AUTOMATIC SPEECH CLASSIFICATION TO FIVE EMOTIONAL STATES BASED ON GENDER INFORMATION

Dimitrios Ververidis and Constantine Kotropoulos

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

ABSTRACT

Emotional speech recognition aims to automatically classify speech units (e.g., utterances) into emotional states, such as anger, happiness, neutral, sadness and surprise. The major contribution of this paper is to rate the discriminating capability of a set of features for emotional speech recognition when gender information is taken into consideration. A total of 87 features has been calculated over 500 utterances of the Danish Emotional Speech database. The Sequential Forward Selection method (SFS) has been used in order to discover the 5-10 features which are able to classify the samples in the best way for each gender. The criterion used in SFS is the crossvalidated correct classification rate of a Bayes classifier where the class probability distribution functions (pdfs) are approximated via Parzen windows or modeled as Gaussians. When a Bayes classifier with Gaussian pdfs is employed, a correct classification rate of 61.1% is obtained for male subjects and a corresponding rate of 57.1% for female ones. In the same experiment, a random classification would result in a correct classification rate of 20%. When gender information is not considered a correct classification score of 50.6% is obtained.

1. INTRODUCTION

This paper is devoted to emotional speech recognition, that deals with automatic classification of utterances into emotional states. Applications of emotional speech recognition can be foreseen in the broad area of human-computer interaction or in measuring immersion in Virtual Reality environments [9]. In [1], 32 statistical properties of energy, pitch, and spectral features of emotional speech have been tested. This initial set of features is augmented to 87 features, as it can be seen in section 2, by including more statistical features of pitch, spectrum, and energy. In section 3 the total set of features is calculated over 500 utterances of the Danish Emotional Speech (DES) database [8]. In section 4, the discrimination capability of each feature is studied. In sections 5 and 6, the joint discrimination capability of several features is assessed.

2. DATA

After a detailed review on available emotional speech databases [4], we decided to work on DES because it was 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 and equally separated to each gender. Speech is expressed in 5 emotional states, such as anger, happiness, neutral, sadness, and surprise.

This work has been partially supported by the FP6 European Union Network of Excellence "Multimedia Understanding through Semantics, Computation and LEarning" (IST-2002-2.3.1.7).

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 duration 15 msec. We assume that the pitch frequencies are limited to the range 60-320 Hz. For estimating the 4 formant contours we use the method proposed in [3]. The method finds the angle of the poles in z-plane 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 feature contours, secondary (statistical) features were extracted from the primary ones. The statistical features employed in our study are grouped in several classes. 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 of 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. - 27. Maximum, mean, median, interquartile range of duration of plateaux at minima
28. - 30. Mean, median, interquartile range of values of plateaux at minima
31. - 35. Maximum, mean, median, interquartile range, upper limit (90%) of duration of plateaux at maxima
36. - 38. Mean, median, interquartile range of values of plateaux at maxima
39. - 42. Maximum, mean, median, interquartile range of durations of rising slopes
43. - 45. Mean, median, interquartile range of values of rising slopes
46. - 49. Maximum, mean, median, interquartile range of durations of falling slopes
50. - 52. Mean, median, interquartile range of values 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. - 62. Maximum, mean, median, interquartile range of durations of plateaux at minima
63. - 65. Mean, median, interquartile range of values of plateaux at minima
66. - 70. Maximum, mean, median, interquartile range, upper limit (90%) of duration of plateaux at maxima
71. - 73. Mean, median, interquartile range of values of plateaux at maxima
74. - 77. Maximum, mean, median, interquartile range of durations of rising slopes
78. - 80. Mean, median, interquartile range of values of rising slopes
81. - 84. Maximum, mean, median, interquartile range of durations of falling slopes
85. - 87. Mean, median, interquartile range of values 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 15 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.

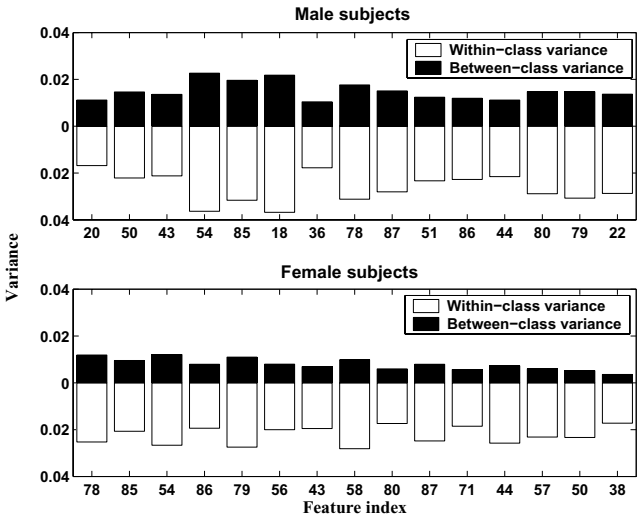

Figure 1: Feature assessment based on the ratio between the between-class variance (σ_b^2) and the within-class variance (σ_w^2) for each gender.

