D. Ververidis and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models," in Proc. IEEE Inter. Symposium on Circuits and Systems (ISCAS), Japan, 2005.

Emotional speech classification using Gaussian mixture models

Dimitrios Ververidis and Constantine Kotropoulos

Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
E-mail: {jimver, costas}@zeus.csd.auth.gr

***Abstract*— In this paper, the classification of utterances into five basic emotional states is studied. A total of 87 statistical characteristics of pitch, energy, and formants is extracted from 500 utterances of the Danish Emotional Speech database. An evaluation of the classification capability of each feature is performed with respect to the probability of correct classification achieved by the Bayes classifier that models the feature probability density function as a mixture of Gaussian densities. Next, the feature subset that yields the highest probability of correct classification is found using the Sequential Floating Forward Selection algorithm. The probability of correct classification is estimated via crossvalidation and the probability density functions are modelled as mixtures of 2 or 3 Gaussian densities. The results demonstrate that the Bayes classifier which employs mixtures of 2 Gaussian densities can achieve a probability of correct classification equal to 0.55, whereas the human classification score is 0.67 for the database considered and the random classification would give a probability of correct classification equal to 0.20.**

I. Introduction

Speech emotion classification can be used for speech normalization in problems such as Automatic Speech Recognition (ASR), because the variety of speech talking styles affects dramatically ASR scores [1]. Furthermore, emotion and affect information extracted from speech is useful in a plethora of applications such as improving human-computer interaction [2]. The global statistical features of energy, pitch, and formants have been valuable for emotion classification [3], [4], [5]. The histograms of statistical characteristics of pitch and energy indicate that the corresponding probability distribution functions (pdfs) can be approximated by mixtures of Gaussian densities [5]. In this paper, an evaluation of the classification capability of global statistical features (Section III) of energy, pitch, and formants is performed with respect to the probability of correct classification achieved by the Bayes classifier that models the pdf of features as a mixture of Gaussian densities. Next, feature selection is performed using the Sequential Floating Forward Selection (SFFS) algorithm. The criterion employed for selecting the best features is the probability of correct classification that is estimated via crossvalidation when the feature pdfs are modelled as mixtures of 2 or 3 Gaussian densities.

The novelty of the proposed approach is in

- the classification of emotions for each gender separately in contrast to [4] where the gender information was not exploited;
- the modelling of the pdf of prosodic features with Gaussian mixtures (This is not the case in [3].);
- the inclusion of the neutral state in the classification in contrast to [6].

The outline of the paper is as follows. Section II describes the database used in our study. In Section III, the total set of features is presented. The discrimination capability of each isolated feature is studied in Section IV. The selection of an optimal set of features by employing the SFFS algorithm is described in Section V and the discrimination capability offered by such a set of features is assessed in Section VI. Finally, conclusions are drawn in Section VII.

II. Data

The Danish Emotional Speech (DES) database [7] has been chosen for our study, because it is easily accessible and well annotated. The data used in the experiments are sentences and words that are located between two silent segments. For example: 'Nej' (No), 'Ja' (Yes), 'Kom med mig' (Come with me!). The total amount of data used is 500 speech utterances (i.e., speech segments between two silence pauses) which are expressed by four professional actors, two males and two females. All utterances are equally separated for each gender. Speech is expressed in 5 emotional states such as anger, happiness, neutral, sadness, and surprise.

III. Feature Extraction

The pitch contour is created from pitch estimates obtained from the peaks of the short-term autocorrelation function of the speech amplitude. The short-term analysis is performed using windows of duration 15 msec. We assume that pitch frequencies are limited to the range 60-320 Hz. For estimating the 4 formant contours, we use a method based on linear prediction analysis. The method finds the angle of the poles in the $\mathcal{Z}$-plane for an all-pole model and considers the poles that are further from zero as indicators of formant frequencies. To estimate the energy contour, a simple short-term energy function has been used. After the evaluation of the primary raw features, secondary statistical features were extracted from the primary ones. The statistical features employed in our study are grouped in several classes. All features are referenced by their corresponding indices hereafter.

