AN INCREMENTAL ANALYSIS OF DIFFERENT FEATURE GROUPS 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 speech features from different feature groups. We use both suprasegmental feature groups like pitch, energy, and duration and segmental feature groups like voice quality, zero crossing rate, and articulation. We want to exploit the selection of the most relevant features from these different feature groups to get a better understanding of the speaker independent emotion recognition. We study how these different feature groups overlap or complement each other. By using the sequential floating forward selection algorithm (SFFS), feature subsets maximizing the classification rate will be generated. For this purpose, we use a Bayesian classifier and a speaker independent cross validation. A detailed study is also done on the relevance of the feature groups for classifying different emotion dimensions known from the psychological emotion research.

1. Introduction

There are many applications of paralinguistic properties in the literature. One well known application is the detection of emotions from the recorded speech signal [1]. Various attempts show quite good results in the case of speaker dependent classification [2], [3]. By generating an emotion model for every speaker in the database, the utterances of a speaker are classified by using his own emotion model. But the performance of many approaches is poor in the case of speaker independent classification. 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 this paper we follow two goals. One is to improve the classification performance of speaker independent emotion recognition by combining conventional prosodic features with additional segmental features. The second one is to get a better understanding of emotion discrimination by analyzing the role and the interaction of different feature groups in emotion recognition.

The paper is organized as follows: The acoustic features extracted for the emotion detection and their grouping are described in the 2. section. In the 3. section, different strategies for an optimum feature composition are discussed and corresponding results of emotion classification are presented. Results for classifying the psychological emotion dimensions can be found in the 4. section. Finally, some conclusions are drawn.

2. Features

In emotion recognition, mainly suprasegmental prosodic features have been used up to now. In our approach, the common prosodic features are combined with additional segmental features describing the zero crossing rate, articulation, and the voice quality. We extracted a total number of 216 speech features belonging to different feature groups. Below these feature groups are briefly described.

2.1. Suprasegmental features

There are three main classes of prosodic features: pitch, energy, and duration (dur). The features are obtained by measuring statistical values that describe the corresponding feature contours. Mean, median, minimum, maximum, range, variance, and the first two derivatives are calculated from the contours. Pitch features are extracted from the intonation contour of a spoken utterance. The raw value of the pitch is calculated for each analysis segment by using the RAPT algorithm which uses normalized cross correlation and dynamic programming. Energy features are derived from the signal energy contour. Duration features count the number of uninterrupted analysis segments of the same voicing type. In all prosodic features, we distinguish between voiced, unvoiced and pause segments. All together there are 146 prosodic features.

2.2. Segmental features

In the family of segmental features we generated 17 features of the zero crossing rate (zcr), 38 features of the articulation (art), and 15 voice quality parameters (vqp). The zero crossing rate counts the number of zero crossings of the speech signal within one analysis segment. Articulation features consist of the frequencies and the bandwidths of the first 4 formants.

In contrast to articulation the voice quality parameters describe the properties of the glottal source and can also be used to detect the phonation type [2]. By inverse filtering, the influence of the vocal tract is compensated to a great content. Phonation is one aspect besides articulation and prosody in generating emotional coloured speech. The feature set we used is a parameterization of the voice quality in the frequency domain by spectral gradients. Its definition and robustness are reported in [4].

2.3. Feature selection

There are three main reasons for reducing the number of features. First, due to the "curse of dimension-

ality" the number of training patterns would have to be enormous if we used all features. Second, the training and classification would take a long time if we used the whole feature set. Third, we will have a better insight into the classification process if we reduce to only the best features.

So the original number of 216 features is reduced by using an iterative selection algorithm. We used the sequential floating forward selection algorithm (SFFS) due to its relatively low computational complexity and good performance. It is an iterative method to find the best subset of features and was first proposed in [5]. In each iteration, a new feature is added to the subset of selected features and then the least significant features are excluded as long as the recognition rate of the classification problem still increases.

