Efficient Recognition of Human Emotional States from Audio Signals

Ernur Sonat Erdem  
Department of Computer Engineering  
Başkent University  
Ankara, Turkey  
sonaterdem@gmail.com

Mustafa Sert  
Department of Computer Engineering  
Başkent University  
Ankara, Turkey  
msert@baskent.edu.tr

Abstract—Automatic recognition of human emotional states is an important task for efficient human-machine communication. Most of existing works focus on the recognition of emotional states using audio signals alone, visual signals alone, or both. Here we propose empirical methods for feature extraction and classifier optimization that consider the temporal aspects of audio signals and introduce our framework to efficiently recognize human emotional states from audio signals. The framework is based on the prediction of input audio clips that are described using representative low-level features. In the experiments, seven (7) discrete emotional states (anger, fear, boredom, disgust, happiness, sadness, and neutral) from EmoDB dataset, are recognized and tested based on nineteen (19) audio features (15 standalone, 4 joint) by using the Support Vector Machine (SVM) classifier. Extensive experiments have been conducted to demonstrate the effect of feature extraction and classifier optimization methods to the recognition accuracy of the emotional states. Our experiments show that, feature extraction and classifier optimization procedures lead to significant improvement of over 11% in emotion recognition. As a result, the overall recognition accuracy achieved for seven emotions in the EmoDB dataset is 83.33% compared to the baseline accuracy of 72.22%.

Keywords—audio based emotion recognition; affective computing; MPEG-7 audio; MFCC; Support Vector Machine

I. INTRODUCTION

Machine analysis and development of automated methods for the recognition of human emotions is an important research field in affective computing. The problem has been researched in various disciplines such as psychology, cognitive science, computer science, and neuroscience. Although the problem has been well studied in the last two decades, most of the past research in machine analysis of human emotion has focused on the visual part and relatively less effort has been paid to the audio modality; specifically vocal analysis of audio signal. Hence, an increased number of studies on audio-based human emotion recognition have emerged in recent years [1]–[5], [8]–[10]. The task of recognizing emotions from audio signals poses several challenges. The first one is to specify robust acoustic features, which are correlated with emotion expressions. The second one is to find optimal parameters of the machine learning algorithm for the classification. And probably the most important one is to decide analysis method of audio segments, i.e., specifying the appropriate window and hope sizes in the feature extraction to reveal vocal characteristics of emotions. In one of the major studies on vocal emotion recognition, Tawari and Trivedi [1] introduce a novel set of features based on cepstrum analysis of pitch and intensity contours. They used the Support Vector Machines (SVM) classifier along with the Sequential Minimal Optimization (SMO) algorithm as the classifier and obtained 84% accuracy on the EmoDB. However, the system needs gender information as an input to reach the reported accuracy. In [10], Yang and Chen represent an emotion as a point in the Cartesian space with valence and arousal as the dimensions and determine the coordinates of a song by the relative emotion of the song with respect to other songs. They reach 0.326 in Gamma statistic performance for valence recognition. Another motivation in vocal emotion recognition is to optimize classifier parameters. Such a study by Li et al. propose a heuristic search algorithm to reduce the selection time of optimal SVM parameters. Wang and Guan introduce a multimodal approach to recognize human emotions from audiovisual signals [6]. They represent audio characteristics with prosodic, formant, and spectral (Mel Frequency Cepstral Coefficient (MFCC)) features and make use of Gabor features for the visual part. The authors propose a multi-classifier scheme for the categorization and obtain a recognition rate of 82.14%. However, the multi-classifier approach introduces some degree of computational complexity to the system.

In this study, we introduce an automated framework for the recognition of seven emotional states (anger, fear, boredom, disgust, happiness, sadness, and neutral) from audio signals and propose empirical methods for feature extraction and classifier optimization to capture the vocal characteristics of these emotions. The framework is discriminative in that a set of model parameters are learned and optimized using EmoDB, yet adaptive to be used with different datasets. The learning model is built upon SVM and the best representative feature (MFCC), which is obtained as a result of extensive experiments using nineteen standalone and joint features, is fed into the classifier.

The paper is organized as follows: In Section 2, the proposed emotion recognition framework is introduced. Experimental evaluation and recognition results are presented...
in Section 3. Finally, concluding remarks and future research directions are given in Section 4.

II. PROPOSED FRAMEWORK

The general steps of the proposed framework are depicted in Figure 1. We describe these steps in the following sections.

