FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 27, N° 3, September 2014, pp. 425 - 433
DOI: 10.2298/FUEE1403425B

RELEVANCE OF THE TYPES AND THE STATISTICAL PROPERTIES OF FEATURES IN THE RECOGNITION OF BASIC EMOTIONS IN SPEECH

Milana Bojanić, Vlado Delić, Milan Sečujski

Faculty of Technical Sciences, University of Novi Sad, Serbia

Abstract. *Due to the advance of speech technologies and their increasing usage in various applications, automatic recognition of emotions in speech represents one of the emerging fields in human-computer interaction. This paper deals with several topics related to automatic emotional speech recognition, most notably with the improvement of recognition accuracy by lowering the dimensionality of the feature space and evaluation of the relevance of particular feature types. The research is focused on the classification of emotional speech into five basic emotional classes (anger, joy, fear, sadness and neutral speech) using a recorded corpus of emotional speech in Serbian.*

Key words: *emotional speech recognition, acoustic features, basic emotions*

1. INTRODUCTION

Basic emotion is a term used in categorical emotion models, among which Ekman's concept of six basic emotions is the most prominent one. His theory of basic emotions, which are "psychological universals and constitute a set of basic, evolved functions that are shared by all humans", is supported with experimental findings of cross-culturally recognized emotions from vocal signals and facial expressions [1].

From the beginning of its development, Emotional Speech Recognition (ESR) studies have used corpora of acted emotional speech since those corpora were easy to collect. Such corpora usually contained several basic emotions reproduced by actors [2].

There are apparently reasonable objections about acted speech corpora, saying that acting emotions is not the same as producing 'spontaneous' emotions and pointing out that within human-machine interaction emotion-related states are much more common than prototypical full-blown emotions (such as those represented in acted speech corpora) [3]. Still, recent research has shown that the relationships between the acted emotions and their acoustic correlates and between real life emotions and their acoustic correlates do not necessarily contradict [4].

Received February 10, 2014; received in revised form March 13, 2014
Corresponding author: Milana Bojanić
University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
(milana.bojanic@uns.ac.rs)

A more flexible solution to the problem of the representation of emotional states is to represent them as points in the continuous 2D space whose co-ordinates are the activation and evaluation involved in the emotional state [5]. Such dimensional models also allow for the mapping of basic emotions into the continuous 2D emotional space [5, 6], thus enabling a broad field of application of the recognition of basic emotions in speech.

The paper summarizes our approach to the recognition of basic emotions in speech, focusing particularly on the improvement of recognition accuracy by lowering the dimensionality of the feature space. Additionally, a feature selection procedure has been performed in order to rank feature types and used statistical functionals. The presented research has been conducted on a corpus of acted emotional speech in Serbian.

The paper is organized as follows. Aspects of the proposed approach that are relevant to the recognition of basic emotions, including acoustic modeling, classification scheme and speech corpus, are presented in Section 2. In Section 3, theoretical background about feature dimensionality reduction techniques is given and their possible benefits are pointed out. Experimental results are shown and discussed in Section 4. Finally, the conclusions are given in Section 5.

2. The Proposed Approach

2.1 The proposed approach to acoustic modeling

The proposed approach to acoustic modeling is based on the statistical analysis of acoustic feature contours [7, 8] and it is performed in three stages, as shown in Fig. 1.

The first stage includes the extraction of acoustic features on a frame basis. These features belong to two acoustic feature sets, namely prosodic and spectral feature set. In the prosodic feature set, pitch and energy are extracted. As to spectral feature set, only the first 12 MFCCs are taken into account in our analysis, since they correspond to slow changes in the spectrum, i.e., the spectrum envelope. The feature contours which correspond to the pitch contour, energy contour and MFCC contour, are, respectively, sequences of short-term pitch, energy and MFCC values extracted on a frame basis.

