Discrimination Capability of Prosodic and Spectral Features for Emotional Speech Recognition

V. Delic[1], M. Bojanic[1], M. Gnjatovic[1], M. Secujski[1], S. T. Jovicic[2]
[1]*Faculty of Technical Sciences, University of Novi Sad,*
Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia, phone: +381 21 485 25 33
[2]*School of Electrical Engineering, University of Belgrade,*
Bulevar Kralja Aleksandra 73, 11000 Beograd, Serbia, phone: +381 11 3292 000
milana.bojanic@uns.ac.rs

***Abstract*—The paper addresses the research question of automatic emotional speech recognition for Serbian. It integrates two research issues: (i) selection of an appropriate feature set, and (ii) investigation of different classification techniques. The paper reports a set of experiments with three feature sets: (i) the prosodic feature set, (ii) the spectral feature set, and (iii) the set of both spectral and prosodic features. The linear Bayes, the perceptron rule and the kNN classifier were considered in all three experiments. The experimental results show that the highest recognition accuracy of 91.5 % was obtained with the third feature set using the linear Bayes classifier.**

***Index Terms*—Emotional speech recognition, prosodic features, spectral features.**

I. INTRODUCTION

Recognition of emotional speech in human-machine interaction is a challenging task. Even in cases when users' emotional state does not lead to the introduction of additional lexical information (e.g., out-of-vocabulary words, etc.), changes in the acoustic features of affective speech may significantly degrade the accuracy of automatic speech recognition (ASR). Therefore, taking into account the changes in acoustic features that indicate emotion may substantially improve human-machine speech-based interfaces. This does not hold only from the aspect of ASR, but also from other functional aspects of dialogue systems (e.g., natural-sounding text-to-speech synthesis). In general, the ability to effectively recognize, track and appropriately respond to the user's emotional state is a crucial feature of emotion-aware human-machine interfaces.

Accurate emotion recognition has an important role in many speech-based applications [1]. For example, in the scope of customer care interactions (engaging a human operator or a conversational agent), emotional speech recognition (ESR) systems may be used to assess customers' satisfaction and quality of service. ESR may also be used for detecting miscommunication in the interaction between the user and automated information services. Furthermore, ESR is important for dialogue systems intended for therapeutic interaction [2], interactive educational systems [3], as well as domains of interaction such as affective attachment and engaging in social interaction.

The paper addresses the restricted research domain of automatic speech-based emotion recognition for Serbian. The presented research is focused on the examination of discrimination capability of a set of features for emotional speech recognition. A total of 384 features have been calculated over a set of 1740 utterances from the Corpus of Emotional and Attitude Expressive Speech ("*Govorna ekspresija emocija i stavova*", in further text: GEES) in Serbian [4]. We consider the five basic emotional states: anger, joy, fear, sadness, and neutral. The statistical properties of energy, pitch, and spectral features of emotional (i.e., non-neutral) speech have been tested and compared to neutral speech. At the methodological level, we integrate two research directions: (i) selection of an appropriate feature set, and (ii) investigation of different techniques for the classification of emotional speech.

The paper is organized as follows. A description of the GEES corpus is given in the next section. Feature extraction and methods for emotion classification are discussed in the following two sections. Finally, the experimental results are reported and discussed.

II. THE DESCRIPTION OF THE EMOTIONAL SPEECH CORPUS

The GEES corpus is the first corpus of emotional and attitude-expressive speech in Serbian, recorded for the purpose of research on emotions in the field of speech technologies. It contains recordings of acted speech-based emotional expressions. A group of six drama students (3 female, 3 male) was engaged to produce emotionally colored utterances, cf. [4]. The subjects were given a set of textual entries – 32 isolated words, 30 short sentences, 30 long sentences, and one passage of 79 words. They were asked to express each textual entry in five emotional states: anger,

Manuscript received March 12, 2012; accepted May 12, 2012.
The presented study is performed as part of the project "Development of Dialogue Systems for Serbian and Other South Slavic Languages" (TR32035), as well as projects III44008 and OI178027, funded by the Ministry of Education and Science of the Republic of Serbia.

joy, fear, sadness and neutral. With respect to lexical content, the entries were out-of-context and emotionally neutral. Therefore, for the subjects, prosody was the only means to express different emotional states.

The GEES corpus was recorded in an anechoic studio at the Faculty of Drama Arts, University of Arts in Belgrade, using a high quality microphone. The corpus was evaluated with respect to the perception of its emotional content. Thirty normal-hearing students participated in the perception test. They were asked to assign exactly one label from the given set *{anger, joy, fear, sadness, neutral}* to each of the recordings.

