Emotion Analysis and Speech Signal Processing

Omar Raghib[1], Eshita Sharma[2], Tameem Ahmad[3], Faisal Alam[4]
Department of Computer Engineering
Z. H. College of Engineering & Technology,
Aligarh Muslim University, Aligarh, India
[1]raghibomar786@gmail.com, [2]eshita96sharma@gmail.com, [3]tameemahmad@gmail.com, [4]alamfaisal654@gmail.com

***Abstract*—In recent times there have been notable advancements in the field of Automatic Speech Recognition (ASR) in terms of the technology used as well as its applications. But still the performance of ASR systems is much inferior as compared to Human Speech Recognition (HSR). One reason being that the ASR systems are not much developed when it comes to the recognition of emotions carried in the speech signal which is done involuntarily in HSR. This paper discusses the series of steps to be followed in order to analyze the speech signal for the recognition of its emotions highlighting some of the best available techniques at present for each step. The disturbances such as background noises incurred during the transmission makes speech processing complex and difficult. In this paper, we present a mechanism to perform the classification of the speech for different emotions.**

Keywords—speech processing, emotion recognition, feature extraction, feature selection, classification techniques

I. INTRODUCTION

While communicating, the emotions associated with the speech are as significant as the words in it to convey the right message. Emotion recognition by the computer based solely on voice is a challenging task in the field of human-computer interaction (HCI). By being adaptive to the emotions of the users, the computers could provide a better and more personalized user experience. There are numerous applications of recognizing the speaker's emotions automatically from his speech signal, for example, it can be used to analyse the degree of satisfaction of the client in automated customer care services. This analysis also helps in adding an emotional factor to artificially synthesized speech. Different studies regarding the extraction of features that are best suited for recognizing emotions are discussed in [1-2]. Various features like prosodic features, vocal tract features and source features have been discussed in [3-4]. The techniques commonly used for emotion recognition from speech include k - Nearest Neighbor (k-NN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Gaussian Mixture Model (GMM) which have been discussed in [5-10].

Emotion recognition is done based on the various features present in the speech signal. A feature may be defined as *"Any distinctive aspect, quality or characteristic which, may be symbolic (i.e., color) or numeric (i.e., height)"* [11]. The n-dimensional vector representing n distinct features of the signal is known as a feature vector and the n-dimensional space created using the feature vector is called the feature space [12].

The approaches followed in an ASR system are acoustic phonetic approach, pattern recognition approach and artificial intelligence approach. Acoustic Phonetic approach assigns appropriate labels to the speech sounds provided, based on the phonemes of the input sound [13-14]. The Pattern Recognition involves two processes: pattern training and pattern testing, in which the classification is based on well formulated mathematical frameworks. Examples of this technique include SVM, HMM, ANN. The artificial intelligence approach is a hybrid of the pattern recognition and acoustic phonetic approach [15] in which the classification is done in a similar fashion to that of human beings applying intelligence (artificial in this case) for analysing, decision making and recognizing the features. Some common examples of this approach include MLP, BPNN, SOM, etc. Though all the three approaches are widely used in various fields of automatic speech recognition but in case of emotion recognition, the second approach i.e. the pattern recognition approach is preferred. This paper discusses the various steps followed in emotion analysis from speech signal using pattern recognition approach. Starting with the database used as training data for the pattern recognition approach, the paper discusses the techniques to be followed beginning with the pre-processing to remove the disturbances followed by feature extraction and then feature selection to get the relevant features for the particular purpose and finally classifying the signal based on the selected features.

II. DATASET

The database providing speech signals for emotion recognition is generally of two types – real and acted. In the real type of databases, speech signals are the recordings of real life conversations such as talk shows whereas in case of acted database, speech signals are the recordings of some professional actors speaking in a certain emotion. There are, however, certain differences in the quality of features present in the two types as in case of acted speech, these features are more prominent and easily distinguishable as compared to the real speech. For example, in [16] the acted speech database contained corpus which included the voices of two male and two female actors classifying four emotions – happy, sad, anger and neutral each being recorded by every individual actor whereas the real emotion database contained voice samples of a single male actor collected from Telugu movies.

