Recognition of Emotions in German Speech Using Gaussian Mixture Models

Martin Vondra and Robert Vích

Institute of Photonics and Electronics, Academy of Sciences of the Czech Republic
Chaberská 57, CZ 18251 Prague 8, Czech Republic
{vich,vondra}@ufe.cz

Abstract. The contribution describes experiments with recognition of emotions in German speech signal based on the same principle as recognition of speakers. The most robust algorithm for speaker recognition is based on Gaussian Mixture Models (GMM). We examine three parameter sets: the first contains suprasegmental features, in the second are segmental features and the last is a combination of the two previous parameter sets. Further we want to explore the dependency of the classification accuracy on the number of GMM model components. The aim of this contribution is a recommendation for the number of GMM components and the optimal selection of speech parameters for emotion recognition in German speech.

Keywords: speech emotions, emotion recognition, Gaussian mixture models.

1 Introduction

Automatic dialogue systems with automatic speech recognition and speech synthesis suffer particularly from lack of emotional intelligence. It seems to be very important to incorporate for more natural interaction of these systems also the recognition of emotions and emotional speech synthesis.

Emotions in human speech are conveyed by complex intentional and unintentional changes of basic speech patterns given primarily by suprasegmental speech characteristics. These parameters are influenced e.g. by the individuality of the speaker, by the language etc.

At present there are no established methods for emotion recognition. In this paper we use a statistical classifier based on Gaussian Mixture Models (GMM) [1], which are commonly used for speaker recognition. Because we don't know which speech parameters characterize emotion in speech, we formulate three sets of speech parameters for the GMM classifier and evaluate the recognition score. Another issue is the optimal number of GMM components for robust emotion classification. For this reason we test also the dependency of the number of GMM components on the recognition score.

In Section 2 the configuration of the experiments is presented. The used emotional speech database, the speech parameter sets, and GMMs are described. The Section 3 describes the results of the experiments, which are represented by the percentage of correctly classified emotional sentences and by the confusion matrices. Section 4 summarizes the achieved results.

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 256–263, 2009.

2 Configuration of Experiment

2.1 Speech Database

For this experiment the German emotional speech database [2] was used. It contains 7 simulated emotions (anger, boredom, disgust, fear, joy, neutral, and sadness), simulated by 10 actors (5 male, 5 female) and 10 sentences (5 short, 5 longer). The complete database was evaluated in a perception test by the authors of the database. Utterances, for which the emotion with which they were spoken, were recognized better than 80% and judged as natural by more than 60% of the listeners, were used for the experiment. The speech material used for our experiment contains 535 utterances. The structure of the used speech material is summarized in Table 1.

Table 1. Structure of the German emotional speech database

Emotions	Number of utterances	Total time [min]
Anger	127	5.59
Boredom	81	3.75
Disgust	46	2.57
Fear	69	2.57
Joy	71	3.01
Neutral	79	3.11
Sadness	62	4.19

The test was performed as cross validated and speaker independent. Configuration of training and testing sets are obvious from Table 2. Speakers are marked by numbers, which mean:

03 – male speaker,
08 – female speaker,
09 – female speaker,
10 – male speaker,
11 – male speaker,
12 – male speaker,
13 – female speaker,
14 – female speaker,
15 – male speaker,
16 – female speaker.

The results for all speakers and all sentences were accomplished in three iterations by changing the speaker groups for training and testing.

Table 2. Configuration of cross validated training and testing sets

Step		Emotion	Training set	Testing set
1	Speakers		10, 11, 12, 13, 14, 15, 16	03, 08, 09
	Number of utterances and total time	Anger	88 utt., 3.91 min.	39 utt., 1.68 min.
		Boredom	62 utt., 2.84 min.	19 utt., 0.91 min.
		Disgust	37 utt., 2.06 min.	9 utt., 0.51 min.
		Fear	58 utt., 2.14 min.	11 utt., 0.43 min.
		Joy	49 utt., 2.13 min.	22 utt., 0.88 min.
		Neutral	49 utt., 1.89 min.	30 utt., 1.22 min.
		Sadness	42 utt., 2.74 min.	20 utt., 1.45 min.
2	Speakers		03, 08, 09, 10, 14, 15, 16	11, 12, 13
	Number of utterances and total time	Anger	92 utt., 4.03 min.	35 utt., 1.56 min.
		Boredom	58 utt., 2.66 min.	23 utt., 1.09 min.
		Disgust	34 utt., 1.95 min.	12 utt., 0.62 min.
		Fear	46 utt., 1.73 min.	23 utt., 0.84 min.
		Joy	51 utt., 2.10 min.	20 utt., 0.91 min.
		Neutral	57 utt., 2.23 min.	22 utt., 0.88 min.
		Sadness	46 utt., 3.10 min.	16 utt., 1.09 min.
3	Speakers		03, 08, 09, 11, 12, 13	10, 14, 15, 16
	Number of utterances and total time	Anger	74 utt., 3.24 min.	53 utt., 2.35 min.
		Boredom	42 utt., 2.00 min.	39 utt., 1.75 min.
		Disgust	21 utt., 1.14 min.	25 utt., 1.43 min.
		Fear	34 utt., 1.27 min.	35 utt., 1.30 min.
		Joy	42 utt., 1.79 min.	29 utt., 1.22 min.
		Neutral	52 utt., 2.10 min.	27 utt., 1.01 min.
		Sadness	36 utt., 2.54 min.	26 utt., 1.65 min.