The results of the perception test showed that average correct identification of emotions was 95%, ranging from 93.33% (fear) to 96.06% (anger) [4]. A substantial level of agreement among the evaluators demonstrated that the corpus contains acoustic variations that are indicative of emotional expression of the five target emotional states.

III. FEATURE EXTRACTION

Since we apply the cepstral method to estimate the fundamental frequency, and use the Mel Frequency Cepstral Coefficients (MFCC) as the selected spectral features, the general procedure of signal preprocessing will be described. Each speech signal is preemphasized and windowed with 25 ms Hamming windows shifted every 10 ms. The Fast Fourier Transform is applied to find the frequency spectrum, and then the frequency axis is warped according to the mel scale. MFCCs are obtained by logarithming the result and applying the Discrete Cosine Transform on it. Only the first 12 coefficients are taken into account in our analyses, since they correspond to slow changes in the spectrum, i.e. the spectrum envelope. The first derivative of cepstral coefficients is estimated in order to model the dynamics of speech, since it carries information important for both speech and emotion recognition.

Additional prosodic features used for emotional speech classification are pitch and energy. Pitch (i.e., the fundamental frequency of phonation *F0*) is the vibration rate of vocal cords. The emotional state of the speaker affects the tension of vocal cords and the subglottal air pressure, which ultimately affects the pitch. Thus, many authors consider it to be the most important prosodic feature for ESR [5], [6]. The fundamental frequency is extracted by the autocorrelation and the cepstrum-based method, using the *openSMILE* toolkit [7]. The voicing probability of a segment is another relevant feature. The fundamental frequency is estimated only for a frame whose voicing probability is above a preset threshold, otherwise it is considered undefined. The root mean square energy is calculated for every frame. The pitch contour and the energy contour are, respectively, sequences of short-term pitch and energy values extracted on a frame basis. The pitch and energy features are finally obtained from these contours by applying so-called static modelling through functionals since they are often reported superior to dynamic classifiers like Hidden Markov Models [5], [7]. We use the following 12 functionals:

1) *Maximum value of the contour;*
2) *Minimum value of the contour;*
3) *Difference between the maximum and the minimum value (i.e., the range);*
4) *Relative position of the maximum value;*
5) *Relative position of the minimum value;*
6) *Arithmetic mean of the contour;*
7) *Slope of a linear approximation of the contour;*
8) *Offset of a linear approximation of the contour;*
9) *Mean squared error computed as the difference of the linear approximation of the contour and the actual contour;*
10) *Standard deviation of the values in the contour;*
11) *Skewness (3rd order moment);*
12) *Kurtosis (4th order moment).*

The GEES corpus contains 348 recorded utterances in each emotion class (with each speaker equally represented), which gives the total number of 1740 utterances used for the feature extraction. According to the studies in [5], [8], [9], the most frequently used acoustic features for ESR are: prosodic features (pitch, intensity, duration), cepstral features (MFCC), spectral features (formant position and bandwidth), and less frequently voice quality features (harmonic-to-noise ratio, jitter, shimmer). While in some studies ESR relies on prosodic and voice quality feature set only [8], and in other on cepstral features only [9], recently, statistical functionals applied on low-level descriptors resulted in very large feature vectors up to a few thousands of prosodic and spectral features, [5]. Our objective was to compare the discrimination capability of prosodic and spectral features used separately, as well as the discrimination capability of their combination.

Three sets of features, based on the twelve functionals, are compared in the experiments reported in the paper. The first feature set (FS1) includes prosodic features: 12 functionals applied on pitch and energy values produce 24 features for each utterance. The second feature set (FS2) includes spectral features: 12 MFCC, their first derivatives, and 12 functionals applied on all of them, which gives 288 features for each utterance. The third set of features (FS3) includes: the spectral features (12 MFCC), the pitch, the voicing probability, the energy, and the zero crossing rate. For these 16 features, the first derivative was estimated, and then 12 functionals were applied on all of them. This results in 384 features for each of the utterances.

IV. CLASSIFICATION TECHNIQUES

For the purpose of emotional speech classification, Linear Discriminant Classifiers (LDC) and k-Nearest Neighbours (kNN) are taken into account. LDCs and kNN classifiers have been used since the very first studies and turned out to be quite successful for both acted and spontaneous emotional speech [5]. We consider three clasifiers. The first classifier is the linear Bayes classifier with the underlying assumption that classes have Gaussian densities and equal covariance matrices. Maximum likelihood estimates of Gaussian density parameters are used. For the second classifier, no assumptions were made about the underlying densities, and linear discriminant functions were derived via

the perceptron rule. Finally, the kNN classifier is a very intuitive method that classifies unlabeled examples based on their similarity to examples in the training set. It implicitly involves non-parametric density estimation, which leads to very simple approximation of the Bayes classifier. In order to improve reliability and performance, and to obtain more efficient models (both in terms of processing speed and memory requirements), Sequential Forward Feature Selection (SFFS) algorithm and Linear Discriminant Analysis (LDA) as a linear feature extraction method were employed.