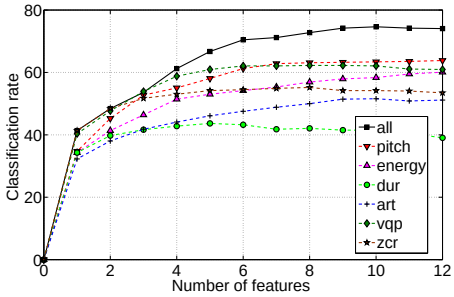
By analyzing the group membership of the selected features we will be able to judge the relevance of individual features and feature groups. We further analyze to what extent the different feature groups overlap or complement each other.

3. Classification

In this section, the results of different classification strategies are presented and the interaction between different feature groups is discussed. We classify six emotions: anger, happiness, sadness, boredom, anxiety, and neutral. We use short acted utterances (approximately between two and five seconds) from the Berlin emotional database [6]. There are 694 utterances, that means over 100 patterns per emotion. For this speaker independent classification, a "leaving-one-speaker-out" cross validation and a Bayesian classifier are used. The class-conditional densities are modelled 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 for this database.

3.1. Classification with all features

Figure 1: Classification with feature groups



The solid line in Fig. 1 shows the average classification rate (over all six emotions) when increasing the number of best features selected from all feature groups. Due to its local maximum at 10 features, we use the best 10 features for classification. The confusion matrix in Table 1 shows that sadness is classified best with 82.4% while happiness shows the worst recognition performance with only 61.8%. The average recognition rate is 74.6%. From Table 1 we observe two hardly distinguishable pairs of emotions: happiness vs. anger and neutral vs. boredom. They will be studied with a special focus in this paper.

Table 1: Classification with the best 10 features

emotion	happy	bored	neutral	sad	angry	anxious
happy	**61.8%**	0.9%	5.5%	0.0%	21.8%	10.0%
bored	0.9%	**75.5%**	10.9%	9.1%	0.9%	2.7%
neutral	3.9%	18.4%	**71.8%**	1.9%	0.0%	3.8%
sad	0.0%	7.6%	4.2%	**82.4%**	0.8%	5.0%
angry	17.6%	0.0%	0.7%	0.0%	**75.7%**	5.9%
anxious	13.8%	0.0%	3.4%	0.0%	3.4%	**79.3%**

3.2. Classification with a single feature group

Now we study the relevance of different feature groups in emotion recognition. Only the best features of the individual feature groups are used for classification. As we see from Fig. 1, the pitch class has the highest relevance. Voice quality and energy features are also quite relevant. Features of articulation and duration seem to be less relevant. But as we will see later, this does not mean that these feature groups do not contribute to the emotion recognition.

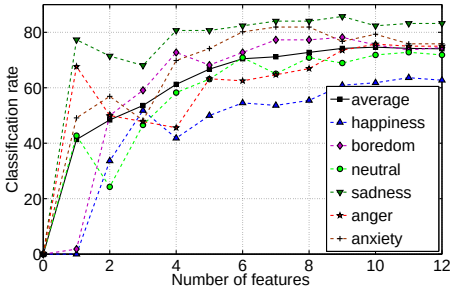
3.3. Incremental classification

Here we study how the recognition rate changes by using more and more features. Two different incremental classification strategies are possible: feature incremental and feature group incremental classification.

3.3.1. Feature incremental classification

Feature incremental means to increase the number of involved features for classification. In this case, we use SFFS to select the best features from all feature groups. Fig. 2 shows both the average recognition rate and the recognition rate of the individual emotions. Interestingly, no emotion shows a monotonically increasing recognition rate. Only for 4 of the added features the recognition rates of happiness and anger increase simultaneously. The same is valid for the emotions boredom and neutral. This is a first indication that good features to achieve a high average classification rate are not always appropriate for the discrimination of specific emotion pairs.