978-1-5386-0814-2/17/$31.00 ©2017 IEEE

Also, in [17] the classification was done only on the basis of acted emotions in the speech database.

III. METHODS

The basic steps for speech recognition in the automatic speech recognition systems are: Pre-processing, Feature extraction, Feature selection and Classification. Each of these steps is, in turn, composed of one or more steps which may be required in some cases while may be optional for some others. *Fig. 1* shows these steps along with some of the best available methods for achieving these steps. The pseudo-code of workflow for emotion recognition from speech signal is shown in *Fig. 2*.

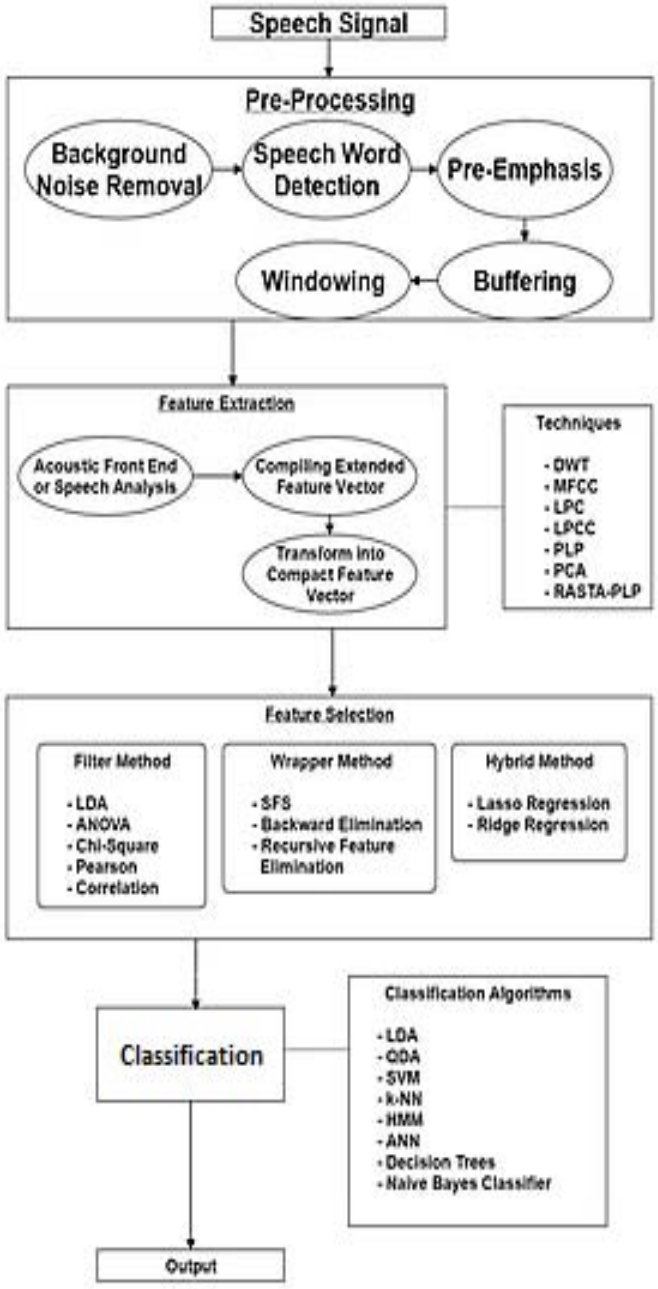

Fig. 1. Basic workflow of emotion analysis.

```
Procedure userInput
        take/get the input signal from user
        call preProcessing on this input signal
end
Procedure preProcessing(input signal)
        remove background noise
        perform speech word detection
        execute pre-emphasis
        buffer the resulting signal
        finally do the windowing of the signal
        call featureExtraction
end

Procedure featureExtraction(pre-processed signal)
        perform the spectro temporal analysis of the
pre-processed signal (Speech analysis)
        compile the extended feature vector obtained in
the previous step
        transform the feature vetor into compact form
        call featureSelection
end

Procedure featureSelection(extracted feature vector)
        select relevant features (for the particular
purpose) from the extracted feature vector
        call classification
end

Procedure classification(selected feature vector)
        identify the set to which the selected feature
vector belongs
end
```

Fig. 2. Pseudo-code.