2.2 Speech Parameters

For our experiment we have chosen three parameter sets. The parameters were computed for short-time speech frames similarly as in speech recognition – one feature vector for each short-time frame. Frame length was set to 25ms with frame shift 10ms. The parameter sets for training of GMMs were obtained by concatenation of feature vectors from all training sentences. This parameterization for emotion recognition is somewhat different from the experiment described e.g. in [3], where global parameters from each utterance are used, e.g. the mean, maximum and minimum values of F0, the maximum steepness and dispersion of F0, etc. We believe that these characteristics are covered in our parameter sets and can be caught by the GMM model.

Our first parameter set contains only suprasegmental features – the fundamental frequency (F0) and the intensity contours. We use the mean subtracted natural logarithm of these parameters. The second parameter set contains only segmental features – 12 mel-frequency cepstral coefficients (MFCC). We do not use mean subtraction of

MFCC. The last parameter set contains the combination of the first and second parameter sets, i.e. the F0 and intensity contours and 12 MFCCs. In each parameter set also the delta and delta-delta coefficients of basic parameters were added.

For F0 estimation the wavesurfer [5] script was used. In unvoiced parts of speech F0 was replaced by piecewise cubic Hermite interpolation (see Matlab function "pchip"). For MFCC computation the Matlab voicebox toolbox [6] was used.

2.3 Gaussian Mixture Models

For recognition of emotions we use GMMs with full covariance matrix [1]. The GMM modeling was implemented in C programming language. For training of GMMs we used the Expectation Maximization (EM) algorithm with 10 iteration steps. The iteration stop based on the difference between the previous and actual probabilities has shown to be uneffective, because the magnitude value depends on the number of training data and we have found out that relative probability in the training has a very similar behavior for different input data. For several initial iterations the probability that GMM parameters belong to training data steeply grows up, then becomes less and after 8 to 10 iterations the probability reaches the limit value. For initialization Vector Quantization (VQ) – K-means algorithm was used.

3 Results

Based on test results a proper number of GMM components is the crucial factor for good recognition score, see Fig. 1. For our tests the optimal number of GMM components has to be selected for the used parameter set. It seems to be convenient to use a smaller number of GMM components (about 4 to 8) than it is used for recognition of speakers (32 and more). In Tables 3, 4, 5 and 6 the confusion matrices are shown, which represent the dependency of classification on the number of GMM components for the third parameter set containing F0, intensity and MFCC.

The values in the confusion matrices (Table 3, 4, 5 and 6) represent the percentage of marked emotional sentences by the GMM classifier as emotions in the top line versus the number of emotional sentences for emotions in the left column. The correctly classified emotions are on the main diagonal of the matrix and the others are the confusions. From the confusion matrices for different numbers of GMM components (Table 3, 4, 5 and 6) it can be seen that if the number of GMM components is growing, the recognition of anger rises, but the recognitions of other emotions fall down (particularly fear, disgust and joy). It is clear that for a greater number of GMM components more confusion occurs.

The dependency of the classification score on parameter sets, when the number of GMM components is constant, is depicted in Fig. 2.It can be seen that a greater parameter set contributes to better recognition score, but the dependency is not as strong as for the number of GMM components (compare with Fig. 1). A more detailed view offers the confusion matrices given in Tables 7, 8 and 9.