A. Spectral features

The set of spectral features is comprised by statistical properties of the first 4 formants and the energy below 250 Hz.

1. Energy below 250 Hz normalized to the length of the utterance
2. - 5. Mean value of the first, second, third, and fourth formant
6. - 9. Maximum value of the first, second, third, and fourth formant
10. - 13. Minimum value of the first, second, third, and fourth formant
14. - 17. Variance of the first, second, third, and fourth formant

B. Pitch features

The following features measure statistical properties of the pitch contour. The plateaux of the contours are detected as follows. The first derivative of the pitch contour is estimated numerically. The plateaux of maxima are located at the inflection points of the first derivative which have a value greater than 45% of the maximum pitch value admitted by the peak of the pitch contour. The inflection points whose pitch is less than 45% of the maximum value form the plateaux of minima.

18. - 22. Maximum, minimum, mean, median, interquartile range of pitch values
23. Pitch existence in the utterance expressed in percentage (0-100%)
24. - 27. Maximum, mean, median, interquartile range of durations for the plateaux at minima
28. - 30. Mean, median, interquartile range of pitch values for the plateaux at minima
31. - 35. Maximum, mean, median, interquartile range, upper limit (90%) of durations for the plateaux at maxima
36. - 38. Mean, median, interquartile range of the pitch values within the plateaux at maxima
39. - 42. Maximum, mean, median, interquartile range of durations of the rising slopes of pitch contours
43. - 45. Mean, median, interquartile range of the pitch values within the rising slopes of pitch contours
46. - 49. Maximum, mean, median, interquartile range of durations of the falling slopes of pitch contours
50. - 52. Mean, median, interquartile range of the pitch values within the falling slopes of pitch contours
53. Number of inflections in F0 contour

C. Intensity (Energy) features

Energy features are statistical properties of the energy contour. The first derivative of the energy contour is estimated numerically. The plateaux of maxima are located at the inflection points of the first derivative which have a value greater than the 45% of the maximum energy value admitted by the peak of the energy contour. The inflection points whose energy is less than the 45% of the maximum value of energy form the plateaux of minima.

54. - 58. Maximum, minimum, mean, median, interquartile range of energy values
59. - 62. Maximum, mean, median, interquartile range of durations for the plateaux at minima
63. - 65. Mean, median, interquartile range of energy values for the plateaux at minima
66. - 70. Maximum, mean, median, interquartile range, upper limit (90%) of duration for the plateaux at maxima
71. - 73. Mean, median, interquartile range of the energy values within the plateaux at maxima
74. - 77. Maximum, mean, median, interquartile range of durations of the rising slopes of energy contours
78. - 80. Mean, median, interquartile range of the energy values within the rising slopes of energy contours
81. - 84. Maximum, mean, median, interquartile range of durations of the falling slopes of energy contours
85. - 87. Mean, median, interquartile range of the energy values within the falling slopes of energy contours

IV. Evaluation of single features

The classification ability of each feature in isolation is rated according to the probability of correct classification achieved by the Bayes classifier when the feature pdf is modelled as a mixture of Gaussian densities. The probability of correct classification was estimated using crossvalidation. In particular, 100 replicas of the classification experiment were performed, where 90% of the data was randomly used for training and 10% for testing. The Expectation Maximization (EM) algorithm [8] is applied to determine the weights and the parameters of the Gaussian densities in the mixture. The implementation of EM algorithm described in [9] was used.

