AUTOMATIC EMOTIONAL SPEECH CLASSIFICATION

Dimitrios Ververidis, Constantine Kotropoulos, Ioannis Pitas

Artificial Intelligence and Information Analysis Laboratory
Department of Informatics, Aristotle Univ. of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
E-mail: {jimver, costas, pitas }@zeus.csd.auth.gr

ABSTRACT

Our purpose is to design a useful tool which can be used in psychology to automatically classify utterances into five emotional states such as anger, happiness, neutral, sadness, and surprise. The major contribution of the paper is to rate the discriminating capability of a set of features for emotional speech recognition. A total of 87 features has been calculated over 500 utterances from the Danish Emotional Speech database. The Sequential Forward Selection method (SFS) has been used in order to discover a set of 5 to 10 features which are able to classify the utterances in the best way. The criterion used in SFS is the crossvalidated correct classification score of one of the following classifiers: nearest mean and Bayes classifier where class pdfs are approximated via Parzen windows or modelled as Gaussians. After selecting the 5 best features, we reduce the dimensionality to two by applying principal component analysis. The result is a 51.6% $\pm$ 3% correct classification rate at 95% confidence interval for the five aforementioned emotions, whereas a random classification would give a correct classification rate of 20%. Furthermore, we find out those two-class emotion recognition problems whose error rates contribute heavily to the average error and we indicate that a possible reduction of the error rates reported in this paper would be achieved by employing two-class classifiers and combining them.

1. INTRODUCTION

Applications of emotional speech recognition can be foreseen in the broad area of human-computer interaction or in measuring presence in Virtual Reality (VR) environments. The efficiency of a VR immersion system is likely to be measured by the correlation of the emotion content of users' speech and the scenario of VR throughout the immersion.

In this paper, the discriminating capability of a set of features for emotional speech recognition is studied. A total of 87 features has been calculated over 500 utterances of the Danish Emotional Speech (DES) database [9]. In [1], 32 statistical properties of energy, pitch, and spectral features of emotional speech have been tested. This initial set of features is now augmented to 87 features by including more statistical features of pitch, spectrum, and energy. It is reported that a Bayes classifier that employs 5 features selected by SFS can achieve 51.6% $\pm$ 3% correct classification score when class pdfs are modeled as Gaussians.

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.

2. DATA

After a detailed review on the available emotional speech databases [4], we decided to work on DES, 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 dig' (Come with me!). The total amount of data used is 500 speech segments (with no silence interruptions), which are expressed by four professional actors, two male and two female. Speech is expressed in 5 emotional states, such as anger, happiness, neutral, sadness, and surprise.

3. FEATURE EXTRACTION

The pitch contour is derived by applying the method described in [2]. The method estimates the pitch from energy peaks of the short-term autocorrelation function computed over a window of 15 $msec$ duration. We assume that the pitch frequencies are limited to the range 60-320 Hz. It is worth noting that the method used for estimating the pitch contour [2] is more robust than others based on prediction analysis [6]. Furthermore, the method in [2] yields similar results to the method in [8]. Accordingly, we believe that this method is quite reliable. For estimating the 4 formant contours, we use the method proposed in [3]. The method finds the angle of the poles in z-space of 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 features, secondary (statistical) features were extracted from the primary ones. The statistical features employed in our study are grouped in several classes as is explained in the sequel. The features are referenced by their corresponding indices throughout the analysis following.

3.1. 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
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

3.2. Pitch features

Pitch features are statistical properties of the pitch contour. The plateaux of the contours are detected as follows. The first and second derivative of the contour are estimated numerically. The derivatives are smoothed with a moving average with a 15 $msec$ window length. If the first derivative is approximately zero and the second derivative is positive, the point belongs to a plateau at a local minimum. If the second derivative is negative, it belongs to a plateau at a local maximum.

18. - 22. Maximum, minimum, mean, median, interquartile range
23. Pitch existence in the utterance expressed in percentage (0-100%)
24. Maximum duration of plateaux at minima
25. Mean duration of plateaux at minima
26. Mean value of plateaux at minima
27. Median duration of plateaux at minima
28. Median value of plateaux at minima
29. Interquartile range of plateaux at minima
30. Interquartile duration of plateaux at minima
31. Maximum duration of plateaux at maxima
32. Mean duration of plateaux at maxima
33. Mean value of plateaux at maxima
34. Median duration of plateaux at maxima
35. Median value of plateaux at maxima
36. Interquartile range of plateaux at maxima
37. Interquartile duration of plateaux at maxima
38. Upper limit (90%) of duration of plateaux at maxima
39. Maximum duration of rising slopes
40. Mean duration of rising slopes
41. Mean value of rising slopes
42. Median duration of rising slopes
43. Median value of rising slopes
44. Interquartile range of rising slopes
45. Interquartile duration range of rising slopes
46. Maximum duration of falling slopes
47. Mean duration of falling slopes
48. Mean value of falling slopes
49. Median duration of falling slopes
50. Median value of falling slopes
51. Interquartile range of falling slopes
52. Interquartile duration range of falling slopes
53. Number of inflections in F0 contour

3.3. Intensity (Energy) features

Energy features are statistical properties of the energy contour.