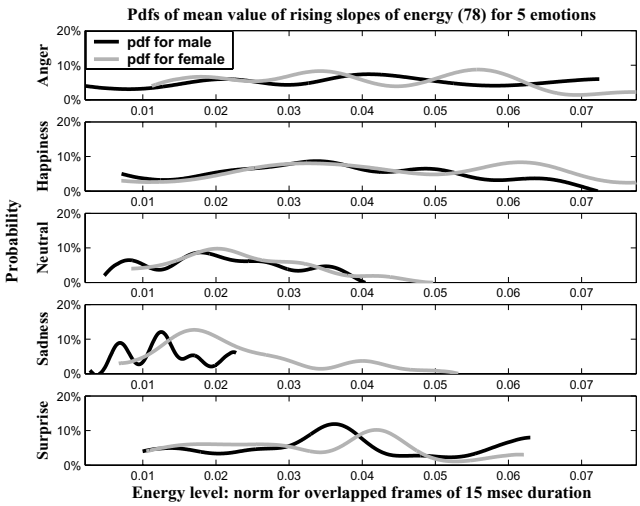

Figure 2: From the inspection of Figure 2 we conclude that a maximum likelihood classifier using feature 78 will classify low energy measurements to neutral and sadness, whereas high energy instances to anger, surprise, and happiness. Males differ from females in sadness and surprise.

In females, energy features dominate in the first 15 positions, whereas in males both energy and pitch features appear. Feature 78 (mean value of rising slopes of energy) shows remarkably good results, namely a correct classification rate of 41% for male, 36% for female, and 40% for both, when it is employed in a Bayes classifier. The class pdfs of feature 78 for the five emotions under study are plotted in Figure 2. We note that the pdf curves are splines fitted to the discrete pdf of each class. The energy level is simply the norm of 15 *msec* frames that overlap by 10 *msec*. Other fea-

tures such as those with indices 54, 56, 58, 79, 85, 86 and 87 behave similarly.

Feature 43 (mean value of rising slopes of pitch) achieves a correct classification rate of 35.9% for male, 39.1% for female, and 34% for both genders, when it is employed in a Bayes classifier. The class pdfs of feature 43 are depicted in Figure 3. Females have increased frequency levels for feature 43 only in the categories of neutral, happiness and sadness. Feature 20 (mean value of pitch contour) achieves a correct classification rate of 37.5% for male and 33.5% for female, when it is employed in a Bayes classifier. The class pdfs of feature 43 are depicted in Figure 4. However, it does not yield significant results when gender information is not used.

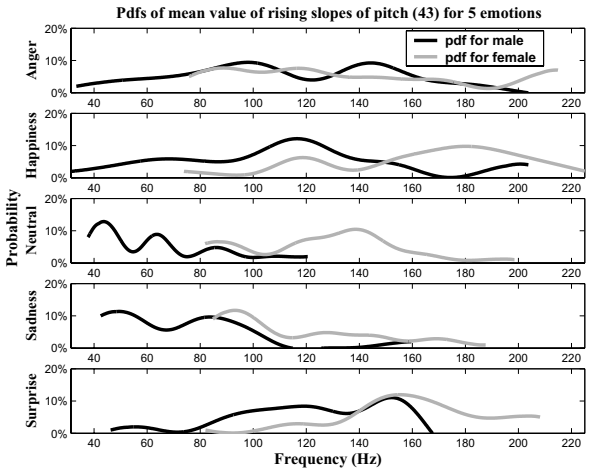

Figure 3: Pdfs of mean value of rising slopes of pitch for 5 emotions.

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 5 demonstrates the correct classification rate obtained for several feature numbers. The SFS is applied to two classifiers, namely the Bayes classifier where class pdfs are approximated via Parzen windows or modeled 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 as best those features selected by SFS for a Bayes classifier when class pdfs are modeled as Gaussians. The features included in the row of Table 1 referred to as "Bayes with Gaussian pdfs (male)" can achieve a 61.1% correct classification rate. The feature selection shown in the row of Table 1 referred to as "Bayes with Gaussian pdfs (female)" can achieve a correct classification rate of 57.1%. When gender information is not considered, the feature selection in the last row of Table 1 achieve a 50.6% correct classification rate.