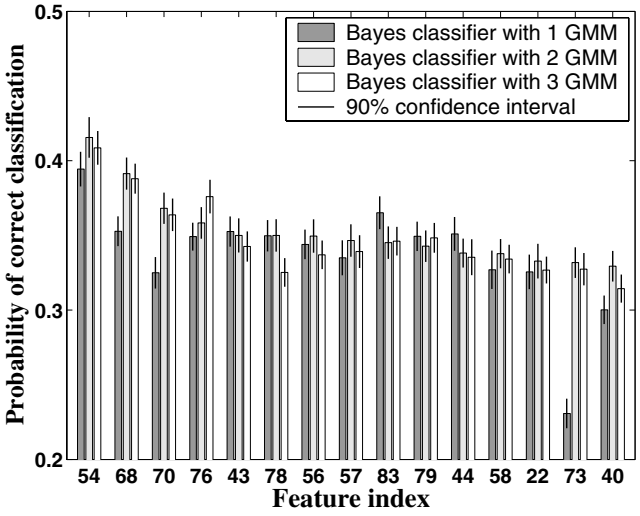

Fig. 1. Single feature evaluation using the probability of correct classification a Bayes classifier yields, when both genders are considered. (1 GMM stands for a pdf approximated by 1 Gaussian density; 2 GMM stands for a pdf approximated by a mixture of 2 Gaussians; 3 GMM stands for a pdf approximated by a mixture of 3 Gaussians.)

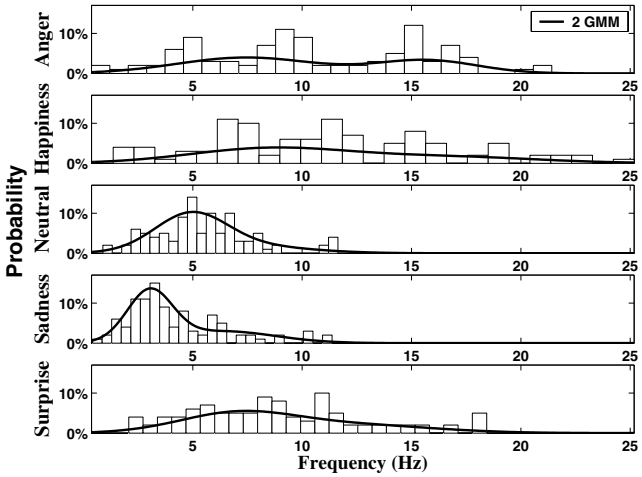

Fig. 2. Pdfs of the maximum value of energy (54) for the 5 emotional states considered in our study.

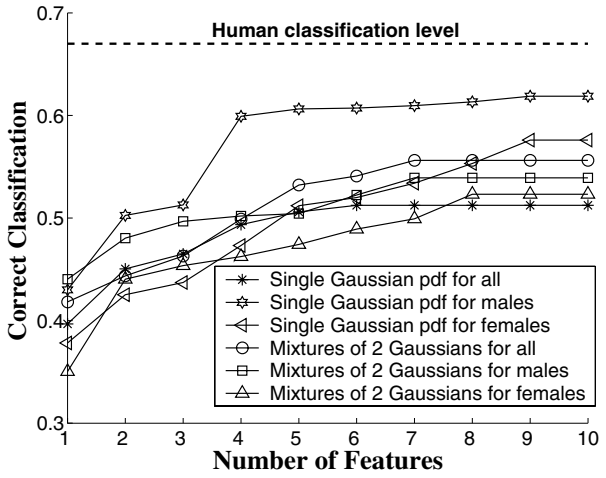

Fig. 3. Probability of correct classification of speech utterances into 5 emotional states for feature sets of increasing cardinality that are determined by the Sequential Floating Forward Selection algorithm applied to the Bayes classifier.

The derived Gaussian Mixture Models (GMMs) are employed in the likelihood computations inside the Bayes classifier to classify each feature to the class that maximizes the likelihood. The iterations of the EM algorithm must not exceed a certain number (usually 20-40) to avoid histogram over-fitting and if the weight of a mixture is not greater than 0.05, then that mixture weight is floored to zero in order to avoid "greedy" classification. In Figure 1, the features are sorted in descending order with respect to the probability of correct classification when the pdfs of emotional speech classes are modelled as mixtures of 2 Gaussians (2 GMMs).

Energy features dominate in the first 15 positions. Feature 54 (maximum value of energy) shows remarkably good results with any pdf modelling. The class pdfs of feature 54, modelled by mixtures of 2 Gaussian densities, are plotted in Figure 2. Other features as those with indices 78, 56, 58, and 79 behave similarly. The pitch features with numbers 44, 43, 40, and 22 achieve also a possibility of correct classification above 0.30 on average.