A. Pre-Processing

The speech signal that is to be processed is in analog form and so pre-processing of speech signal is required to transform the continuous-time signal to discrete form which is suitable for further digital processing. In order to convert the signal in a form that is convenient for further processing, the pre-processing is done which aims at deriving a parameter set which is free from noise and variations in amplitude that may be introduced in the transmission medium. It also tries to remove potentially malicious signals [18]. As pointed out in [19], signal pre-processing is achieved in five steps namely: Background Noise Removal, Speech Word detection (End Point Detection), Pre-emphasis, Buffering (Frame Blocking) and Windowing. Hence, in a speech recognition system, pre-processing is primarily used for increasing the efficiency of the following feature extraction and classification thereby

improving the performance of speech recognition. Pre-processing results in compressed and filtered frames of speech that acts as an input to the subsequent feature extraction stage.

B. Feature Extraction

After the pre-processing step, feature extraction is the second component of automatic speech recognition (ASR) systems. The windowed and enhanced speech signal obtained after pre-processing still has a high dimensionality which is an undesirable trait for subsequent selection and classification stages as it would result in a higher word error rate. For better results, we need to extract only those features (from the pool of features in the signal) which we believe would be suitable for differentiating sounds during classification. The aim of feature extraction algorithms is, therefore, to derive a characteristic feature vector having lower dimensionality (than the input signal) containing only the extracted features that should result in a better classification of sound. As discussed in [20], feature extraction is performed in three steps. The first step, known as the acoustic front end or speech analysis, extracts raw features that determine the envelope of the power spectrum of the signal and is achieved by performing the spectro temporal analysis of the input signal. An extended feature vector comprising of dynamic and static features is compiled at the second stage of feature extraction. The extended feature vector obtained in the second step is transformed into more compact and robust feature vector in the third and final step.

Some of the commonly extracted features in speech related discrimination in time and frequency domain include:

1) Spectrul Flux: It represents the rate of change of power spectrum of the signal and is calculated by taking the summation of the difference between adjacent frames.

2) Spectral Roll-Off: It specifies the percentage of frequencies that are concentrated below a given energy threshold which is usually set to 95% according to [28].

3) Low Energy Ratio: Represents the number of the frames in which the effective or root mean square (RMS) energy of the particular frame is less than the average energy of the input signal.

4) Spectral Centroid: It specifies the center of mass in the frequency spectrum.

5) Number of Zero Crossings: It represents the number of zero crossings of a particular frame in time domain and thus measures dominant frequency in the signal.

The formulas for calculation of these features can be found in [20]. Mean and variances of these features are usually computed for better results. The commonly used feature extraction techniques in speech recognition systems are as follows [21-22].

- Discrete Wavelet Transform (DWT)
- Mel-Frequency Cepstrum Coefficients (MFCC)
- Linear Predictive Coding (LPC)
- Linear Prediction Cepstral Coefficients (LPCC)
- Perceptual Linear Prediction (PLP)
- Principal Component analysis (PCA)
- Relative Spectral (RASTA-PLP)

According to [20], in speech classification and recognition applications, Mel-frequency cepstrum coefficients have proved to be extremely successful.

C. Feature Selection

The accuracy of the classifiers can be greatly influenced by the number of features present and so the presence of large number of irrelevant features may result in less accurate classification thereby making effective classification practically impossible. Feature Selection is used to select the subset of most relevant features (for the particular purpose) from the previously extracted feature set thus increasing the efficiency of the particular classification. As mentioned in [23], feature selection aims at achieving the following objectives: to convert the feature set into a form having reduced size than the original one which would enhance the performance for predicting the required characteristics, to make the prediction process computationally fast and efficient, and to provide a better understanding of the underlying process that generated the required data. Feature selection can be implemented by three methods:

- Filter method
- Wrapper method
- Hybrid/embedded method

Filter type methods select variables regardless of the model. Some common filter methods include: Linear discriminant analysis (LDA), Pearson's Correlation, Analysis of variance (ANOVA) and Chi-Square. Unlike filter approaches, subsets of variables are determined in wrapper methods that detect the possible interactions between variables. Sequential Forward Search (SFS), Backward Elimination and Recursive Feature elimination are most popular wrapper methods. Embedded/Hybrid methods are those that combine the qualities of filter and wrapper methods. Examples of embedded methods include Lasso regression and Ridge regression.