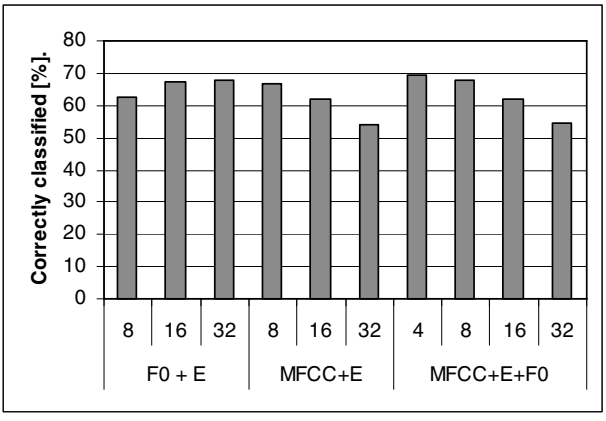

Fig. 1. Dependency of correctly classified emotions on the number of GMM components and used parameters

Table 3. Confusion matrix for 4 GMM components

	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Anger	**84.25**	0.00	1.57	0.79	12.60	0.79	0.00
Boredom	1.23	**75.31**	11.11	0.00	0.00	3.70	8.64
Disgust	4.35	0.00	**71.74**	0.00	10.87	8.70	4.35
Fear	13.04	1.45	4.35	**30.43**	26.09	15.94	8.70
Joy	28.17	0.00	5.63	1.41	**60.56**	4.23	0.00
Neutral	0.00	18.99	10.13	1.27	0.00	**65.82**	3.80
Sadness	0.00	8.06	0.00	3.23	0.00	0.00	**88.71**

Table 4. Confusion matrix for 8 GMM components

	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Anger	**94.49**	0.00	0.00	0.79	4.72	0.00	0.00
Boredom	2.47	**64.20**	11.11	0.00	0.00	17.28	4.94
Disgust	19.57	4.35	**56.52**	0.00	10.87	8.70	0.00
Fear	30.43	0.00	1.45	**28.99**	13.04	18.84	7.25
Joy	39.44	0.00	4.23	4.23	**49.30**	2.82	0.00
Neutral	1.27	10.13	11.39	0.00	0.00	**73.42**	3.80
Sadness	0.00	9.68	3.23	3.23	0.00	0.00	**83.87**

Table 5. Confusion matrix for 16 GMM components

	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Anger	**98.43**	0.00	0.00	0.00	1.57	0.00	0.00
Boredom	4.94	**62.96**	3.70	0.00	2.47	19.75	6.17
Disgust	36.96	0.00	**28.26**	0.00	15.22	15.22	4.35
Fear	43.48	0.00	0.00	**14.49**	14.49	18.84	8.70
Joy	56.34	0.00	0.00	1.41	**42.25**	0.00	0.00
Neutral	3.80	18.99	5.06	0.00	5.06	**64.56**	2.53
Sadness	0.00	9.68	0.00	4.84	0.00	0.00	**85.48**

Table 6. Confusion matrix for 32 GMM components

	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Anger	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00
Boredom	12.35	**54.32**	0.00	0.00	2.47	23.46	7.41
Disgust	54.35	0.00	**17.39**	0.00	10.87	13.04	4.35
Fear	59.42	0.00	0.00	**5.80**	11.59	14.49	8.70
Joy	80.28	0.00	0.00	1.41	**18.31**	0.00	0.00
Neutral	10.13	25.32	1.27	0.00	3.80	**55.70**	3.80
Sadness	0.00	11.29	0.00	3.23	0.00	0.00	**85.48**

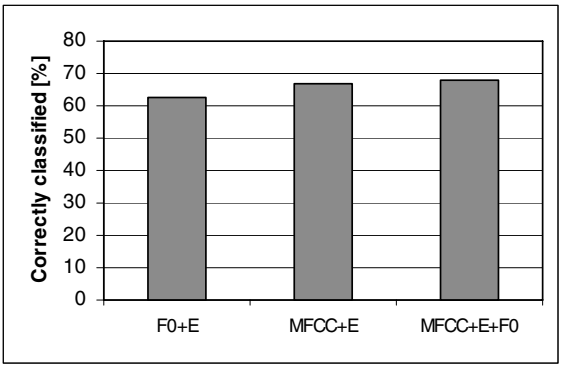

Fig. 2. Dependency of correctly classified emotions on parameter sets for 8 GMM components

Table 7. Confusion matrix for F0 and intensity parameter set and 8 GMM components

