AUTOMATIC RECOGNITION OF EMOTIONS FROM SPEECH

Martin Gjoreski[1], *Hristijan Gjoreski*[2], *Andrea Kulakov*[1]
[1]Faculty of Computer Science and Engineering, Rugjer Boshkovikj 16, 1000 Skopje, Macedonia;
[2]Department of Intelligent Systems, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
e-mail: martin.gjoreski@gmail.com, hristijan.gjoreski@ijs.si, andrea.kulakov@finki.ukim.mk

ABSTRACT

This paper presents an approach to recognition of human emotions from speech. Seven emotions are recognized: anger, fear, sadness, happiness, boredom, disgust and neutral. The approach is applied on a speech database, which consists of simulated and annotated utterances. First, numerical features are extracted from the sound database by using audio feature extractor. Next, the extracted features are standardized. Then, feature selection methods are used to select the most relevant features. Finally, a classification model is trained to recognize the emotions. Three classification algorithms are tested, with SVM yielding the highest accuracy of 89% and 82% using the 10 fold cross-validation and Leave-One-Speaker-Out techniques, respectively. "Sadness" is the emotion which is recognized with highest accuracy.

1 INTRODUCTION

Human capabilities for perception, adaptation and learning about the surroundings are often three main compounds of the definition about what intelligent behavior is. In the last few decades there are many studies suggesting that one very important compound is left out of this definition about intelligent behavior. That compound is emotional intelligence. Emotional intelligence is the ability of one to feel, express, regulate his own, to recognize and handle the emotional state of others. In psychology the emotional state is defined as complex state that results in psychological and physiological changes that influence our behaving and thinking [1].

With the recent advancements of the technology and the growing research areas like machine learning, audio processing and speech processing, the emotional states will be inevitable part of the human-computer interaction. There are more and more studies that are working on providing the computers with abilities like recognizing, interpretation and simulation of emotional states.

In this research we present an approach for automatic recognition of emotions from speech. The goal is to recognize the emotional state that is experiencing the speaker. Furthermore, the focus is on how something is said, and not what is said. Besides this approach where only the speaker's voice is analyzed, there are several different approaches for recognizing the emotional state. In some approaches the voice and the spoken words are analyzed [2]. Some are focused only on the facial expressions [3]. Some are analyzing the reactions in the human brain for different emotional states [4]. Also there are combined approaches where combination of the mentioned approaches is used [5].

In studies where human emotions are analyzed mainly two methodologies are used. In the first methodology the emotions are viewed as discrete and completely distinct classes that are universally recognized [6]. In the second methodology the emotional states are represented in 2D or 3D space where parameters like emotional distance, level of activeness, level of dominance and level of pleasure can be observed [7]. In this research the discrete methodology will be used, so the emotional states will be represented as 7 classes: anger, fear, sadness, happiness, boredom, disgust and neutral.

The remainder of this paper is organized as follows. Next section is a brief overview of speech emotion analysis. Then, the methodology used for the process of emotion classification is presented. In the next section, the experimental setup and the results are presented. Finally, the conclusion and a brief discussion about the results is given.

2 SPEECH EMOTION ANALYSIS

Speech emotion analysis refers to usage of methods to extract vocal cues from speech as a marker for emotional state, mood or stress. The main assumption is that there are objectively measureable cues that can be used for predicting the emotional state of the speaker. This assumption is quite reasonable since the emotional states arouse physiological reactions that affect the process of speech production. For example, the emotional state of fear usually initiates rapid heartbeat, rapid breathing, sweating and muscle tension. As a result of these physiological activities there are changes in the vibration of the vocal folds and the shape of the vocal tract. All of this affects the vocal characteristics of the speech which allows to the listener to recognize the emotional state that the speaker is experiencing [8]. The basic speech audio features that are used for speech emotion recognition are: fundamental frequency (human perception for fundamental frequency is pith), power, intensity (human perception for intensity is loudness), duration features (ex. rate of speaking) and vocal perturbations. The main question is: Are there any objective voice feature profiles that can be used for speaker emotion recognition? A lot

studies are done for the sake of providing such feature profiles that can be used for representation of the emotions, but not always the results are consistent. For some basic problems like distinguishing normal speech from angry speech or distinguishing normal speech from bored speech the experimental results converge [9]. The problem arises when we have to distinguish emotional states like anger from happiness or fear from happiness. By using the basic speech audio features for describing these emotional states, the feature profiles will be quite similar so distinguishing them is hard.

In the last few years, new method is introduced where static feature vectors are obtained by using so called acoustic Low-Level Descriptors (LLDs) and descriptive statistical functionals [10]. By using this approach a big number of large feature vectors is obtained. The downside is that not all of the feature vectors are of good value, especially not for emotion recognition. For that reason a feature selection method is often used.