6. CONFUSION MATRICES

In order to figure out the misclassifications introduced by a Bayes classifier, we compare the confusion matrices of Bayes classifiers (Tables 2, 3 and 4) to the confusion matrix of humans (Table 5). The latter confusion matrix was obtained from [8]. The confusion matrices in Tables 2, 3 and 4 have been calculated by taking the average classification rate of a

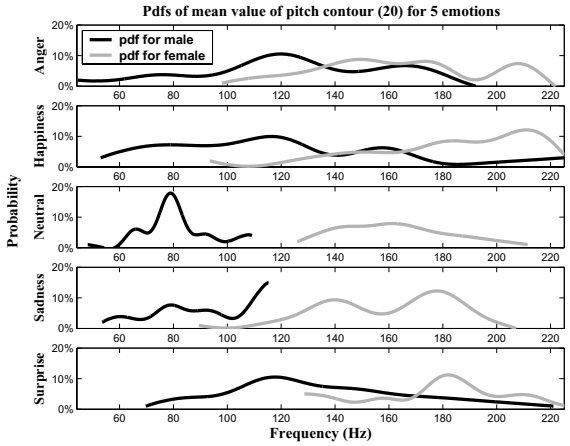

Figure 4: Males are separated from females at the states happiness, neutral, sadness, and surprise using the mean value of pitch contour.

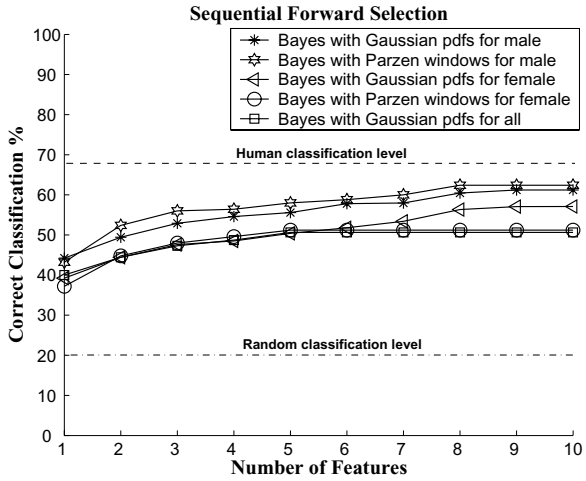

Figure 5: 10 best features selected by the sequential forward selection algorithm using as criterion the correct classification rate for each classifier.

Bayes classifier with Gaussian pdfs for the feature selections shown in Table 1 over 225 replicas of the experiment using crossvalidation, where 90% of the data were used for training and 10% for validation.

When gender information is not taken into consideration then a Bayes classifier that employes Gaussian pdfs achieves a 50.6% correct classification rate (see Table 2). When gender information is included then the classifier achieves a 61.1% and 57.1% correct classification rate for males and females, respectively.

From the diagonal entries in Table 2 we find out that the automatic speech emotion classification system commits gross errors on the emotional states of happiness and anger, when gender information is not included. The numbers in boldface indicate the cases where a Bayes classifier is more than twice as errorful as the human subjects. From the diagonal entries in Table 3, we notice that severe errors exist only in the emotional states of anger and sadness for females. From the inspection of the diagonal entries in Tables 2 and Tables 4, we find out that the errors for happiness and anger are signifi-

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 with Gaussian class pdfs (male)	54	43	74	81	21	78	8	69	18	-
Bayes with Parzen windows (male)	54	74	20	67	86	58	17	30	-	-
Bayes with Gaussian class pdfs (female)	43	78	20	25	10	77	6	17	82	45
Bayes with Parzen windows (female)	43	80	18	39	86	-	-	-	-	-
Bayes with Gaussian class pdfs	78	18	45	26	7	-	-	-	-	-

cantly reduced when utterances of male subjects are classified.

Table 2: Confusion matrix of a Bayes classifier when gender information is not exploited. When crossvalidation is used, a correct classification rate of 50.6% is obtained.

Classification rates of a Bayes classifier for subjects of both genders

Stimuli	Response (%)				
	Neutral	Surprise	Happiness	Sadness	Anger
Neutral	51	**15**	2	28	4
Surprise	5	64	7	**9**	**14**
Happiness	9	24	36	**13**	**18**
Sadness	17	6	2	70	5
Anger	12	19	**26**	**12**	31

7. DISCUSSION

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. The rates reported in Tables 3 and 4 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.

Table 3: Confusion matrix of a Bayes classifier applied to utterances of female subjects. The result is a 57.1% correct classification rate.

Classification rates of a Bayes classifier for female subjects

Stimuli	Response (%)				
	Neutral	Surprise	Happiness	Sadness	Anger
Neutral	55	**13**	6	20	6
Surprise	12	61	14	6	7
Happiness	12	11	54	4	**18**
Sadness	13	4	4	58	**21**
Anger	6	10	**18**	9	57

Table 4: Confusion matrix of a Bayes classifier for male subjects. The result is a 61.1% correct classification rate.

Classification rates of a Bayes classifier for male subjects

Stimuli	Response (%)				
	Neutral	Surprise	Happiness	Sadness	Anger
Neutral	67	3	8	20	1
Surprise	3	60	18	6	**13**
Happiness	**18**	13	43	6	**21**
Sadness	11	3	2	80	3
Anger	6	**19**	**13**	6	56

Table 5: 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

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

[9] M. C. Whitton, "Making virtual environments compelling," *Communications of the ACM*, vol. 46, no. 7, pp. 40-46, July 2003.