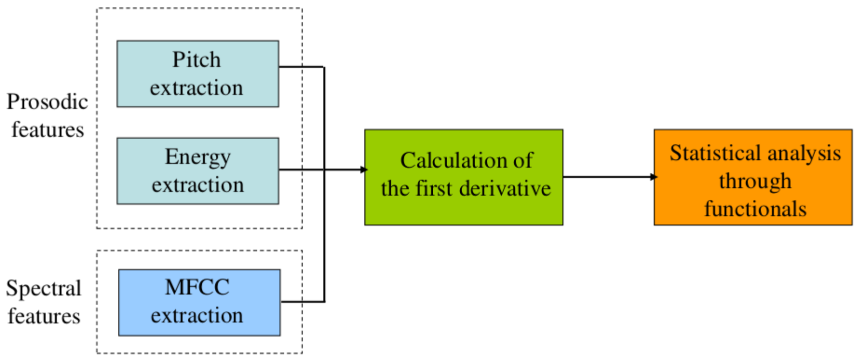

Fig. 1 Feature extraction process in three stages

The extracted features are forwarded to the second stage, in which the first derivative of the acoustic features is calculated in order to model the dynamics of speech. The first derivative carries the information about the dynamics of emotional speech, which is useful in emotional speech classification [4].

The third stage of the feature extraction process involves a statistical analysis of the feature contours. The final feature set is obtained from the feature contours by applying so-called static modeling through functionals [9].

In the literature, larger numbers of statistical features are analyzed [10, 11]. Our selection of statistical functionals was guided by the principle that chosen statistical features should describe the variations and follow the trend of changes of acoustic features correlated with different types of emotional speech. At the same time, since it was impossible to predict which statistical characteristics would be the most effective, the proposed set of statistical features included 12 features, bearing in mind that if particular information in the feature vector showed to be redundant and aggravating for classification, an efficient subset of features would be extracted using a dimensionality reduction technique.

The proposed set of 12 statistical functionals has been chosen from three groups of functionals which are the most frequently used [9]. These groups and their corresponding functionals are [7]:

1. The first four moments (mean, standard deviation, skewness and kurtosis),
2. Extrema and their positions (minimum, maximum, range, relative position of minimum and relative position of maximum),
3. Regression coefficients (the slope and the offset of the linear regression of the contour) and regression error (the mean squared error between the regression curve and the original contour).

By applying the proposed procedure, three sets of features have been extracted [7]. The first feature set includes only prosodic features (pitch and energy) and it will be referred to as prosodic feature set (P-FS). The second feature set includes only spectral features (12 MFCC); this set will be referred to as spectral feature set (S-FS). Finally, the third feature set includes both prosodic and spectral features, and additionally the voicing probability and the zero crossing rate. For the mentioned 16 features, the first derivative is calculated, and then 12 functionals are applied on all of them, resulting in 384 features extracted for each utterance. The third feature will be referred to as prosodic-spectral feature set (PS-FS).

2.2 Classification scheme

For the purpose of emotional speech classification, we have considered the linear discriminant classifier (LDC) and the *k*-nearest neighbours classifier (kNN), as they belong to well known and simple classifiers, which have been used by other researchers for this purpose and which have proved to be successful for both acted and spontaneous emotional speech [9]. As for LDC, two classification schemes have been considered. The first one is the linear Bayes classifier with the underlying assumption that classes have Gaussian densities and equal covariance matrices. The second one is the derivation of linear discriminant functions via the perceptron rule [12]. In the latter case, no assumptions have been made about the underlying class densities.