54. - 58. Maximum, minimum, mean, median, interquartile range
59. Maximum duration of plateaux at minima
60. Mean duration of plateaux at minima
61. Mean value of plateaux at minima
62. Median duration of plateaux at minima
63. Median value of plateaux at minima
64. Interquartile range of plateaux at minima
65. Interquartile duration range of plateaux at minima
66. Maximum duration of plateaux at maxima
67. Mean duration of plateaux at maxima
68. Mean value of plateaux at maxima
69. Median duration of plateaux at maxima
70. Median value of plateaux at maxima
71. Interquartile range of plateaux at maxima
72. Interquartile duration range of plateaux at maxima
73. Upper limit (90%) of duration of plateaux at maxima
74. Maximum duration of rising slopes
75. Mean duration of rising slopes
76. Mean value of rising slopes
77. Median duration of rising slopes
78. Median value of rising slopes
79. Interquartile range of rising slopes
80. Interquartile duration range of rising slopes
81. Maximum duration of falling slopes
82. Mean duration of falling slopes
83. Mean value of falling slopes
84. Median duration of falling slopes
85. Median value of falling slopes
86. Interquartile range of falling slopes
87. Interquartile duration range of falling slopes

4. EVALUATION OF SINGLE FEATURES

In order to study the classification ability of each feature, a rating method has been implemented. Each feature is evaluated by the ratio between the between-class variance (σ_b^2) and the within-class variance (σ_w^2). The between class variance measures the distance between the class means, whereas the within-class variance measures the dispersion within each class [7]. The best features should be characterized by a large σ_b^2 and a small σ_w^2. The 16 features with the highest ration (σ_b^2/σ_w^2) are shown in Figure 1, where both σ_b^2 and σ_w^2 are depicted. The evaluation is rather qualitative than quantitative, because it implies indirectly that classification information is enclosed in a single feature. We note that y axis has positive values and it is not symmetrical.

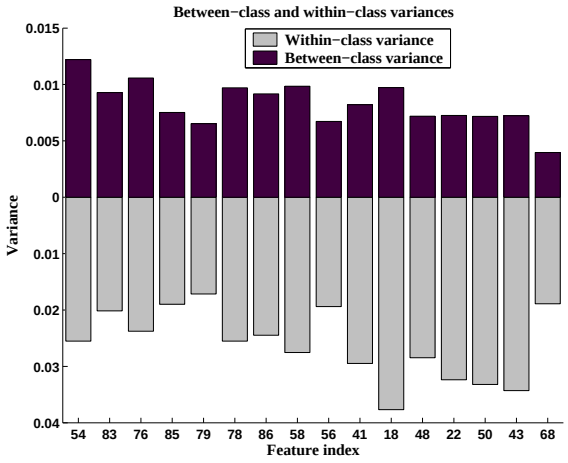

Fig. 1. Feature assessment based on the ratio between the between-class variance (σ_b^2) and the within-class variance (σ_w^2).

Energy features dominate in the 16 first positions. Feature 76 (mean value of rising slopes of energy) shows remarkably good results, namely 40% correct classification rate when it is employed in a Bayes classifier, as is depicted in Figure 4. The class pdfs of feature 76 for the five emotions under study are plotted in Figure 2. The energy level is simply the norm of 15 $msec$ frames that over-

lap by 10 $msec$. Other features such as those with indices 54, 58, 78, 83, 86, 85, 56, and 79 behave similarly.

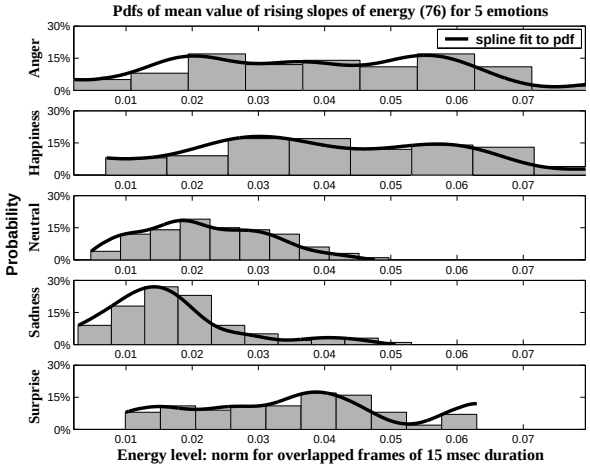

Fig. 2. From the inspection of Figure 2 we conclude that a maximum likelihood classifier using feature 76 will classify low energy measurements to neutral and sadness, whereas high energy instances to anger, surprise, and happiness.