3 THE APPROACH

Figure 1 shows the whole process of the speech emotion classification used in this research. An emotional speech database is used, which consists of simulated and annotated utterances. Next, feature extraction is performed by using open source feature extractor. Then, the extracted features are standardized. After standardization, feature selection methods are used for decreasing the number of features and selecting only the most relevant ones. Finally, the emotion recognition is performed by a classification model.

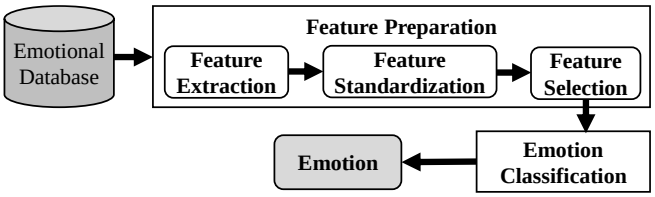

Figure 1: *Scheme for speech emotion classification.*

3.1 Emotional Database

There are several emotional speech databases that are extensively used in the literature [11]: German, English, Japanese, Spanish, Chinese, Russian, Dutch etc. One of the main characteristics of an emotional speech database is the type of the speech: whether it is simulated or it is extracted from real life situations. The advantage of having a simulated speech is that the researcher has a complete control over the emotion that it is expressed and complete control over the quality of the audio. However, the disadvantage is that there is loss in the level of naturalness and spontaneity. On the other hand, the non-simulated emotional databases consist of a speech that is extracted from real life scenarios like call-centers, interviews, meetings, movies, short videos and similar situations where the naturalness and spontaneity is kept. The disadvantage is that in these databases there is not a complete control over the expressed emotions. Also the low quality of the audio can be problem.

For this research the Berlin emotional speech database [12] is used. It consists of 535 audio files, where 10 actors (5 male and 5 female) are pronouncing 10 sentences (5 short and 5 long). The sentences are chosen so that all 7 emotions that we are analyzing can be expressed. The database is additionally checked for naturalness by a human expert. The utterances that were rated with more than 60% naturalness and from which the expressed emotion was recognized with more than 80%, were included in the final database.

3.2 Feature Preparation

The feature extractor tool used in this research is openSMILE (Open Speech and Music Interpretation by Large Space Extraction) [13]. It is a tool for signal processing and machine learning. We extracted 1582 features in total [14]. The LLDs that openSMILE is using are computed from basic features (pitch, loudness, voice quality) or representations of the audio signal (cepstrum, linear predictive coding).

On these LLDs functionals are applied and static feature vectors are produced, therefore static classifiers can be used. The functionals that are applied are: extremes (position of mix/min value), statistical moments (first to forth), percentiles (ex. the first quartile), duration (ex. percentage of time the signal is above threshold) and regression (ex. the offset of a linear approximation of the contour).

After the feature extraction the feature vectors are standardized so the distribution of the values of the feature vectors is with mean equal to 0 and standard deviation equal to 1. Next, a method for feature selection is used. Features are ranked with algorithms for feature ranking and experiments are done with varying number of top ranked features. For ranking the features two different algorithms are used, gain ratio [15] and ReliefF [16]. Both algorithms are used as they are implemented in Orange software packet for machine learning and data mining [17].

3.3 Emotion Classification

Once the features are extracted, selected and standardized, they are used to form the feature vector database. That is a database in which each data sample is an instance, i.e., feature vector. Additionally, each instance is labeled with the emotion. After this the instances are used to train a classification model in order to recognize emotions out of a speech data.

4 EXPERIMENTS

Three types of experiments are performed. In the first type, tests for comparison of three classification algorithms are done. The algorithm with the highest accuracy is further evaluated with 2 evaluation techniques: 10 fold cross-

validation and Leave-One-Speaker-Out (LOSO) cross-validation.

4.1 Comparison of Classification Algorithms

Three classification algorithms are compared: KNN [18], SVM [19] and Naïve Bayes [20]. They are used as implemented in the Orange machine learning toolkit. The data is split 70-30, i.e., 70% of the data is used as training, and the remaining 30% is used for testing. Tests are performed with varying number (50, 100, 200, 300, 400, 500, 750, 1000 and 1582) of top ranked features by gain ratio. The results (shown in Figure 2) show that the SVM has the highest accuracy, i.e., 91% when the top ranked 500 features are used. By using the top ranked 300 features the drops to 88%.