A. Preprocessing and Temporal Segmentation

This stage involves with the preprocessing and the temporal segmentation procedures. In the preprocessing stage, the sampling rate, bit-depth, and the number of channels of an input audio signal are converted to 16 kHz, 16 bits, and to mono signal for the analysis, respectively. The common approach in audio analysis tasks is the integration of segment-level features over some period of time in order to comprise a single feature vector and feed into a machine learning algorithm for categorization. The typical approach consists of applying standard window and hop (shift) sizes for the analysis. Then features are extracted from each of these windows. However, the duration of desired information in an audio signal depends on the application. For instance, the durations of laughter and gun-shot events in an audio signal may be quite different. Therefore, proper selection of analysis window for desired events effects the robustness of feature representation. Figure 2 illustrates the relation between the window and the hop sizes. Assume that each interval is 1 millisecond (ms) long and we define the window and the hop sizes as 5ms and 3ms, respectively. Let $E_i$ is an audio event we desire to recognize and $w_i$ represents an analysis window. Here, the window size of 5ms is adequate to capture the audio events $E_1$ and $E_2$, but the $E_3$, since the selected window size is not long as much as $E_2$ to capture its characteristics. If we consider the selected hop size, which is 3ms, we may not be able to capture the characteristics of $E_2$, since its duration is less than the hop size granularity. To overcome this problem, we perform empirical analysis to define the correct size and combinations of these parameters.

B. Feature Extraction and Representation

In order to capture the characteristics of emotional states in audio signals, we make use of nineteen audio features (15 standalone and 4 joint) in the experiments. Among these, ten spectral and prosodic features are selected from MPEG-7 audio family [7]. The MFCC feature, which is closest to the human perception, is selected due to success in automatic speech recognition applications [5], [9], [13]. The rest of the standalone features are Linear Predictive Coding (LPC), Zero Crossing Rate (ZCR), Fundamental Frequency (F0), and Matching Pursuit (MP). In addition, four different joint features are also used in the experiments. To determine efficient combinations among these, we concatenated a robust spectral feature (e.g., MFCC, MPEG-7 ASF) with the prosodic features (MPEG-7 ASS and MPEG-7 ASC) to enhance the expressive capability of the feature. We decide robust spectral features by examining the confusion matrix of each standalone feature as described in [13]. The feature sets are extracted from each predefined analysis window of an audio clip and represented with $n$-dimensional vectors, where $n$ is the number of analysis frames in the processed audio clip. To feed such $n$-dimensional vectors into the classifier, we compute mean and standard deviation of each vector that represent an analysis window. This operation, together with the selection of adequate window and hop sizes are also useful for reducing the computational/memory complexity.

Let the dimension of a feature vector $F$ is $n \times m$, where $n$ is the number of feature vectors and $m$ is the number of feature coefficients as illustrated in eq. (1). If we extract 13-coefficient MFCC feature from an audio clip of length 20sec, using a window size of 30ms, a hop size of 10ms, then the size $F$ is determined as $n = 20_{sec}/10_{ms} = 2000$; and the size of $F$ is $2000 \times 13$.

$$F = \begin{bmatrix} F_{1,1} & \cdots & F_{1,m} \\ \vdots & \ddots & \vdots \\ F_{n,1} & \cdots & F_{n,m} \end{bmatrix}_{n \times m} \quad (1)$$

If we store each sample with 2-bytes, then we need $\approx 52kB$ of memory for a 20sec long audio clip. On one hand, it may be necessary to calculate feature vectors using small hop size (e.g., 10ms) for a reliable detection. On the other hand, we may not need such a small granularity of 10ms for the detection of audio events. In order to define the window and hop sizes adequately, we carry out empirical analysis over the training dataset and pick the window and hop size combination based on the highest recognition accuracy.

C. Classifier Design

The classification of emotions from audio clips is performed using the SVM due to the success in pattern recogni-
tion applications. The inputs of an SVM in training phase are n-dimensional feature vectors, which represent the emotional states. Since SVM is a binary classifier and we have seven emotions to recognize, a separate SVM model is constructed for each emotional state. We use the Radial Basis Function (RBF) as the kernel trick scheme. In this high-dimensional feature space a linear classifier is built to perform predictions by using the LibSVM library [11]. In the prediction phase, one-versus-all (ova) approach is used for the multi-class problem. In this approach, the decision is made by a winner-takes-all-strategy, i.e., the classifier with the highest output function defines the category. For testing, we used k-fold cross validation method, where k is set to 10. In the study, we considered to optimize four different parameters, namely penalty constant (C), gamma (γ), tolerance of termination criterion epsilon (ε), and the kernel coefficient (r).

III. Experimental Results and Evaluation

We used the well known Berlin database of emotional speech (EmoDB) for performance evaluations [12]. In order to evaluate the framework, we perform three experiments in order. In the first experiment, we optimize the SVM parameters using the 10-fold cross validation method and the MFCC as the feature. In the second step, we experiment different window and hop sizes to increase the recognition rate. Thirdly, we evaluate nineteen feature sets with the SVM classifier to measure their expressive power.