The mean value of rising slopes of pitch (41) has achieved a 34% correct classification rate with a Bayes classifier. This feature is valuable, because it can make a good separation between surprise and neutral emotions as can be seen in Figure 3. Another valuable feature is maximum value of pitch (18) which can separate anger from surprise.

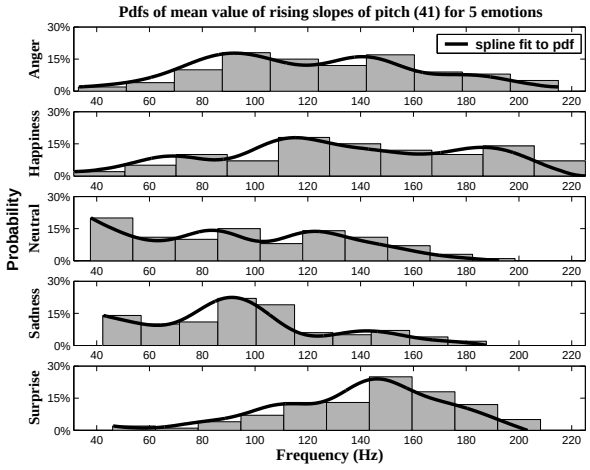

Fig. 3. Surprise is separated from neutral and sadness using the mean value of rising slopes of pitch.

5. AUTOMATIC FEATURE SELECTION

The (SFS) algorithm is used for automatic feature selection [5]. The criterion employed is the correct classification rate achieved by the selected features. Figure 4 demonstrates the correct classification rate obtained for several feature numbers. The SFS is applied to three classifiers, namely the nearest mean and Bayes classifier where class pdfs are approximated via Parzen windows or modelled as Gaussians. The correct classification rate is calculated by crossvalidation where 90% of the data were used for training and 10% for validation. We have chosen the features selected by SFS with a Bayes classifier when class pdfs are modelled as Gaussians as a criterion. The result is that features 76 (mean value of rising slopes of energy), 18 (maximum value of pitch) , 44 (interquartile range of rising slopes of pitch), 27 (median duration of plateaux at minima of pitch) and 7 (maximum value of the second formant) can achieve $51.6\% \pm 3\%$ correct classification rate at 95% confidence interval. Features 76 and 18 can achieve together a correct classification rate of 45%. Table 1 enlists the indices of the ten best features found for each classifier by the SFS method.

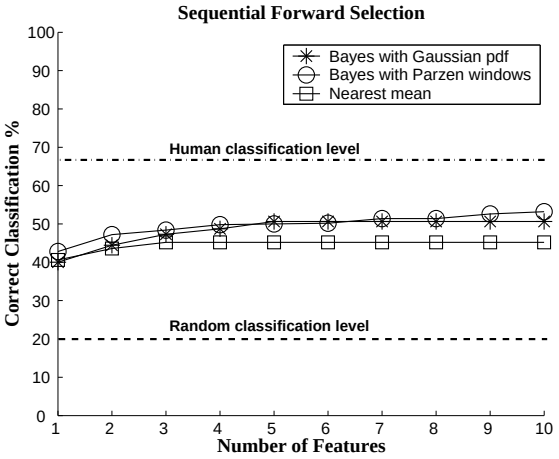

Fig. 4. Selecting 10 features with the sequential forward selection method. A Bayes classifier where class pdfs are approximated via Parzen windows outperforms the other classifiers but a Bayes classifier where class pdfs are modelled as Gaussians finds more distinct classes in the 2 dimensional space (see Figure 5).

Table 1. 10 best features selected by the sequential forward selection algorithm using as criterion the correct classification rate for each classifier.

	Forward selection steps									
Classifier\Step	1	2	3	4	5	6	7	8	9	10
Bayes	76	18	44	27	7	-	-	-	-	-
Parzen	54	81	18	42	79	43	57	47	33	67
Nearest mean	54	39	1	-	-	-	-	-	-	-

6. REPRESENTATION TO A TWO-DIMENSIONAL SPACE

After selecting the best five features, namely those with indices 76, 18, 44, 27 and 7, Principal Component Analysis (PCA) was used in order to reduce the dimensionality from five dimensions (5D) to two dimensions (2D) and to represent the samples in a 2D space. Only the samples which belong to the interquartile range of the pdf for each class are shown in Figure 5.