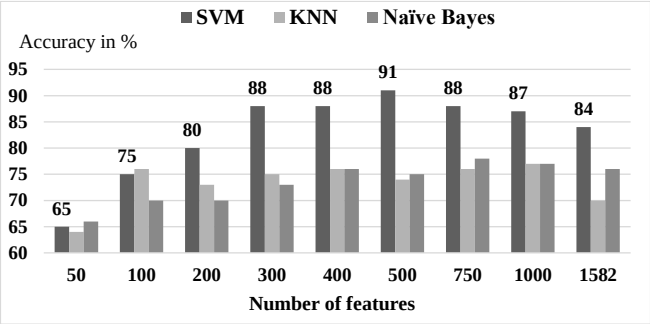

Figure 2: *SVM, KNN and Naïve Bayes classification accuracy for varying number of features.*

4.2 10 Fold Cross-Validation

We further evaluated the SVM with the 10 fold cross-validation technique. The results are shown in Figure 3. The highest accuracy of 89% is obtained by using top ranked 750 features. By using the top ranked 300 features the average accuracy is 87%, which is significantly high performance with such a low number of features.

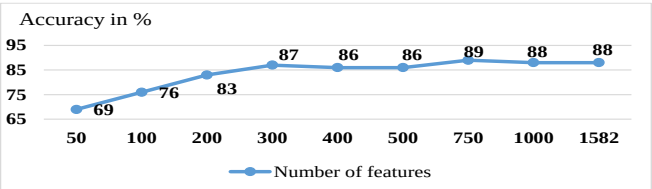

Figure 3: *SVM classification accuracy for 10 fold cross-validation with varying number of features.*

Additional analysis of the performance is performed by analyzing the recognition results for each emotion individually. The results achieved for the top ranked 750 features are shown in Figure 4. The highest accuracy per class is achieved for the class "sadness" (97%). On the contrary, the lowest accuracy per class is achieved for the class "happiness" (68%).

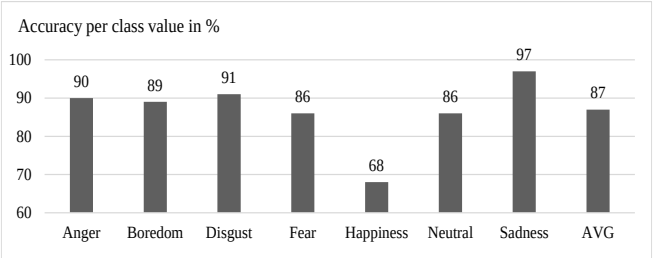

Figure 4: *SVM accuracy per class for 10 fold cross-validation with top ranked 750 features.*

4.3 Leave-One-Speaker-Out Cross-Validation

If the system for speech emotion recognition is supposed to work in an environment where it does not have any information about the speaker, LOSO is the best approach for testing the accuracy of the system.

The LOSO validation approach means that the train data consists of 9 speakers and the remaining one is used for testing. This is repeated 10 times, each time using different speaker's data for testing. Figure 5 shows the results that are obtained with the LOSO technique. The testing speaker is represented on the x-axis. The varying color represents the number of top ranked features (by ReliefF) used. The highest average accuracy of 82% is obtained by using top ranked 1000 features. Also we can see that the accuracy depends mainly from the speaker that is used as test data.

For the experiments about the accuracy per class for each of the 7 emotional states, top ranked 1000 features (by ReliefF) are used. The results are shown in Figure 6. The highest accuracy per class of 94% was achieved for the class "sadness" and the lowest accuracy per class of 70% was achieved for the class "fear".

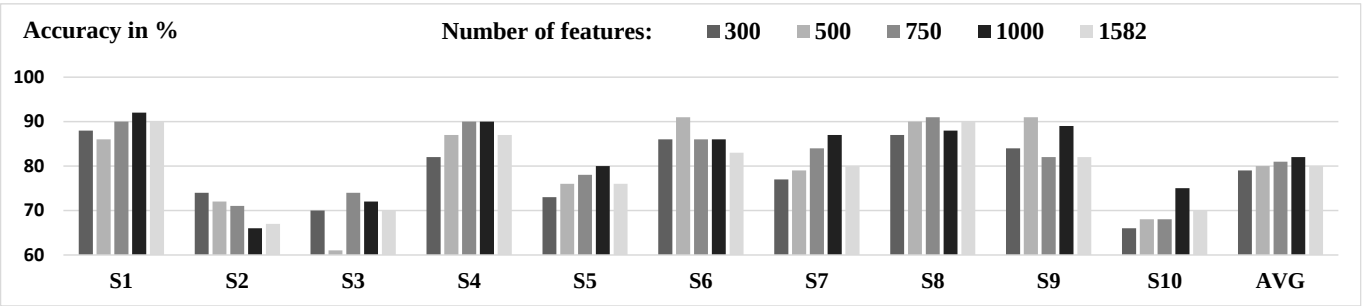

Figure 5: *SVM classification accuracy for LOSO with varying number of features*

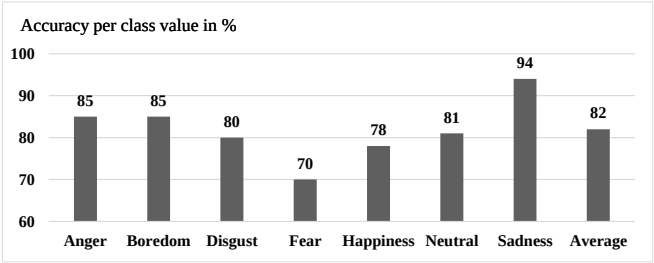

Figure 6: *SVM accuracy per class value for Leave-One-Speaker-Out cross-validation with top ranked 1000 features.*

