ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.7

Recognition and Classification of Human Emotion from Audio

Anuja Pawar
ME Student
Department of Computer Engineering
D. Y. Patil college of Engineering, pune, India

Abstract:
In this paper, the audio emotion recognition system is proposed that uses a mixture of rule-based and machine learning systems to increase the efficacy in the recognition of audio paths. The audio path is intended using a mixture of input prosodic features (pitch, zero crossing rates (ZCR), log- energy and Teager energy operator (TEO)) and spectral features (i.e. Mel-scale frequency cepstral coefficients). Mel-Frequency Cepstral Coefficients (MFCC) feature extraction method used for speech feature extraction as well as existing research goals to identify performance enhancement parameters. After the MFCC feature extraction, these features are passed to three parallel sub- paths which use feature extraction and classification techniques (i.e. BDPCA+LSLDA+RBF). In addition, SVM classifier is pre- sented with BDPCS and LSLDA and BDPCA+LSLDA+SVM) for evaluation of emotion. The extracted audio features are passed into an audio feature fusion section that uses a set of rules to decide the most expected emotion contained in the audio signal. The performances of the proposed audio path and the final system are evaluated on standard databases of audio clips extracted from the video.

Index Terms: Emotion recognition, audio-visual processing, rule-based, machine learning, multimodal system.

I. INTRODUCTION

Emotion recognition is an automated process to identify the affective state of a person and has gained the increasing attention of researchers in the human-computer interaction (HCI) field for various applications like automotive safety, gaming experiences, the mental diagnosis in military service, customer services, etc. Over the decades, several research efforts have been conducted for audio-visual emotion recognition. In the survey, three main approaches can be broadly distinguished: (i) audio based approaches, (ii) visual-based approaches, and (iii) audio-visual approaches. Initial works focused on treating the audio data and visual data modalities separately. The audio-based emotion recognition efforts are based on extracting and recognizing the emotional states contained in the human speech signal. An important issue is the selection of the salient features to be used for discriminating the different emotions. Prosodic and spectral features are two different types of features to be useful for recognizing emotion in speech. Examples of commonly used prosodic features are pitch and energy and examples of commonly used spectral features are Mel-scale frequency cepstral coefficients (MFCC). Although prosodic features are commonly used in many works some researchers have demonstrated the usefulness of spectral features for speech emotion recognition. The monograph work further investigated combining different types of features like prosodic and spectral features for audio-based emotion recognition. The visual-based emotion recognition efforts are based on extracting and recognizing the emotional states contained in the human facial expression. An example is a recent work by Tawari Trivedi which used a representation of image sequences by weighted sums of registered face images where the weights are derived using auditory features. Some other examples are the studies by which recommended facial expression recognition that based on local binary patterns. A recent effort proposed a bimodal emotion recognition system which utilized Kernel Cross-Modal Factor Analysis. Their system achieved 72.47% and 82.22% recognition rates when evaluated using the eNTERFACE05 and RML audio-visual emotion databases respectively. Other recent works on audio-visual emotion recognition include the works. From the above, we observed that the potential to improve the existing performance of the audio-visual emotion recognition is still far from a solved problem due to the restrictions in the accuracy recognition. In this paper, adapted an approach for audio-based rule development based on the works which were motivated by the psychological study of emotions by Schlosberg in which emotions are represented in three dimensions: activation (arousal), potency (power) and evaluation (pleasure). In these works, the authors remarked that the optimal feature set to be used strongly depends on the emotions to be separated and that using one global feature set for the discrimination of all emotions is suboptimal. They proposed a three-stage classification technique to recognize six emotions (anxiety, happiness, anger, neutral, boredom and sadness) using a set of rules and showed that their three-stage rule-based approach could outperform the single-stage approach. In this work, a different set of rules is proposed to recognize a different set of six emotions (anger, happy, sad, disgust, surprise and fear). The next sections of the paper are organized as follows: Section II gives the essential literature survey. Section III addresses existing. Section IV introduces the proposed architecture overview. Section V shows mathematical model for important processing functions. Section VI describes assumptions expected results. Section VII accomplishes the paper.