V. Automatic Feature Selection

An improved version of the SFS algorithm is used for automatic feature selection [10]. The simple SFS algorithm is subjected to nesting problems. It selects a feature that might not improve the criterion being optimized yielding a "dead end". The SFFS does a step backward by removing the last selected feature that led to a "dead end" and a step forward by selecting the second best feature. The criterion employed in the SFFS is the probability of correct classification achieved by the Bayes classifier that employs Gaussian mixture modelling of the pdf in the n-dimensional space with a full covariance matrix. The probability of correct classification obtained for several feature numbers is plotted in Figure 3.

TABLE I

10 best features selected by the Sequential Floating Forward Selection algorithm using as criterion the probability of correct classification.

Classifier\Step	1	2	3	4	5	6	7	8	9	10
Bayes with class pdfs modelled by 1 Gaussian for both genders	54	1	39	78	21	33	-	-	-	-
Bayes with class pdfs modelled by 1 Gaussian for males	54	22	58	1	8	76	32	11	10	-
Bayes with class pdfs modelled by 1 Gaussian for females	44	78	38	1	54	20	39	4	5	-
Bayes with class pdfs modelled by mixtures of 2 Gaussian densities for both genders	54	20	1	39	76	8	2	-	-	-
Bayes with class pdfs modelled by mixtures of 2 Gaussian densities for males	54	31	10	1	76	17	43	-	-	-
Bayes with class pdfs modelled by mixtures of 2 Gaussian densities for females	54	20	32	10	83	1	3	78	-	-

TABLE II

Probability of correct classification of the Bayes classifier when the class pdfs are modelled as a single Gaussian density or mixtures of Gaussian densities.

	Single Gaussian	2 GMMs	3 GMMs
Both genders	0.514	0.556	0.528
Males	0.618	0.539	0.485
Females	0.576	0.523	0.432

The ten best features selected by the SFFS algorithm when the experiments are conducted for both genders and for males and females separately are shown in Table I.

As can be seen from Table II, when the class pdfs are modelled by a single Gaussian distribution, the Bayes classifier achieves a 0.514 probability of correct classification for both genders. When the experiments are conducted separately for each gender then the Bayes classifier using a single Gaussian distribution achieves a probability of correct classification equal to 0.618 for males and 0.576 for females, respectively.

This is an indication that gender information influences positively the emotion classification.

When the class pdfs are modelled by mixtures of 2 Gaussian densities the probability of correct classification is found to be 0.556 for both genders. That is, an improvement of 4.2 % has been obtained. Accordingly, the modelling of pdfs using GMMs is proven beneficial. When we take into account the gender information and repeat the modelling of class pdfs with GMMs for 2 or 3 densities, no gains are obtained. Therefore, it seems to us that modelling the class pdfs with a single Gaussian density is adequate when the emotional speech classification experiments are performed separately for each gender.

TABLE III

CONFUSION MATRIX OF A BAYES CLASSIFIER WHEN CLASS PDFS ARE MODELLED AS A SINGLE GAUSSIAN DENSITY.

Stimuli	Response				
	Neutral	Surprise	Happiness	Sadness	Anger
Neutral	0.518	**0.094**	**0.111**	0.154	**0.121**
Surprise	0.082	0.478	0.273	0.021	**0.144**
Happiness	0.019	0.232	0.467	0.017	**0.263**
Sadness	**0.217**	**0.099**	**0.086**	0.510	**0.094**
Anger	0.022	0.133	**0.196**	0.053	0.594

TABLE IV

CONFUSION MATRIX OF A BAYES CLASSIFIER WHEN CLASS PDFS ARE MODELLED AS MIXTURES OF 2 GAUSSIAN DENSITIES.

Stimuli	Response				
	Neutral	Surprise	Happiness	Sadness	Anger
Neutral	0.521	**0.089**	**0.103**	0.198	0.086
Surprise	0.063	0.552	0.208	0.05	**0.204**
Happiness	0.072	0.198	0.497	0.021	**0.209**
Sadness	0.204	**0.078**	**0.071**	0.586	0.058
Anger	0.04	0.096	**0.198**	0.059	0.604

TABLE V

CLASSIFICATION RATES BY HUMANS.