2.3 Emotional speech corpus

The research was conducted on the Corpus of Emotional and Attitude Expressive Speech (GEES, according to the Serbian acronym), which is the first speech corpus recorded in Serbian for the purpose of research on acoustic manifestations of emotions in human speech in the context of speech technology [13]. It contains recordings of acted speech-based emotional expressions corresponding to five basic emotional states: anger, joy, fear, sadness, and neutral, reproduced by six actors (3 female, 3 male). The underlying textual material is emotionally neutral with respect to lexical content and for the purpose of this study a section of the corpus including 30 short and 30 long sentences was used. The reported human recognition accuracy for this corpus is 94.7%. To avoid an imbalance between male and female speakers, an equal portion of the material from each emotional class belonging to each speaker was chosen and a total of 1740 sentences (75 minutes of speech) have been processed. Both training and test sets included utterances from all speakers. Therefore, these experiments belong to the case of speaker dependent emotion recognition.

3. Dimensionality Reduction

Dimensionality reduction can be performed through feature extraction or feature selection. While feature extraction employs a mapping (usually linear) of a given feature space onto a lower dimensional space, creating a feature subset which is a combination of existing features, feature selection involves a selection of a subset from the existing features without any transformation.

3.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a linear feature extraction technique whose goal is the enhancement of the class-discriminatory information in a lower dimensional feature space. Fisher's LDA for a two-class problem is based on a search for a projection that maximizes the ratio of between-class to within-class scatter. The solution is in a specific choice of direction for the projection of the data where the examples from the same class are projected so as to be very close to each other and, at the same time, the projected class means are projected so as to be as far from each other as possible [14].

Fisher's LDA generalizes easily for a C class problem (in our case C = 5 since we deal with 5 emotional classes). Since the projection is no longer a scalar (it has $C-1$ dimensions), the determinants of the scatter matrices are used to obtain an objective function. Between-class scatter matrix represents the scatter of the class mean vectors around the mixture mean, defined as:

$$S_B = \sum_{i=1}^{C} N_i\,(\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}}, \tag{1}$$

where $\mu_i = \frac{1}{N_i}\sum_{x \in C_i} x$ is the mean vector of each class in the original feature space x, and

$\mu = \frac{1}{N}\sum_{\forall x} x$ is the mean vector of the mixture distribution.

A within-class scatter matrix shows the scatter of samples around their respective class mean vectors, and is expressed by:

$$S_W = \sum_{i=1}^{C} S_i \ , \tag{2}$$

where

$$S_i = \sum_{x \in C_i} (x - \mu_i)(x - \mu_i)^{\mathrm{T}} \ . \tag{3}$$

It can be shown that the optimal projection matrix $W^* = [w_1^* \mid w_2^* \mid \ldots \mid w_{C-1}^*]$ is the one whose columns are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigenvalue problem [14]:

$$(S_B - \lambda_i S_W)\, w_i^* = 0 \ . \tag{4}$$

The projections with maximum class separability information are the eigenvectors corresponding to the largest eigenvalues λ_i of the matrix $S_W^{-1} S_B$.

3.2 Feature selection

The drawback of feature extraction methods is that they are not very appropriate for feature mining, as the original features are not retained after the transformation [9]. In order to gain an insight into the significance of particular features, feature selection was used. We adopted Sequential Forward Feature Selection (SFFS) as the search strategy and wrapper based evaluation as the objective function. SFFS starts the selection with an empty set and sequentially adds the feature that results in the highest value of the objective function when combined with the already selected features [15]. In the case of wrappers the objective function is a classifier which evaluates feature subsets by their recognition rate on test data employing cross-validation. In our case, linear Bayes classifier was selected as the wrapper as it had shown the best performance in previous recognition tests [7]. Ideally, feature selection methods should not only reveal the single most relevant attribute (or groups thereof), but they should also decorrelate the feature space [9]. Feature selection results in a reduced, interpretable set of significant features; their counts and weights in the selection set allow us to draw conclusions on the relevance of the feature types they belong to [16]. The feature set used in our feature selection experiments was PS-FS. Since it is a combination of both prosodic and spectral features, the relevance of particular feature types within PS-FS was expected to be evaluated.