II. REVIEW OF LITERATURE

In the literature review, topical methods over the emotion recognition are discussed. Some of them are as follows.
K. P. Seng et al. [1] proposes an audio-visual recognition system to recognize human emotion that uses a mixture of rule-based as well as machine learning methods to increase the recognition efficacy in the audio as well as video paths. The audio-visual path is designed using the Bidirectional Principal

Component Analysis (BDPCA) for the reduction of dimensionality and Least-Square Linear Discriminant Analysis (LSLDA) for discrimination of class. G. Chetty et al. [2] analyzes the strong points and the limits of systems built only on facial expressions or acoustic data. Moreover, it analyses two different methods decision level and feature level integration that are used to fuse these two (i.e. facial expressions or acoustic data) modalities. Authors propose a novel multilevel fusion method for improving the person dependent as well as person independent classification performance for various emotions. H.W. Kung et al. [3] projected a Dual Subspace Nonnegative Graph Embedding (DSNGE). This technique represents images having expressions using identity and expression subspaces. The identity subspace describes identity-dependent presence variations however the expression subspace describes identity-independent expression variations. S. Zhalehpour et al. [4] presented a framework for recognition of multimodal emotion based on a dissimilarity matrix method for automatic peak frame collection from audio-visual video sequences. S. Poria et al. [5] suggest a multimodal data removal agent, which infers as well as aggregates the semantic and affective data related with user created multimodal data in contexts such as e-health, e-learning, automatic tagging for video content and human computer interaction. In [6] C. H. wu et al. presented a review on the theoretical as well as practical effort offering different and comprehensive views of the latest study in emotion recognition from bimodal data including facial and vocal expressions. In [7] projected novel architecture of intellectual audio emotion recognition system. This architecture completely utilizes prosodic and spectral features in its design module.H. Gaidhane et al. [8] presented a simple method for face recognition amongst numerous human faces. This method is relying on the covariance matrix, polynomial coefficients, and algorithm on common eigenvalues. The advantage of this approach is to the identify similarity among human faces without computing actual eigenvalues and eigenvectors. A symmetric matrix is evaluated using the polynomial coefficients-based companion matrices of two comparative images. Cheng Q et al. [9] introduced a minimax structure for multiclass classification, which is appropriate to common data including, imagery and other types of high-dimensional data. C. Fadil et al. [10] presented the emoFBVP database of multimodal (face, voice, body sign and physiological signals) recordings of artists enacting different emotion expressions. They also define four Deep Belief Network (DBN) models and demonstrate that these models produce robust multimodal features for emotion classification in an unsupervised manner. A. Tawari et al. [11] presented a facial expression recognition framework by using audio-visual data analysis. They also suggest modeling the cross-modality data correlation, however, allowing them to be preserved as asynchronous streams. Yi-Ren Yeh et al. [12] recommend an innovative rank-one update technique by means of an easy class indicator matrix. Least-Square Linear Discriminant Analysis (LSLDA) model can be drawn-out to address the learning process of concept drift, in which the newly conventional information exhibit with regular or abrupt variations in distribution. S. Dobriek et al. [13] presents multimodal recognition of emotion system abusing audio and video information. The method first processes both sources of information separately to generate corresponding matching scores and after that combines the evaluated matching scores to get a classification decision. For the video part processing of the system, a novel approach is developed which rely on image set matching to recognize emotion. The presented method avoids the necessity for detecting and tracking particular facial landmarks during the given video sequence, which signifies a common source of error in video based emotion recognition systems, therefore, enhances robustness to the video dispensation chain. In [14] Z. Xie et al. purposes at providing the overall theoretical examination for the problem of multimodal data fusion and applying novel data theoretic tools in a multimedia application. The greatest important issues for data fusion include feature revolution and reduction of feature dimensionality. Md. A. Hossan et al. [15] offered new MFCC feature extraction method established on distributed Discrete Speaker authentication tests are projected based on three unlike feature extraction approaches including conventional MFCC, Delta- Delta MFCC and distributed DCT-II based Delta-Delta MFCC with a Gaussian Mixture Model (GMM) classifier, Cosine Transform (DCT-II). M. Lugger and B. Yang [16] presented a methodology for classification of speaker independent emotions. Authors used a huge set of voice quality constraints along with standard prosodic features. In entirely classification readings, they used the SFFS algorithm to decrease the feature number to 25. They moreover observed that generally, a multi-stage classification achieves better than a flat classifier.