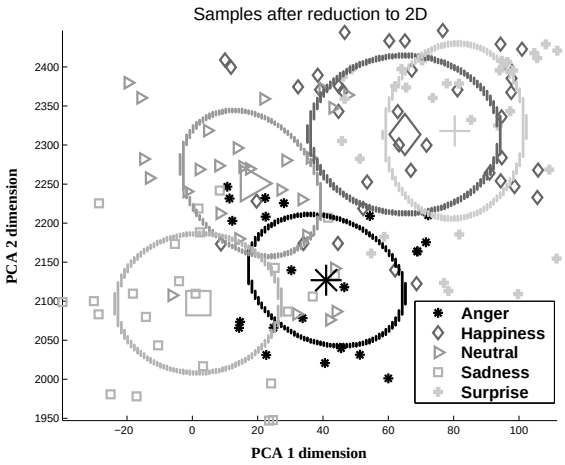

Fig. 5. Partition of the 2D domain into five emotional states derived by SFS and PCA. The samples which belong to the interquartile range of each pdf are shown. The big symbols denote the mean of each class. The ellipses denote the 60% likelihood contours for a 2-D Gauss model.

7. OPEN PROBLEMS AND FUTURE WORK

In order to figure out the misclassifications introduced by a Bayes classifier, we compare the confusion matrix of a Bayes classifier (Table 2) to the confusion matrix of humans (Table 3). The latter confusion matrix was obtained from [9]. From the diagonal entries in Table 2 we find out that the automatic speech emotion classification system commits gross errors on the emotional states happiness and anger. The numbers in boldface indicate the cases where a Bayes classifier is more than twice as errorful as the subjects. From the inspection of the off-diagonal entries we find out that:

- neutral is more frequently misclassified as surprise;
- surprise is more frequently misclassified as sadness or anger;
- happiness is more frequently misclassified as sadness or anger;
- anger is more frequently misclassified as happiness or sadness;

The rates reported in Table 2 can be further improved by analyzing the properties of the above mentioned two-class problems. The features which can separate two classes could be different from those which separate 5 classes. By designing proper decision fusion algorithms, we may combine several two-class classifiers and the overall system could outperform the rates obtained by the five-class classifiers.

8. CONCLUSIONS

This study was based on features related to the energy, the pitch and the formants of a speech signal in order to classify the emotional content of speech. Our analysis has verified the previously published result in [1] that features 79 (interquartile range of rising slopes of energy), 86 (interquartile range of falling slopes) and 22 (interquartile range of pitch contour) were among the sixteen most powerful ones as can be seen from the indices of the sixteen best features in Figure 1.

Table 2. Confusion matrix for a Bayes classifier using features 76, 18, 44, 27, 7. The result is a correct classification rate of 54% when all data are used for training and testing. When crossvalidation method used, a correct classification rate of 51.6% is obtained.

Confusion matrix for Bayes classifier

Stimuli	Response (%)				
	Neutral	Surprise	Happiness	Sadness	Anger
Neutral	56	**13**	3	25	3
Surprise	6	65	5	**9**	**15**
Happiness	9	24	39	**14**	**14**
Sadness	17	6	1	72	4
Anger	14	14	**20**	**12**	40

Table 3. Correct classification rates by humans.

Stimuli	Response (%)				
	Neutral	Surprise	Happiness	Sadness	Anger
Neutral	60.8	2.6	0.1	31.7	4.8
Surprise	10	59.1	28.7	1.0	1.3
Happiness	8.3	29.8	56.4	1.7	3.8
Sadness	12.6	1.8	0.1	85.2	0.3
Anger	10.2	8.5	4.5	1.7	75.1

9. REFERENCES

[1] S. McGilloway, R. Cowie, E. Douglas-Cowie et al., "Approaching automatic recognition of emotion from voice: A rough benchmark", in *Proc. ISCA Workshop Speech and Emotion*, pp. 207-212, Newcastle, 2000.

[2] P. Loizou, "Colea: A MATLAB software-tool for Speech Analysis", University of Arkansas, May 2003, http://www.utdallas.edu/ loizou/speech/colea.htm

[3] L. Arslan, "Speech toolbox in MATLAB", Bogazici University, http://www.busim.ee.boun.edu.tr/ arslan/

[4] D. Ververidis and C. Kotropoulos, "A State of the Art Review on Emotional Speech Databases", in *Proc. 1st Richmedia Conference*, Lausanne, Switzerland, pp. 109-119, October 2003.

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

[6] S. Malcolm, "Auditory toolbox in MATLAB", version 2, University of Purdue, http://rvl4.ecn.purdue.edu/ malcolm/interval/1998-010/

[7] K. Fugunaka, *Introduction to Statistical Pattern Recognition*, N.Y.: Academic Press, 1990.

[8] K. Sjolander and J. Beskow, "Wavesurfer - an open source speech tool", *in Proc. ICSLP 2000*, Beijing.

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