	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Anger	**55.12**	0.00	6.30	3.94	33.86	0.79	0.00
Boredom	0.00	**83.95**	3.70	1.23	1.23	6.17	3.70
Disgust	0.00	6.52	**50.00**	10.87	10.87	6.52	15.22
Fear	10.14	4.35	13.04	**47.83**	11.59	10.14	2.90
Joy	19.72	0.00	4.23	8.45	**67.61**	0.00	0.00
Neutral	1.27	15.19	8.86	13.92	5.06	**51.90**	3.80
Sadness	0.00	1.61	4.84	0.00	3.23	6.45	**83.87**

Table 8. Confusion matrix for MFCC parameter set and 8 GMM componets

	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Anger	**92.91**	0.00	0.00	0.00	7.09	0.00	0.00
Boredom	2.47	**66.67**	12.35	0.00	0.00	12.35	6.17
Disgust	17.39	4.35	**54.35**	2.17	10.87	8.70	2.17
Fear	27.54	2.90	1.45	**24.64**	18.84	17.39	7.25
Joy	39.44	0.00	2.82	5.63	**49.30**	2.82	0.00
Neutral	1.27	15.19	6.33	2.53	2.53	**69.62**	2.53
Sadness	0.00	9.68	0.00	4.84	0.00	0.00	**85.48**

Table 9. Confusion matrix for MFCC, F0 and intensity parameter set and 8 GMM components

	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Anger	**94.49**	0.00	0.00	0.79	4.72	0.00	0.00
Boredom	2.47	**64.20**	11.11	0.00	0.00	17.28	4.94
Disgust	19.57	4.35	**56.52**	0.00	10.87	8.70	0.00
Fear	30.43	0.00	1.45	**28.99**	13.04	18.84	7.25
Joy	39.44	0.00	4.23	4.23	**49.30**	2.82	0.00
Neutral	1.27	10.13	11.39	0.00	0.00	**73.42**	3.80
Sadness	0.00	9.68	3.23	3.23	0.00	0.00	**83.87**

4 Conclusion

If we summarize the results, we can state that the number of GMM components must be optimally adjusted for the used parameter set. The parameter set is not such an important factor, but the number of used parameters has a main impact on computational requirements. It seems to be convenient to use mainly suprasegmental parameters, because the correctly classified emotions based on these parameters only are quite high compared to the classification based only on segmental parameters. The best classification score was achieved for the combination of segmental and suprasegmental parameters and for the GMM model with 4 components.

The best recognized emotions were mostly anger and sadness, followed by boredom and neutral. More difficult were disgust and joy and the most difficult seems to be fear. Joy generates most confusion and is recognized as anger, and fear is recognized as anger. This can be influenced by the used speech database which is not balanced. A similar tendency of results was also achieved in classifications accomplished in [3], where the same emotional speech database was used. In [4] the database of emotional speech of Mandarin and Burmese speakers [7] with slightly different emotions was used and their results were for a different classification technique quite similar to ours. If we compare our results with results published in [3], we have achieved a slightly better recognition score for the optimal number of GMM components and for F0, intensity and MFCC parameter set. Some differences between our results and those in [3] can be found in confusions.

If we compare our results with the perception test made in [2] (where the recognition score of human listeners is above 80%) we can say that recognition of anger, sadness and in some cases of boredom approaches that of human listeners. For other emotions the described automatic recognition is much worse than that of human listeners.

Generally it can be said that the GMM classifier has a good differentiation property between emotions with high stimulation (anger, fear, disgust, joy) and between emotions with small stimulation (boredom, neutral, sadness), but shows poor differentiation between emotions with similar stimulation, like fear and anger. This can be given by very similar distributions of selected parameters, which can cause overlapping of GMM models. Therefore our future work will focus on finding parameters which would have appropriate distributions for better separation by the classifier.

Acknowledgement. This work was supported within the framework of COST2102 by the Ministry of Education, Youth and Sport of the Czech Republic, project number OC08010.

References

1. Reynolds, D.A.: Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17, 91–108 (1995)
2. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech. In: Proc. Interspeech 2005, Lisbon, Portugal, September 4-8 (2005)
3. Truong, K.P., Leeuven, D.A.: An 'open-set' detection evaluation methology for automatic emotion recognition in speech. In: ParaLing 2007: Workshop on Paralinguistic Speech - between models and data, Saarbrücken, Germany (2007)
4. Morrison, D., Wang, R., De Silva, L.C.: Ensemble methods for spoken emotion recognition in call-centers. Speech Communication 49 (2007)
5. Sjölander, K., Beskow, J.: Wavesurfer, `http://www.speech.kth.se/wavesurfer/`
6. Brookes, M.: VOICEBOX: Speech Processing Toolbox for MATLAB, `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`
7. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden markovov models. Speech Communication 41, 603–623 (2003)