III. PROPOSED SYSTEM

A. System Overview

The proposed system consist the audio recognition approach to detect emotion from the voice of any individual. This approach uses training and testing approach for emotion detection from audio which is extracted from the video. The training part contains an Audio Features Analyzer to extract some important prosodic features such as pitch, log-energy, zero-crossing rate (ZCR), Teager energy operator (TEO) and MFCC. Then these audio features are stored into the database with the annotation of emotion in an audio clip. In testing part, for the audio path, the first sub-path A1 is designed based on the audio prosodic features. It contains an Audio Features Analyzer to extract some important prosodic features such as pitch, log-energy, zero-crossing rate (ZCR) and Teager energy operator (TEO). It also contains a module called Audio Feature-Level Fusion to perform fusion at the feature level. The second sub-path A2 is designed based on audio spectral features. This path consists of Melscale frequency cepstral coefficients (MFCC) feature extraction followed by three parallel sub-paths for three sets of emotion groups. An Audio Decision-Level Fusion module is also proposed to fuse the information from both sub-paths A1 and A2. A decision- making mechanism is included in the fusion module to decide the most likely audio emotion. For classification, we proposed an optimal data fusion technique for training two-class RBF classifiers for the audio path and SVM classifier for analysis.

B. Proposed Architecture Diagram

Figure 1 shows the proposed approach for emotion recog-nition.

IV. EMOTION DETECTION PROCESSING

The audio path can be seen in the bottom block diagram in Fig. 1. The Path A2 contains three sub-paths in parallel for three emotion groups. Each emotion group is designed to contain two emotion classes. The three emotion groups are Group 1, Group 2 and Group 3.

A. ***Preprocessing:*** The Voice Activity Detector (VAD) is used for pre- processing the speech signal to eliminate the

background noise and segment out the non-speech portions of the audio signal. The VAD technique uses the short-time zero-crossing rate (STZCR) features. Using this unwanted signal frames (e.g. silence or unvoiced) are then segmented out.

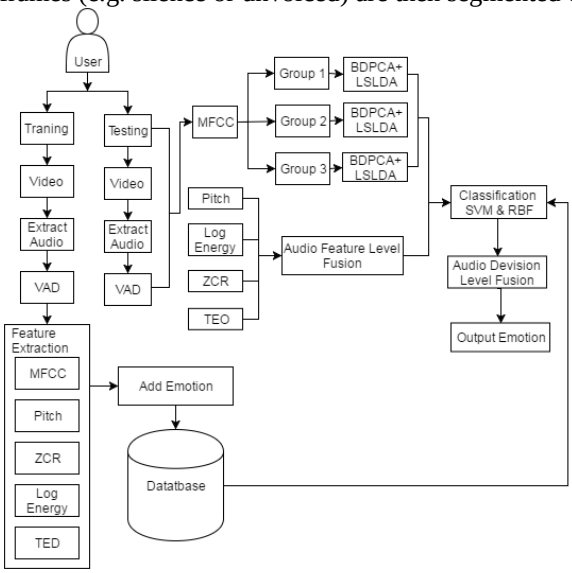


Figure.1. Architecture Diagram of Proposed System

B. Audio Analyzer Feature Level Fusion (Path A1)

The Audio Feature Analyzer extracts important features such as pitch, log-energy, ZCR, and TEO from the input signal. The pitch extraction method calculates the distance between the zero crossing points of the signal. The log-energy indicates the total squared amplitude in a segment of speech. The zero crossing rates (ZCR) computes the weighted average of the number of times the speech signal changes sign within a particular time window. The TEO detects the nonlinear component which changes appreciably between different emotional speech signals. The MFCC spectral features from the Path A2 will be used to recognize the individual emotion within the group. The MFCC spectral features from the Path A2 will be used to recognize the individual emotion within the group. For example, the disgust and surprise emotions should be put into different groups since they can be discriminated using the TEO and log-energy features.

C. Emotion Groups Classification and Extraction (Path A2)

The Path A2 flow begins with processing the MFCC from the speech The MFCC feature has been identified to be one of the most influential audio features. After the MFCC feature extraction, these features are passed to three parallel sub- paths which use similar feature extraction and classification techniques as for the visual path (i.e. BDPCA+LSLDA+RBF and BDPCA+LSLDA+SVM). SVM and RBF classifier is used to classify emotion.

D. Audio Decision-Level Fusion

The Audio Decision-Level Fusion module makes the final decision based on the outputs from the Path A1 and the Path A2. Output emotion is recognizing from training data stored in the database.

V. MATHEMATICAL MODEL

a. Set Theory

$S = \{s1, s2, s3, , sn\}$- set of audio signal.

$E = log10^{PN}$

$x2$

b. Compute energy as:

$I = \{i1, i2, i3, , in\}$ - set of frames extracted from audio.