Stimuli	Response				
	Neutral	Surprise	Happiness	Sadness	Anger
Neutral	0.608	0.026	0.001	0.317	0.048
Surprise	0.10	0.591	0.287	0.010	0.013
Happiness	0.083	0.298	0.564	0.017	0.038
Sadness	0.126	0.018	0.01	0.852	0.03
Anger	0.102	0.085	0.045	0.017	0.751

VI. CONFUSION MATRICES

The confusion matrix of the Bayes classifier is improved when the class pdfs are modelled by mixtures of 2 Gaussian densities (Table IV), instead of single Gaussian density (Table III) as it can be inferred from the comparison of their diagonal elements. From the inspection of the diagonal entries in Tables III and IV, we find out that the errors for surprise, happiness, and sadness are significantly reduced when the class pdfs are modelled as mixtures of 2 Gaussian densities . However, the probability of the correct classification of sadness and anger in the case of modelling the pdfs by 2 Gaussian densities are still significantly lower compared to the human rates in Table V. The latter table is originally found in [7]. The numbers in boldface indicate the cases where the Bayes classifier is more than twice as errorfull as the human subjects.

VII. CONCLUSION

By modelling the class pdfs with mixtures of Gaussian densities instead of a single Gaussian density, a gain of 4% is obtained. However, if the experiments are conducted separately for each gender there is no gain. This observation leads to the conclusion that each gender introduces its Gaussian distribution in the pdf. From the bank of the 87 global statistical features, the maximum value of energy, the energy below 250 Hz normalized to the length of the utterance, and the mean value of rising slopes of pitch are the most valuable.

ACKNOWLEDGMENT

This work has been partially supported by the research project 01E312 “Use of Virtual Reality for training pupils to deal with earthquakes” financed by the Greek Secretariat of Research and Technology.

REFERENCES

[1] S. E. Bou-Ghazale and J. H. L. Hansen, “A comparative study of traditional and newly proposed features for recognition of speech under stress,” *IEEE Trans. Speech and Audio Process.*, vol. 8, no. 4, pp. 429-442, July 2000.

[2] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, “Emotional speech: Towards a new generation of databases,” *Speech Communication*, vol. 40, pp. 33-60, 2003.

[3] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, “Approaching automatic recognition of emotion from voice: A rough benchmark”, in *Proc. ISCA Workshop Speech and Emotion*, pp. 207-212, Newcastle, 2000.

[4] B. Schuller, G. Rigoll, and M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid Support Vector Machine - Belief Network architecture,” in *Proc. 2004 IEEE Int. Conf. Acoustics, Audio and Signal Processing, (ICASSP 2004)*, vol. 1, pp. 577-580, Montreal, May 2004.

[5] D. Ververidis and C. Kotropoulos, “Automatic Speech Classification to five emotional states based on gender information,” in *Proc. 12th European Signal Processing Conf.*, pp. 341-344, Vienna, September 2004.

[6] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, “A corpus-based speech synthesis system with emotion,” *Speech Communication*, vol. 40, pp. 161-187, 2003.

[7] I. S. Engberg and A. V. Hansen, “Documentation of the Danish Emotional Speech Database (DES),” Internal AAU report, Center for Person Kommunikation, Denmark, 1996.

[8] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, Series B, vol. 1, no. 39, pp. 1–38, 1977.

[9] F. Vojtech and H. Vaclav, “Statistical Pattern Recognition Toolbox for Matlab,” Research reports of CMP, Czech Technical Univ. in Prague, No. 8, 2004.

[10] P. A. Devijver and J. Kittler, *Pattern Recogntion: A Statistical Approach*, London: Prentice-Hall International, 1982.

[11] D. Ververidis, C. Kotropoulos, and I. Pitas, “Automatic emotional speech classification,” in *Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 593-596, Montreal, May 2004.