4. EXPERIMENTAL RESULTS

The focus of the research was on the investigation of a possible improvement of recognition accuracy in the case of a reduced feature space in the task of basic emotions classification. Therefore, the performances of each classifier were tested in two ways: (1) using 3 extracted feature sets (P-FS, S-FS, PS-FS), and (2) using 3 feature sets obtained after LDA feature reduction has been applied on the 3 initial feature sets. The experiments were carried out using 3 classification techniques (the kNN classifier, the linear Bayes classifier and the perceptron rule).

Table 1 shows the class and average recognition rate of the kNN classifier (k = 9) in case of 3 feature sets, before LDA (originally extracted feature sets) and after LDA (original feature space reduced to 4 projection vectors). It can be observed that rather poor performance of kNN in the case of all three original sets has been significantly improved in the reduced feature space. The highest improvement has been achieved in the case of prosodic-spectral feature set (an increase from 39.9% to 91.3% average recognition rate), which could be explained by the fact that the performance of the kNN classifier is affected by the high dimensionality, which is particularly apparent in case of PS-FS.

Table 1 Recognition accuracy of kNN classifier using 3 feature sets (before and after feature reduction using LDA)

	Class recognition rate [%]					
Feature set	Anger	Fear	Joy	Neutral	Sadness	Average
P-FS	44.3	23.9	39.1	25	44.5	35.4
P-FS reduced with LDA	53.7	51.2	52.3	53.2	61.2	54.3
S-FS	73.9	56.9	35.1	58.1	37.1	52.2
S-FS reduced with LDA	81.3	92.8	81.6	95.7	93.9	89.1
PS-FS	57.8	37.9	32.2	23.6	33.3	39.9
PS-FS reduced with LDA	86.8	93.7	83.6	95.9	96.3	91.3

Table 2 shows the class and average recognition rate of the linear Bayes classifier in case of three feature sets, before LDA (initially extracted feature sets) and after LDA (original feature space reduced to 4 projection vectors). An improvement of recognition accuracy is obtained only in the case of prosodic feature set (P-FS). This improvement amounts to about 5%, which is a rather moderate increase compared to the results in Table 1, where the improvement is about 19%. As to S-FS and PS-FS there were no improvements, which is probably due to good linear separability in the original feature space (resulting in high recognition rates using non-reduced S-FS and PS-FS).

Table 2 Recognition accuracy obtained with 3 feature sets (before and after feature reduction using LDA) and with the linear Bayes classifier

	Class recognition rate [%]					
Feature set	Anger	Fear	Joy	Neutral	Sadness	Average
P-FS	51.4	43.7	46.8	45.4	62.4	49.9
P-FS reduced with LDA	51.4	53.4	46.8	56.6	69.8	55.6
S-FS	85.1	91.7	81	95.9	93.9	89.5
S-FS reduced with LDA	85.1	91.4	80.7	96.5	94.3	89.6
PS-FS	88.8	92.5	84.2	97.1	94.8	91.5
PS-FS reduced with LDA	88.2	92.5	85.3	95.9	95.7	91.5

The class and average recognition rate of the perceptron rule in two test conditions (3 feature sets before LDA and 3 feature sets after LDA) are given in Table 3. Slight improvements of recognition accuracy are noticeable in the case of all three reduced feature sets. The improvement is the lowest in case of P-FS.

When these three classifiers are compared, it can be noted that a substantial improvement of recognition accuracy has been achieved for the simplest classifier, namely kNN. Using the PS-FS reduced using LDA, kNN achieves the accuracy almost equal to the best result in our experiments (91.5%). This holds for the perceptron as a classifier, although the relative improvement of the average performance of the perceptron is much smaller.