$E = \{e1, e2, e3, en\}$- set of emotions.

$Sf = \{sf1, sf2, sf3, sfn\}$- set of audio signal features.

c. Functional Dependency

Where N denotes the number of frames and x denotes the sample of the speech.

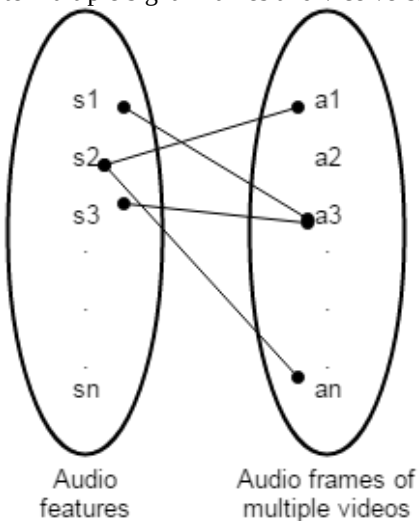
d. Compute zero crossing rate (ZCR) as:

$1if\ x(n) \geq 0$

Audio features are computed for each frame (signal frame) extracted from that audio. Thus features of audio are function-

$sgn\ \{x\} = -1 \quad if \quad x(n) < 0$

ally depend on audio signal. Same audio features are belongs to multiple signal frames and vice versa.



Where n refers to the current sample.

e. An energy tracking operator for the speech signal is computed as:

$\Psi\ [s(n)] = s^2(n) - s(n+1)s(n-1)$

Where, $\Psi\ [s(n)]$ are the coefficients and $s(n)$ is the sampled speech signal.

4. Audio Decision-Level Fusion.
5. Classify audio features.
6. Recognize the output of audio emotion.

VI. ANALYSIS AND RESULTS

C. Process

1. Select audio and process it by framing, audio signal S

is the set of frames. Thus it can be represented as in (1).

$$S = {}^{n}_{i=0}\mathrm{P}\, Ii \quad (1)$$

2. Extract noise and segment out the non-speech portions of the audio signal using Voice Activity Detector (VAD).

a. Break audio signal into frames.

b. Extract features from each frame.

c. Train a classifier on a known set of speech and silence frames.

d. Classify unseen frames as speech or silence.

3. Audio Analyzer Feature Level Fusion (Path -A1).

. Let the speech signal $x(m)$ in the time domain is first divided into n number of frames by windowing $w(n)$ and is denoted as $s(m)$. Then the observed periodicity R is:

$$R(k) = {}^{L-k-1}_{m=0}\mathrm{P}\ s(m)s(m+k)$$

Where *L* denotes the window length, and *k* refers to the representation of the pitch period of a peak.

A. *Dataset*

For evaluation, video dataset is used which consist multiple videos having different emotions like disgust, sad, surprise, fear, angry and happy.

B. Expected results

To evaluate the performance, RBF and SVM classifiers are used. The expected results are evaluated according to classification outcome and time complexity. Time requires to process audio clips with different emotion and size are computed for proposed and existing system as shown in figure

2. To compute elapsed time for the process, it uses the formula: Total time = end time - start time.

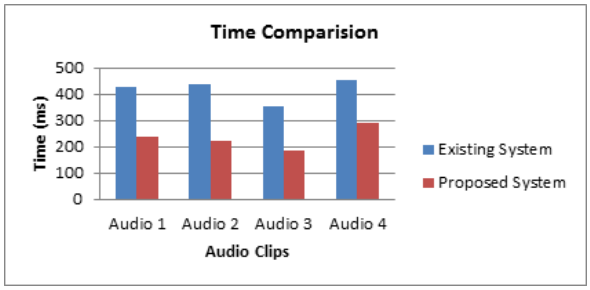


Figure.2. Time comparision graph for RBF and SVM

Time evaluation is analyzed by comparing time require for each algorithm. The expected results shows that time requires for RBF algorithm processing and SVM algorithm processing for testing audio. Figure 3 represents expected time require for audio processing in the existing and proposed system. This time is calculated in milliseconds.