In every experiment we used 10-fold partition of the data set to estimate the recognition accuracy of a particular classifier for the given feature set.

V. RESULTS

We recall that our research integrates two research directions: (i) selection of a feature set, and (ii) investigation of different techniques for classification of emotional speech. Therefore, the research was designed as a set of experiments with three different feature sets: the first experiment with the feature set FS1 (prosodic features only), the second experiment with the feature set FS2 (spectral features only) and finally, the third experiment with the feature set FS3 (both spectral and prosodic features). In all experiments, the following classification techniques were considered: the linear Bayes classifier, the perceptron rule, and the kNN classifier.

The results obtained from the first experiment are summarized in Table I. It can be observed that the linear Bayes classifier outperformed the perceptron rule. The kNN classifier was tested for different values of neighbours (k), and the highest recognition accuracy was obtained for k=9. Still, even in that case, the kNN classifier was less accurate than the two aforementioned classifiers. However, when LDA was applied to reduce the feature dimension to 4, the recognition accuracy of the kNN classifier was improved, especially for the emotional state of fear and the neutral emotional state.

TABLE I. RECOGNITION ACCURACIES ON FEATURE SET FS1 (ESTIMATED THROUGH 10-FOLD CROSS-VALIDATION) FOR THE SUBJECT-INDEPENDENT TEST.

Emotion Classifier	Class Recognition Rate, in % (Feature Set 1)					
	Anger	Fear	Joy	Neutral	Sadness	Average
Linear Bayes	52.9	43.7	44.6	43.1	64.4	49.7
Perceptron rule	23.0	49.1	37.4	32.5	54.0	39.2
kNN (k=9)	43.1	26.7	36.8	23.9	41.7	34.4
kNN+LDA(4)	53.4	50.9	51.2	52.3	61.8	53.9

The results obtained from the second experiment are summarized in Table II. It can be noted that the recognition accuracies of the observed classifiers under this experimental setting are significantly higher than under the first experimental setting. This implies that the feature set FS2 provides better discrimination capability for the adopted 5-class emotion classification task than the feature set FS1. This observation is in line with the findings that the distribution of the spectral energy across the speech range of frequency is a possible measure of the emotional content of speech. The highest recognition accuracy of the kNN classifier under this condition was obtained for k=11.

TABLE II. RECOGNITION ACCURACIES ON FEATURE SET FS2 (ESTIMATED THROUGH 10-FOLD CROSS-VALIDATION) FOR THE SUBJECT-INDEPENDENT TEST.

Emotion Classifier	Class Recognition Rate, in % (Feature Set 2)					
	Anger	Fear	Joy	Neutral	Sadness	Average
Linear Bayes	85.9	91.1	79.6	95.7	94.3	89.3
Perceptron rule	77.9	80.2	75.9	89.4	86.2	81.9
kNN (k=11)	73.6	58.9	35.9	59.2	35.1	52.5

The results obtained from the third experiment are summarized in Table III. The recognition accuracies of the linear Bayes classifier and the perceptron rule in the third experiment are significantly higher than in the first experiment, and slightly higher than in the second experiment. However, even in the latter case, it may be considered to be a noticeable improvement if we keep in mind that the obtained recognition accuracy comes close to the human accuracy of 95%, as reported in [5]. The feature set FS3 provides the best discrimination capability for the adopted 5-class emotion classification task.

TABLE III. RECOGNITION ACCURACIES ON FEATURE SET FS3 (ESTIMATED THROUGH 10-FOLD CROSS-VALIDATION) FOR THE SUBJECT-INDEPENDENT TEST.

Emotions Classifier	Class Recognition Rate (Feature Set 3)					
	Anger	Fear	Joy	Neutral	Sadness	average
Linear Bayes	88.8	92.5	84.2	97.1	94.8	91.5
Perceptron rule	79.6	87.1	81.0	91.7	95.4	87.0
kNN (k=9)	56.6	37.1	33.3	25.9	33.1	37.2
kNN (k=9)+SFFS	53.1	30.7	41.1	41.3	65.8	46.4

On the other hand, the highest recognition accuracy of the kNN classifier in the third experiment was only 37.2%, obtained for k=9. This rather poor performance may be explained by the fact that the kNN classiffier was affected by the high dimensionality. To overcome this problem, the kNN classifier was tested using the best 35 features selected by the SFFS algorithm. The overall recognition accuracy increased to 46.4%.