Table 3 Recognition accuracy using 3 feature sets (before and after feature reduction using LDA) and with the perceptron rule as the classifier

	Class recognition rate [%]					
Feature set	Anger	Fear	Joy	Neutral	Sadness	Average
P-FS	34.8	29.3	36.2	21.3	56.9	35.7
P-FS reduced with LDA	16.9	33.1	42.2	33	62.6	37.6
S-FS	79.9	81.9	72.1	89.7	87.9	82.3
S-FS reduced with LDA	78.2	90.5	80.5	91.1	93.4	86.7
PS-FS	83.9	88.2	77.1	91.4	93.7	86.9
PS-FS reduced with LDA	86.8	94.2	82.8	93.9	94.5	90.5

Employing LDA, the original feature space is transformed to a new one, making it impossible to interpret the relevance of particular feature types. For an insight into the list of the most relevant features in the original (untransformed) feature space, SFFS (Sequential Forward Feature Selection) has been applied. The wrapper for SFFS is the linear Bayes classifier since it had the best recognition results. The number of selected features has been preset to 35. For the interpretation of results, three indicators have been used. The first indicator of the relevance of a feature type is the number (#) of the features selected by SFFS. The other two indicators are so called 'share' and 'portion', as described in [16]. With 'share', the count of the selected feature type is normalized by the total number of features in the reduced set (#/35 in our experiment). With 'portion', the same number is normalized by the cardinality of a feature type in the original feature set (#/#total). For each feature type, the 'share' indicator displays its percentage in modeling our 5-class problem, while the indicator 'portion' gives the percentage of the total number of the feature type which contributes to the modeling of the problem.

The results of the selection of 35 features from PS-FS and the effectiveness of each feature type are displayed through 3 indicators in Table 4. The observed feature types from PS-FS are: zero crossing rate (ZCR), energy, pitch (plus voicing probability) and MFCC. Columns '#Total' and '#' show the total number and the number of selected features per each feature type, respectively.

From Table 4 it can be observed that the most selected features ('share'=77.1%) belong to the MFCC type. The second important feature type is energy ('share'=11.4%). The third and the fourth feature type are ZCR and pitch, respectively.

As regards the indicator 'portion', the list of feature types can be arranged in the following way: from the total feature set energy is selected with the highest percentage (16.7%), followed by ZCR (12.5%). Although the MFCC feature type is the most frequent one in the selected feature set, only 9.4% of the total number of MFCC is selected. The pitch feature type is selected by the lowest rate (2.1%).

Table 4 Summary of feature selection results (35 features selected using SFFS), displayed with respect to feature types

	ZCR	Energy	Pitch	MFCC
#Total	24	24	48	288
SFFS				
#	3	4	1	27
share [%]	8.6	11.4	2.9	77.1
portion [%]	12.5	16.7	2.1	9.4

Table 5 summarizes the results of the feature selection distributed along groups of used statistical functionals: moments, extrema and regression coefficients. The features derived via moments are the most frequent among the selected features ('share'=57.1%), followed by the features derived via extrema (22.9%) and the features derived via linear regression (20%). Observing the 'portion' of the total number of features in each group of functionals, the most highly ranked are moments, followed by regression functionals and extrema, in that order.

Table 5 Summary of feature selection results, distributed along groups of used statistical functionals

	Moments	Extrema	Regression
#Total	128	160	96
SFFS			
#	20	8	7
share [%]	57.1	22.9	20
portion [%]	15.6	5	7.3

5. Conclusion

The paper gives an outline of a system for the recognition of basic emotions in speech, with particular emphasis on the extracted acoustic feature sets, classification schemes and emotional speech corpus. The paper discusses the obtained improvement of the recognition accuracy in a lower dimensional feature space obtained by applying Linear Discriminant Analysis. The most substantial improvement of the recognition accuracy has been achieved for the simplest classifier in our experiments, namely the kNN classifier. A combination of kNN with a reduced prosodic-spectral feature set nearly approaches the best results obtained in the experiments (the accuracy of 91.5%).