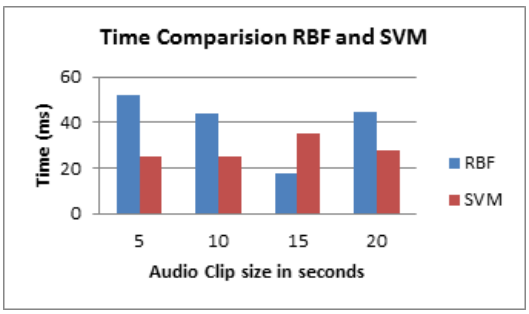


Figure. 3. Time comparision graph existing and proposed

VII. CONCLUSION

This paper has proposed a combined rule-based and ma- chine learning approach to solving the audio emotion recog- nition problem extracted from the video stream. The main challenges were to determine the suitable feature extraction techniques and optimal data fusion techniques previous to clas- sification. To address these challenges, the BDPCA+LSLDA proposed for feature extraction. For classification, an optimal data fusion technique using training an RBF neural and SVM classifiers are used and results are analyzed using time, accuracy parameters. BDPCA and LSLDA are combining with these classifiers to compute emotions involved in extracted audio signal.

VIII. ACKNOWLEDGMENT: The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance.

IX. REFERENCES

[1]. K. P. Seng, L. Ang, and C. S. Ooi, A Combined Rule-Based Machine Learning Audio-Visual Emotion Recognition Approach, IEEE Tran. On affective computing, TAFFC-2015-01-0016.R2.

[2]. G. Chetty, M. Wagner, and R. Goecke, ”A multilevel fusion approach for audiovisual emotion recognition”, Emotion Recognition: A Pattern Analysis Approach (Ed. A. Konar and A. Chakraborty), pp. 437-460, 2015.

[3]. H.-W. Kung, Y.-H. Tu, and C.-T. Hsu, ”Dual subspace nonnegative graph embedding for identity-independent expression recognition,” IEEE Trans. Inform. Forensics Sec., vol. 10, no. 3, pp. 626-639, 2015.

[4]. S. Zhalehpour, Z. Akhtar, and C.E. Erdem, ”Multimodal emotion recogni- tion based on peak frame selection from video” Signal, Image and Video Processing, pp. 1-8, 2015.

[5]. S. Poria, E. Cambria, A. Hussain, and G.-B. Huang, ”Towards an intelligent framework for multimodal affective data analysis,” Neural Networks, vol. 63, pp. 104-116, 2015.

[6]. C-H. Wu, J-C. Lin, and W-L. Wei, ”Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies,” APSIPA Trans. Signal and Information Processing, vol. 3, pp. 1-18, 2014.

[7]. C.S. Ooi, K.P. Seng, L.M. Ang, and L.W. Chew, ”A new approach of audio emotion recognition,” Expert Systems with Applications, vol. 41, pp. 5858-5869, 2014.

[8]. V.H. Gaidhane, Y.V. Hote, and V. Singh, ”An efficient approach for face recognition based on common eigenvalues,” Pattern Recognition, vol. 47, pp. 1869-1879, 2014.

[9]. Q. Cheng, H. Zhou, J. Cheng, and H. Li, ”A minimax framework for classification with applications to images and high dimensional data,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 11, pp. 2117-2130, 2014.

[10]. C. Fadil, R. Alvarez, C. Martinez, J. Goddard, and H. Rufiner, ”Multi- modal emotion recognition using deep networks”, in Proc. Latin Ameri- can Congress on Biomedical Engineering, pp. 813- 816, 2014.

[11]. A. Tawari and M.M. Trivedi, ”Face expression recognition by cross modal data association,” IEEE. Trans. Multimedia, vol. 15, no. 7, pp. 1543-1552, 2013.

[12]. Y.-R. Yeh and Y.-C.F. Wang, ”A rank-one update method for least squares linear discriminant analysis with concept drift,” Pattern Recogni- tion, vol. 46, pp. 1267-1276, 2013.

[13]. S. Dobrisek, R. Gajsek, F. Mihelic, N. Pavesic, and V. Struc,”To- wards efficient multi-modal emotion recognition,” Int. Journal Advanced Robotic Systems, vol. 10, no. 53, pp. 1-10, 2013.

[14]. Z. Xie and L. Guan, ”Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools”, in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), pp. 1-6, 2013

[15]. M.A. Hossan, S. Memon, and M.A. Gregory, ”A novel approach for MFCC feature extraction,” in Proc. 4th International Conference on Signal Processing and Communication Systems, pp. 1-5, 2010

[16]. M. Lugger and B. Yang, ”Psychological motivated multi-stage emotion classification exploiting voice quality features,” Speech Recognition: Technologies and Applications (Ed. F. Mihelic and J. Zibert), pp. 395- 410, 2008.