5 CONCLUSION

The results showed that SVM outperforms the KNN and Naïve Bayes. By using the top ranked 500 features by gain ratio, SVM achieved the highest accuracy of 91%.

In addition, the 10 fold cross-validation of the SVM showed that highest accuracy of 89% was achieved by using the top 750 ranked features. By using the top 300 ranked features the accuracy was 87%. This is the so-called "knee" on the graph, which represents the best tradeoff between the number of features and the achieved performance.

Regarding the accuracy for each of the 7 emotions, experiments were performed with the top ranked 750 features by gain ratio. The best recognized emotion was the "sadness", with 97%; and the worst recognized emotion was the "happiness" with 68% accuracy.

With LOSO cross-validation, the SVM achieved highest accuracy of 82% by using the top 1000 ranked features. By using the top 500 ranked features the accuracy was 80%. Regarding the accuracy per emotion, experiments were performed with the top ranked 1000 features. The highest accuracy per class (emotion) of 94% was achieved for the class "sadness" and the lowest for the class "fear" 70%.

The results showed that the classifier achieves better accuracy with the 10 fold cross-validation technique compared to the LOSO validation technique. The reason for this is that with the 10 fold cross-validation the training and the testing data usually contain data samples of the same speaker. This is not the case if the system is intended to be used in real life for users not known in advance. However, a hybrid approach that includes a calibration phase at the beginning (for example asking the user to record several data samples) is considered for future work.

References

[1] D. G. Myers. Theories of Emotion. Psychology: Seventh Edition. New York NY: Worth Publishers. 2004.

[2] V. Perez-Rosas, R. Mihalcea. Sentiment Analysis of Online Spoken Reviews. Interspeech, 2013.

[3] A. Halder, A. Konar, R. Mandal, A. Chakraborty. General and Interval Type-2 Fuzzy Face-Space Approach to Emotion Recognition. IEEE Transactions on Systems, Man, and Cybernetics,43 (3), 2013.

[4] R. Horlings, D. Datcu, L. J. M. Rothkrantz. Emotion recognition using brain activity. Proceeding CompSysTech '08 Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing, 2008.

[5] A. Metallinou, S. Lee, S. Narayanan. Audio-Visual Emotion Recognition Using Gaussian Mixture Models for Face and Voice. Multimedia. 2008. ISM 2008. IEEE International Symposium on Multimedia, 2008.

[6] P. Ekman. Emotions in the Human Faces. 1982.

[7] James A. Russell. A circumplex model of affect. 1980.

[8] P. N. Juslin, K. R. Scherer. Vocal expression of affect. In J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.). The new handbook of methods in nonverbal behavior research, pp. 65-135, 2004.

[9] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. Speech Communication 40: 227–256. 2003

[10] M. E. Mena. Emotion Recognition From Speech Signals, 2012.

[11] D. Ververidis, C. Kotropoulos. A review of emotional speech databases. In: PCI 2003. 9th Panhellenic Conference on Informatics., pp. 560–574, 2003.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss. A Database of German Emotional Speech. 2005. In: Proc. Interspeech. pp. 1517–1520.

[13] F. Eyben, M. Wöllmer, B. Schuller. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. 2010.

[14] F. Eyben, F. Weninger, M. Wollmer, Bjorn Schuller. openSmile Documentation. Version 2.0.0., 2013.

[15] H. Deng, G. Runger, E. Tuv. Bias of importance measures for multi-valued attributes and solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN2011). 2011

[16] I. Kononenko, E. Simec, M. Robnik-Sikonja. Overcoming the myopia of inductive learning algorithms with RELIEFF. Applied Intelligence, Forthcoming.

[17] J. Demšar, B. Zupan. Orange: From experimental machine learning to interactive data mining. White Paper (http://www.ailab.si/orange). Faculty of Computer and Information Science. University of Ljubljana.

[18] D. Aha, D. Kibler. Instance-based learning algorithms. 1991, Machine Learning. 6:37-66.

[19] N. Cristianini, J. Shawe-Taylor. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000.

[20] R. Stuart, N. Peter. Artificial Intelligence: A Modern Approach. Second Edition, Prentice Hall.