	0.0%	7.6%
angry	*18.5%*	0.0%	0.7%	0.0%	**67.6%**	13.2%
anxious	7.8%	0.0%	1.7%	3.4%	12.1%	**75.0%**

Table 1. Classification with prosodic features only

not differ in the activation dimension and so prosodic features can not adequately distinguish between them. On the other hand, sad is classified best with 77.3%. In this database, sadness is spoken very slowly and also with long pauses. Hence, duration features work very well to recognize sad utterances.

4.2. Classification with voice quality parameters

Now we test the feasibility of voice quality parameters as features for the emotion classification. We used all eight parameters from [1] and got the worse results shown in Table 2.

Emotion	happy	bored	neutral	sad	angry	anxious
happy	**41.8%**	0.9%	4.6%	1.8%	30.0%	20.9%
bored	0.9%	**56.4%**	17.3%	16.3%	0.0%	9.1%
neutral	1.0%	22.3%	**53.4%**	11.7%	0.0%	11.6%
sad	0.9%	17.6%	5.9%	**70.6%**	0.0%	5.0%
angry	19.1%	0.7%	0.0%	0.0%	**76.5%**	3.7%
anxious	12.1%	2.6%	3.4%	2.6%	9.5%	**69.8%**

Table 2. Classification with voice quality parameters only

But the classification result is very good for the three emotions: anger, sadness, and anxiety which are most coloured with nonmodal voice qualities. The corresponding confusion matrix for these three emotions is depicted in Table 3. The

Emotion	sad	angry	anxious
sad	**95.8%**	0.0%	4.2%
angry	0.0%	**94.1%**	5.9%
anxious	6.0%	12.1%	**81.9%**

Table 3. Voice quality parameters only for 3 classes

reason for this good result is that these three emotions differ

considerably in the phonation type. For the production of a sad emotional state, a creaky phonation is often used. Rough voice is usually used to support an angry emotional state. The anxious emotion shows sometimes parts of breathy voice. For the recognition of these different voice quality classes defined by J. Laver [11], the voice quality parameters are predestinated. Very good speaker dependent and independent recognition rates for voice qualities have been reported in [1]. Two questions arise that we would like to answer in the sequel: Do the voice quality parameters contain some new information that is not included in the prosodic features? And how can we combine both feature types to get the best classification result?

4.3. Classification with combined feature sets

Below we classify with both prosodic and voice quality features. The first result shown in Table 4 is the classification rate we would obtain by an ideal combination of a prosodic and a voice quality classifier. The prosodic classifier uses the best eight prosodic features selected by SFFS and the voice quality classifier uses all eight voice quality parameters. P stands for the event "correctly classified by prosodic features" and V stands for the event "correctly classified by voice quality parameters". The second row in Table 4 shows the rate of patterns that are classified correctly by both classifiers. For the emotions happiness, boredom, and neutral the correctly classified patterns are quite disjoint, while for sadness, anger, and anxiety they are strongly overlapping. The third and fourth row show the patterns that are correctly classified by prosodic features but not by voice quality parameters and vice versa. In general, the prosodic classifier performs better. But for the emotions happiness, neutral, and anger the voice quality classifier contributes to an improvement between 14.6% and 21.8%. For anger, the voice quality classifier even outperforms the prosodic one. In the last row of Table 4 the overall classification rate for P OR V is given. This implies that we would have complete knowledge of which classifier performs correctly for every single given pattern. We only get a misclassification when both classifiers are wrong. One can interpret this as a reference value for the classification rate with both feature sets. It corresponds to an overall recognition rate of 80.2% that is at the level of human recognition rate. Clearly, the voice quality features improve considerably the classification beyond the prosodic information. The gain is biggest for the problematic classes happiness and anger.

Emotion	happy	bored	neutral	sad	angry	anxious
P AND V	20.0%	44.6%	38.8%	59.7%	58.8%	61.2%
P AND $\overline{V}$	28.2%	23.6%	23.3%	17.6%	8.8%	13.8%
$\overline{P}$ AND V	21.8%	11.8%	14.6%	10.9%	17.7%	8.6%
P OR V	**70.0%**	**80.0%**	**76.7%**	**88.2%**	**85.3%**	**83.6%**

Table 4. Reference value for the classification rate of a combined classifier with prosodic and voice quality features

4.3.1. Seperate SFFS-based combination

The best four prosodic and the best four voice quality features separately selected by SFFS are used for classification. Table 5 shows the resulting confusion matrix. We observe that the fusion of both feature classes improves slightly the classification rate. The average overall recognition rate is 71.0%.

Emotion	**happy**	bored	neutral	sad	angry	anxious
happy	**51.8%**	0.9%	5.5%	0.0%	28.2%	13.6%
bored	1.8%	**74.6%**	17.3%	4.5%	0.0%	1.8%
neutral	3.8%	20.5%	**68.0%**	5.8%	0.0%	1.9%
sad	0.0%	9.2%	3.4%	**82.4%**	0.0%	5.0%
angry	20.6%	0.7%	0.0%	0.0%	**75.8%**	2.9%
anxious	10.4%	0.8%	8.6%	0.8%	7.8%	**71.6%**

Table 5. Classification with four prosodic and four voice quality features which are separately selected by SFFS

4.3.2. Joint SFFS-based combination

In Table 6, the best eight features out of all (prosodic and voice quality features) are jointly selected by SFFS. Among them, there are six prosodic and two voice quality features. With the overall recognition rate of 72.8%, this approach outperforms the result of Table 5. The reason that anger is worse classified is that less voice quality information is used in this approach.