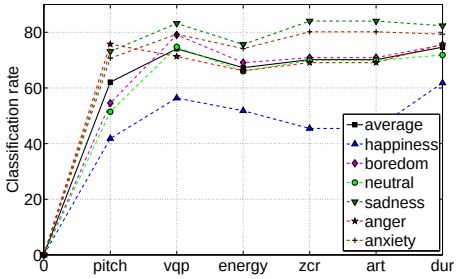
Figure 2: Feature incremental classification



3.3.2. *Feature group incremental classification*

Feature group incremental means to add a whole feature group to the existing ones. At each step, the SFFS selects a fixed number of 10 best features from an increasing number of feature groups. The order of added feature groups corresponds to their performance in Fig. 1. By using features from the pitch group only, we achieve an average classification rate of 62.1%. By using both pitch and voice quality features, we reach 74.1%, see Fig. 3. There are two reasons for the decreasing recognition rate when adding the energy group. First SFFS algorithm runs into a dead end and can not find the optimum feature set. The second reason is that pitch and energy features strongly overlap. Interestingly, the least significant feature group duration in Fig. 1 shows a higher improvement of the average recognition rate than energy, zero crossing rate, and articulation. This phenomenon indicates that representation is not equal discrimination. Though the duration feature group alone is not able to represent any emotional state, it seems to contain a few discriminative features which are not covered by the other feature groups.

Figure 3: Feature group increment. classification



3.4. Best feature sets

Table 2 shows the first 10 best features for the overall classification, together with their group membership and incremental recognition rate. The features are selected by SFFS from all feature groups. Clearly, pitch and voice quality features dominate. In the case of pitch, mainly measurements of statistical dispersion like standard deviation and interquartile range are relevant and not the mean values.

Table 2: Best 10 features for overall classification

No.	Feature	Group	Rate
1	EnMShimmer	energy	41.4%
2	PitchVStd	pitch	48.4%
3	VqpVoicingRatio	vqp	53.6%
4	VqpErr3	vqp	61.2%
5	PitchMRaise	pitch	66.7%
6	VqpJitter	vqp	70.5%
7	ZcrDiffIqr	zcr	71.2%
8	PitchVDiffIqr	pitch	72.8%
9	DurVStd	duration	74.2%
10	EnAReg	energy	74.6%

Table 3 and 4 show the first 10 best features to discriminate between happiness and anger as well as boredom and neutral, respectively. We see that the feature sets are totally different. Energy and articulation features dominate in Table 3. The articulation features there describe the statistical dispersion of the formant contours and are called articulation accuracy. In comparison, the discrimination between boredom and neutral is mainly based on energy features in Table 4.

The conclusion is that different emotional states are represented by different features and even different feature groups. This makes the design of a high-precision multi-emotion recognizer particularly difficult.

Table 3: Best 10 features for happiness vs. anger

No.	Feature	Group	Rate
1	ZcrMedian	zcr	73.6%
2	Form2FreqMin	art	74.8%
3	VqpSKG	vqp	75.2%
4	PitchVLast	pitch	75.3%
5	EnAFall	energy	76.8%
6	EnVDiffMean	energy	78.9%
7	Form1FreqIqr	art	78.5%
8	EnMDiffStd	energy	80.0%
9	EnV2S	energy	80.0%
10	Form2FreqMax	art	80.9%

Table 4: Best 10 features for boredom vs. neutral

No.	Feature	Group	Rate
1	EnMShimmer	energy	76.1%
2	EnADurFallMean	energy	81.7%
3	DurAud2Tot	duration	83.6%
4	ZcrMax	zcr	86.4%
5	PitchVDiffMean	pitch	87.3%
6	EnARegNeg	energy	89.2%
7	EnMSpec3	energy	89.7%
8	PitchVDiff2Iqr	pitch	90.1%
9	VqpVoicingRatio	vqp	91.0%
10	EnADurFallMedian	energy	90.6%

4. Classification of emotion dimensions

Psychological research in the area of emotion production proposes to locate the different emotions in a two- or three-dimensional space [7]. The most common dimensions are activation, evaluation, and potency. On the other side, 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 like happiness and anger can not be well separated. We now classify different psychological emotion dimensions instead of emotional states. Fig. 4 shows the 3-dimensional approach we used for classification. Once again we want to find the most important feature groups by using the feature group incremental classification. In