Fig. 1 shows the average ESR rates for the three classifiers considered in this study. It can be observed that the linear Bayes and perceptron rule classifiers show better performance than the kNN classifier under all three experimental settings. The difference is the most noticeable in the third experiment (using the feature set FS3).

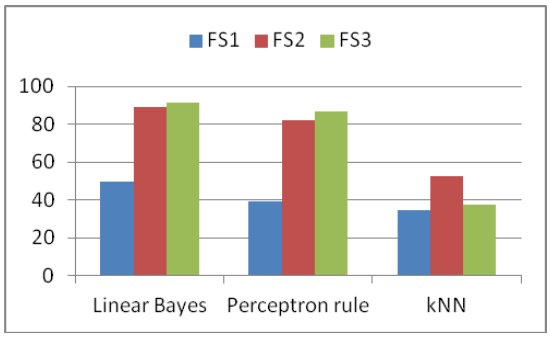

Fig. 1. Average emotion speech recognition rates for the three feature sets and the three classifiers (Bayes, perceptron, kNN).

Emotion class recognition rates of the linear Bayes classifier for the three feature sets are represented in Fig. 2. This classifier showed the best performance when both

spectral and prosodic features were used. The performance is slightly degraded in the case when only spectral features were used, but is still significantly better than in the case when only prosodic features were used.

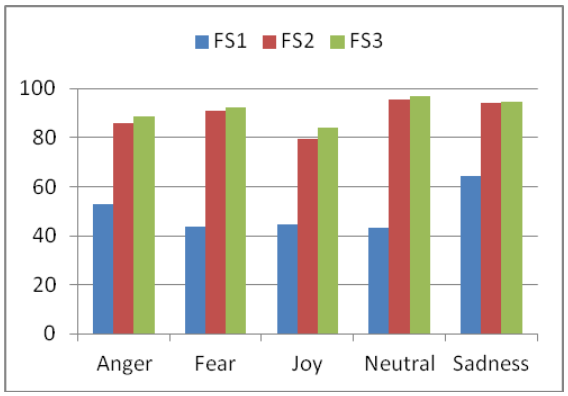

Fig. 2. Emotion class recognition rates of the linear Bayes classifier for the three feature sets.

VI. CONCLUSIONS

The reported research was designed as a set of experiments with three different feature sets. In the first experiment, the feature set included prosodic features only, while in the second experiment, it included spectral features only, and in the third experiment, it included both spectral and prosodic features. The experimental results showed that the highest recognition accuracy was obtained in the third experiment. However, a relatively small difference in the performances between the second and the third experiment indicates that prosodic and spectral features are highly correlated. Among the observed classifiers, the linear Bayes classifier showed the best performance. Its overall recognition rate in subject and gender independent tests was 91.5 %.

Future research directions include the implementation of emotion recognition within ASR modules, as well as integration with speaker recognition [11].

REFERENCES

[1] M. Bojanić, V. Delić, "Automatic Emotion Recognition in Speech: Possibilities and Significance", *Electronics*, Faculty of Electrical Engineering, University of Banjaluka, vol. 13, no. 2, pp. 35–40, 2009.

[2] O. A. Schipor, S. G. Pentiuc, M. D. Schipor, "The Utilization of Feedback and Emotion Recognition in Computer based Speech Therapy System", *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering)*, no. 3, pp. 101–104, 2011.

[3] C. C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan, "Emotion Recognition Using a Hierarchical Decision Tree Approach", *Speech Communication*, Netherlands: Elsevier BV, no. 53, pp. 1162–1171, 2011.

[4] S. T. Jovičić, Z. Kasić, M. Djordjević, M. Rajković, "Serbian emotional speech database: design, processing and evaluation", in *Proc. of the SPECOM 2004*, St. Peterburg, 2004, pp. 77–81.

[5] 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*, Netherlands: Elsevier BV, no. 53, pp. 1062–1087, 2011.

[6] Y. Li, Y. Zhao, "Recognizing emotions in speech using long-term and short-term features", in *Proc. of the ICSLP 1998*, Sydney, Australia, 1998, pp. 2255–2258.

[7] F. Eyben, M. Woellmer, B. Schüller, "OpenSmile – the munich versatile and fast open-source audio feature extractor", in *Proc. of the ACM Multimedia 2010*, Florence, Italy, 2010, pp. 1459–1462.

[8] R. Fernandez, R. Picard, "Recognizing affect from speech prosody using hierarchical graphical models", *Speech Communication*, Netherlands: Elsevier BV, no. 53, pp. 1088–1103, 2011.

[9] T. Nwe, S. Foo, L. De Silva, "Speech emotion recognition using hidden Markov models", *Speech Communication*, Netherlands: Elsevier BV, no. 41, pp. 603–623, 2003.