Feature selection algorithm has been employed in order to evaluate the relevance of the feature types and their statistical properties in the given task of the recognition of 5 basic emotions. In descending order of relevance, the features are: MFCC, energy, zero crossing rate and pitch. Observing the ratio of selected features to the total number of features in each feature type, features related to the energy are the most usually selected. The results of the feature selection distributed along groups of used statistical functionals imply that moments are the most relevant statistical features, although the extrema, regression coefficients and regression error also play notable roles.

Combining chosen prosodic and spectral features, represented by appropriate statistical features, even with a most simple classification scheme (such as kNN) the recognition results comparable with more complex systems can be achieved.

Acknowledgement: *The research presented in this paper has been carried out within the project "The development of dialogue systems for Serbian and other south Slavic languages" (TR32035), supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia.*

REFERENCES

[1] D.A. Sauter, F. Eisner, P. Ekman, S. Scott, "Crosscultural recognition of basic emotions through non-verbal emotional vocalizations", *Proceedings of National Academy of Sciences of the USA*, vol. 107(6), pp. 2408-2412, 2010.
[2] D. Ververidis, C. Kotropoulos, "Emotional speech recognition: Resources, features and methods", *Speech Communication*, vol. 48, pp. 1162-1181, 2006.
[3] S.L. Lutfi, F. Fernandez-Martinez, J.M. LucasCuesta, L. Lopez-Lebon, J.M. Montero, "A satisfaction-based model for affect recognition from conversational features in spoken dialog systems", *Speech Communication*, vol. 55, pp. 825-840, 2013.
[4] M.E. Ayadi, M.S. Kamel, F. Karray, "Survey on speech emotion recognition: Features, classification schemes and databases", *Pattern Recognition*, vol. 44, pp. 572-587, 2011.
[5] N. Fragopanagos, J.G. Taylor, "Emotion recognition in human-computer interaction", *Neural Networks*, vol 18, pp. 389-405, 2005.
[6] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth, "Acoustic emotion recognition: a benchmark comparison of performances", IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009, Italy, 2009, pp. 552-557.
[7] V. Delić, M. Bojanić, M. Gnjatović, M. Sečujski, S.T. Jovičić, "Discrimination capability of prosodic and spectral features for emotional speech recognition", *Electronics and Electrical Engineering*, Kaunas Technologija, vol. 18, no. 9, pp. 51-54, 2012.
[8] M. Bojanić, *Extraction and selection of feature set for automatic emotional speech recognition*. Ph.D. dissertation, Dept. Elect. Eng., Faculty of Technical Sciences, University of Novi Sad, 2013.
[9] B. Schüller, A. Batliner, S. Steidl, D. Seppi, "Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge", *Speech Communication*, vol. 53, pp. 1062-1087, 2011.
[10] C.M. Lee, S.S. Narayanan, "Toward detecting emotions in spoken dialogs", *IEEE Transactions Speech Audio Processing*, vol. 13, no. 2, pp. 293-303, 2005.
[11] H. Altun, G. Polat, "New frameworks to boost feature selection algorithms in emotion detection for improved human computer interaction", *LNCS*, vol. 4729, Berlin-Heidelberg: Springer, pp. 533-541, 2007.
[12] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, 2nd edition. Wiley, New York, 2000.
[13] S.T. Jovičić., Z. Kašić, M. Djordjević, M. Rajković, "Serbian emotional speech database: design, processing and evaluation", Proceedings of International Conference on Speech and Computer (SPECOM 2004), St Peterburg, 2004, pp.77–81.
[14] K. Fukunaga, Introduction to Statistical Pattern Recognition. Academic Press, 1990.
[15] P. Pudil, J. Novovicova, J. Kittler, "Floating search methods in feature selection", *Pattern Recognition Lett.*, vol. 15, pp. 1119-1125, 1994.
[16] A. Batliner et al., "Whodunnit – Searching for the most important feature types signalling emotion-related user states in speech", *Computer Speech and Language*, vol. 25, pp. 4-28, 2011.