Figure 4: Three-dimensional psychological emotion space and 6 basic emotions

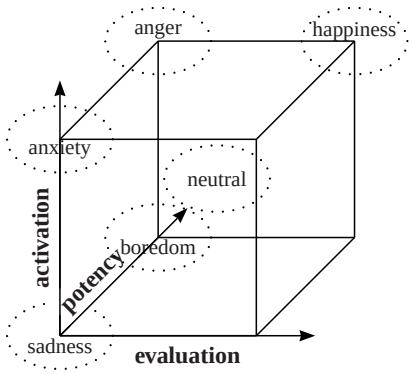
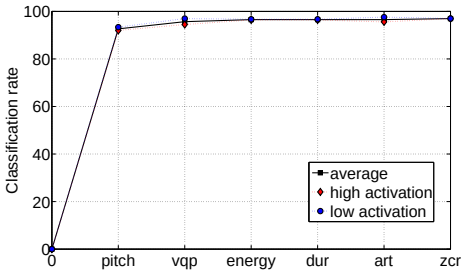


Table 5, 6, and 7, the solid line corresponds to the average recognition rate while the dashed lines correspond to high respectively low values of the corresponding emotion dimensions.

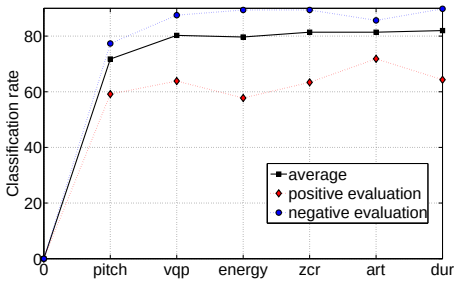
Emotion with a high activation level are anger, happiness, and anxiety. On the other hand, neutral, boredom, and sadness have low activation. As we see from Fig. 5, the activation can be nearly perfectly classified (95%) using only pitch features. There is also almost no difference between the recognition rate of high and low activation.

Figure 5: Classification of the activation



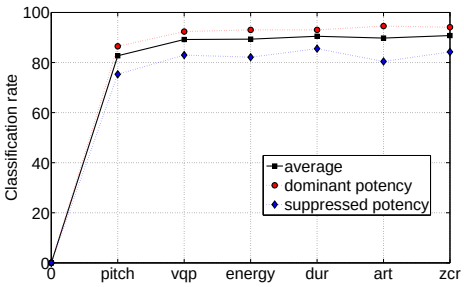
Neutral and happiness are positive emotions in comparison to the negative emotions anger, sadness, boredom, and anxiety. In this dimension we have to discriminate happiness from anger as well as boredom from neutral. Thus we achieve the worst recognition rate with only 82.0% for all feature groups. As we see from Fig. 6, only zero crossing rate features can support pitch and voice quality features.

Figure 6: Classification of the evaluation



Dominant emotion with a high level of potency are happiness, anger, and boredom. The emotions sadness, anxiety, and neutral show low levels of potency. Fig. 7 shows that voice quality parameters improve the classification of potency from about 82% to 89%. The other feature groups do not improve the classification rate.

Figure 7: Classification of the potency



5. Conclusion

We analyzed the features and feature groups for the speaker independent classification of 6 emotions. We used two different incremental classification strategies to find both the most relevant features and feature groups. In summary, the selection of the best features for classification strongly depends on the emotions to be classified. Pitch and voice quality are in average the most relevant feature groups. To distinguish between happiness and anger, both energy and articulation are relevant. The discrimination between boredom and neutral mainly relies on energy features. It is interesting that in all cases features of the statistical dispersion are more relevant than mean values. We also classified different psychological emotion dimensions. Activation can almost be perfectly classified by using only pitch features. The evaluation dimension is classified worst.

Finally, we mention that a better understanding of the relevant features and feature groups could help us to design a cascaded emotion recognizer that improves both the average classification and that of specific emotion pairs.

6. REFERENCES

[1] A. Batliner, S. Steidl, and B. Schuller, "Combining efforts for improving automatic classification of emotional user states language technologies," *IS-LTC*, 2006.

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

[3] Nogueiras et . al., "Speech emotion recognition using hidden Markov models," *Eurospeech 2001*, pp. 2679–2682, 2001.

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

[5] 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.

[6] Burkhardt et. al., "A database of German emotional speech," *Proceedings of Interspeech*, 2005.

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