1) SVM Parameter Optimization: We start with the default SVM parameters, which are C=1, r=0, ε = 0.001, and γ = 1/7 and apply standard window (30ms) and hop (10ms) sizes and using the MFCC feature. As a result we achieve an average recognition accuracy of 72.22% with the default values. After performing cross-validation we obtain the maximum accuracy with the parameters C=20, γ = 0.2, ε = 0.04, and r=1. Therefore, the best values of the SVM parameters are picked as C=20, γ = 0.2, ε = 0.04, and r=1 as depicted in Table I.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Recognition Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C=10</td>
<td>80.0010</td>
</tr>
<tr>
<td>C=20</td>
<td>82.0510</td>
</tr>
<tr>
<td>ε=0.02, r=0</td>
<td>82.2642</td>
</tr>
<tr>
<td>ε=0.04, r=1</td>
<td>83.1481</td>
</tr>
<tr>
<td>ε=0.05, r=1</td>
<td>82.6205</td>
</tr>
<tr>
<td>ε=0.10, r=1</td>
<td>82.2432</td>
</tr>
</tbody>
</table>

2) Window and hop size analysis: In order to define the window and hop sizes adequately, we experiment different window and hop sizes to obtain best accuracy using the MFCC feature with the SVM classifier. We start with a window size of 10ms and hop size of 5ms. This setup provides an average recognition accuracy of 84.10%. When we increase hop size to 10ms and leave the window size as is, the recognition accuracy decrease to 81.86%. This result show that non-overlapping analysis window decrease the recognition accuracy. If we set up a window size of 20ms with 10ms overlap, the average recognition accuracy indeed attains to its highest value of 84.29%. It is obvious that when we increase the window size and also increase the hop size, the average recognition accuracy is decreasing compared with the initial window and hop sizes (Figure 3).

Figure 3. Performance of different window and hop sizes.

This may happen due to length of analysis window, i.e., longer window sizes cannot capture emotion characteristic as much as the small ones. To sum up, the recognition accuracy is higher when the window size is between 10ms and 20ms, but decreasing for window sizes greater than 30ms.

3) Feature analysis: Figure 4 shows the results of the selected standalone and joint feature performances for emotion recognition task on the EmoDB dataset. It is clear that, the standalone and scalar features (e.g., ASC, ASS, HSC, HSD, HSS, HSV, SC, ZCR, and F0) cannot provide enough contribution for our setup. Despite the ASP feature is represented as a vector, it also gives a poor performance similar to the standalone features. The MFCC feature set provides the second highest average recognition accuracy of 83.15% among the others. This result is reasonable since the MFCC feature is close to human perception and robust in capturing the short-term power spectrum of a sound. Other standalone features, such as MPEG-7 ASF, MP, MPEG-7 ASE, and LPC are also demonstrate good performances compared with the MFCC (Fig. 4). When considered the combination of these standalone features with prosodic features (ASS and ASC), we obtain the highest average recognition accuracy of 83.33% with the MFCC+ASS+ASC combination.

The confusion matrix of the MFCC feature is given in Table II. Each column of the matrix represents the instances in a predicted class, whereas the rows represent the instances
in an actual class. All correct guesses are located in the diagonal of the table and are marked by asterisks. The rest of the confusion matrices are not included in the paper due to the space limitation. The MFCC feature is robust in almost every emotion state for many the improvement is very significant after optimization, such as boredom (↑ 7), sadness (↑ 7), anxiety/fear (↑ 6), etc. An interesting result is that, the recognition accuracy of happiness decreases (↓ 3). It is mostly confused with anger. Similarly, the neutral state is mostly confused with the boredom state. We believe that it is due to the posed audio clips in the dataset. We see that the ASF feature gives the second highest recognition accuracy of 64.15% among the standalone features. On the other hand, the performance of the joint feature, namely MFCC+ASS+ASC slightly outperforms the MFCC feature. Despite the enhancement, we can ignore the joint feature due to the introduced complexity at the feature extraction stage. As a result, the three states, namely anger, anxiety/fear, and sadness have similar characteristics and make the recognition task difficult for all features using the SVM.

IV. CONCLUSION

This paper introduces an automated framework for the recognition of seven emotional states (anger, fear, boredom, disgust, happiness, sadness, and neutral) from audio signals and proposes empirical methods for feature extraction and classifier optimization to capture the vocal characteristics of these emotions. In the experiments, discrete emotional states from EmoDB dataset, are recognized and tested based on 19 audio features (15 standalone, 4 joint) by using the SVM classifier. Our experiments show that, the methods introduced in the paper provide significant improvement of over 11% in emotion recognition. Additionally, the MFCC, ASF, LPC, and MP features can be used interchangeably in audio-based emotion recognition tasks, since they are quite successful in capturing the characteristics of emotions.

Our future work aims to adapt this framework to handle continuous emotions from audio signals.

REFERENCES