Emotion	happy	bored	neutral	sad	angry	anxious
happy	**55.5%**	0.0%	10.0%	0.0%	21.8%	12.7%
bored	2.7%	**77.3%**	10.0%	7.3%	0.9%	1.8%
neutral	2.9%	15.5%	**71.0%**	1.9%	2.9%	5.8%
sad	0.9%	4.2%	4.2%	**84.0%**	1.7%	5.0%
angry	25.1%	0.0%	1.4%	0.0%	**66.9%**	6.6%
anxious	6.9%	0.0%	1.7%	2.6%	6.9%	**81.9%**

Table 6. Classification with six prosodic and two voice quality features jointly selected by SFFS

4.3.3. Cascaded classification

The main drawback of the previous approaches is that we do not consider which type of features classifies better for which emotions. The fundamental observation that prosodic features are very powerful in discriminating different levels of activation and voice quality features perform better in discriminating the other emotion dimensions leads to the following cascaded strategy.

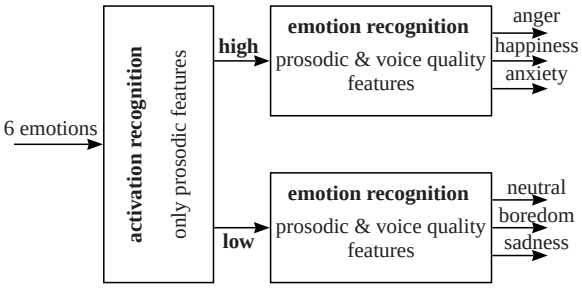

Fig. 2. Two-step approach of emotion recognition

As shown in Fig. 2, we separate the classification process in two steps. In the first step, we classify for two different

activation levels. One class including anger, happiness, and anxiety has a high activation level. 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 95.5% on average with eight prosodic features only. Table 7 shows the confusion matrix. Here, we have seen that including voice quality features will not contribute to any improvements.

Activation	high	low
high	**96.1%**	3.9%
low	5.1%	**94.9%**

Table 7. Classification of two activation levels

In the second step, we classify in each activation class the real emotions. That means, all patterns that were classified to high activation in the first step are classified to anger, happiness, and anxiety. Similarly, all patterns that were decided to have a low activation in the first step were classified to neutral, boredom, and sadness. For this second step, the joint SFFS-based combination of prosodic and voice quality features was used. Here, there is a clear advantage of including voice quality features. Table 8 and Table 9 show the classification results for the second step.

Emotion	happy	angry	anxious
happy	**57.5%**	30.2%	12.3%
angry	13.4%	**79.9%**	6.7%
anxious	7.4%	8.3%	**84.3%**

Table 8. Classification of emotions with high activation

Emotion	bored	neutral	sad
bored	**81.0%**	14.3%	4.7%
neutral	16.0%	**78.7%**	5.3%
sad	10.3%	6.0%	**83.7%**

Table 9. Classification of emotions with low activation

By combining both steps in Fig. 2, the overall confusion matrix is shown in Table 10. The average recognition rate is 74.5% and outperforms the combination strategies in 4.3.1 and 4.3.2.

5. CONCLUSION

We presented an approach of speaker independent emotion classification. We used the SFFS algorithm to reduce the feature number. We have also used the Fisher feature transform instead of SFFS feature selection. Though the Fisher transform generally performs better than SFFS for the same number of final features, no significant changes could be observed in this study. We showed that parameters of voice quality supply a contribution in addition to the well known prosodic features. They deliver information concerning the phonation type used by the speaker that is not contained in the

Emotion	happy	bored	neutral	sad	angry	anxious
happy	**55.5%**	0.0%	3.6%	0.0%	29.1%	11.8%
bored	3.6%	**77.3%**	13.6%	4.6%	0.0%	0.9%
neutral	1.0%	14.6%	**71.8%**	4.8%	1.0%	6.8%
sad	0.0%	10.1%	5.9%	**81.5%**	0.0%	2.5%
angry	13.3%	0.7%	0.7%	0.0%	**78.7%**	6.6%
anxious	6.9%	1.7%	4.3%	0.9%	7.8%	**78.4%**

Table 10. Cascaded classification

prosodic features. An intelligent combination of the discriminating power of prosody and voice quality yields in an improved classification performance. In our two-step approach, we could raise the average recognition rate from 66.7% to 74.5%. This improvement could be even larger by using emotional databases that make more use of different voice qualities in the production of emotions.

6. ACKNOWLEDGEMENTS

We would like to thank Stefan Uhlich and Fabian Friedrichs for their support in implementation and evaluation.

7. REFERENCES

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

[2] S. McGilloway, R. Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, "Approaching automatic recognition of emotion from voice: A rough benchmark," *ISCA Workshop Speech and Emotion*, pp. 737–740, 2000.

[3] A. Nogueiras, A. Moreno, A. Bonafonte, and J. Marino, "Speech emotion recognition using hidden Markov models," *Eurospeech 2001*, pp. 2679–2682, 2001.

[4] 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.

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

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

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

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

[9] 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.

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," *Proceedings of Interspeech*, 2005.

[11] John Laver, *The phonetic description of voice quality*, Cambridge University Press, 1980.