D. Classification

Classification basically assigns a class, from among the classes created using the database as training data, to the input signal based on its selected feature vector. An algorithm used for classification is called classifier. The most commonly used classifiers for emotion recognition from speech signal include Support Vector Machine (SVM), Hidden Markov Model (HMM) and Artificial Neural Networks. The merits and demerits of which have been discussed in [24]. Other classifiers like k- Nearest Neighbor may also be used but they are usually less efficient in case of emotion recognition from speech signal as in [17] where SVM and k-NN were applied to emotion recognition system for better improved results. Paper [17], using their dataset in Polish language, SVM produced

better results where the mean accuracy of SVM was found to be 79.2 whereas the mean accuracy of k-NN was only 75.6.

IV. CONCLUSION

This work presents steps and study of techniques used for emotion recognition from speech signals. Although huge amount of researches are being carried out in this area but there are still many possibilities for improvements. A general workflow of techniques to be used for emotion recognition of speech signal have been discussed in this paper but all of these steps might not always be necessary (though their inclusion has no adverse effects on the results) like in case of [17], no pre-processing was done as only those signals were considered which were free from noise. Also different classification techniques discussed in the paper produced results with different accuracies in different cases like in [17], SVM produces results with accuracies ranging from 64.1% to 90.7%. Also, going unconventionally, feature selection combined with representation as discussed in [25] can be used for emotion recognition to improve the accuracy of the system. The emotions are also language specific and it is important to consider the appropriate measures for each language.

References

[1] T.L. Pao, C. H. Wang, and Y. J. Li, "A Study on the Search of the Most Discriminative Speech Features in the Speaker Dependent Speech Emotion Recognition," in 2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming, 2012, pp. 157–162.

[2] H. Atassi, A. Esposito, and Z. Smekal, "Analysis of high-level features for vocal emotion recognition," in 2011 34th International Conference on Telecommunications and Signal Processing (TSP), 2011, pp. 361–366.

[3] D. Kamińska, T. Sapiński, and A. Pelikant, "Comparison of perceptual features efficiency for automatic identification of emotional states from speech," in 2013 6th International Conference on Human System Interactions (HSI), 2013, pp. 210–213.

[4] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," Int. J. Speech Technol., vol. 15, no. 2, pp. 99–117, Jan. 2012

[5] Garg, Anjali, and Poonam Sharma. "Survey on acoustic modeling and feature extraction for speech recognition." Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on. IEEE, 2016.

[6] Khan, Ayesha and Uzzaman Khan, Yusuf. "Time Domain based Seizure Onset Analysis of Brain Signatures in Pediatric EEG", 4th International Conference on "Computing for Sustainable GlobalDevelopment" (INDIACom-2017), 2017.

[7] K. V. K. Kishore and P. K. Satish, "Emotion recognition in speech using MFCC and wavelet features," in Advance Computing Conference (IACC), 2013 IEEE 3rd International, 2013, pp. 842–847.

[8] S. A. Rieger, R. Muraleedharan, and R. P. Ramachandran, "Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers," in 2014 9th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2014, pp. 589–593.

[9] E. Bozkurt, E. Erzin, C. E. Erdem, and A. T. Erdem, "Use of Line Spectral Frequencies for Emotion Recognition from Speech," in 2010 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 3708–3711.

[10] T. L. Pao, Y. T. Chen, J. H. Yeh, Y. M. Cheng, and C. S. Chien, Feature Combination for Better Differentiating Anger from Neutral in Mandarin Emotional Speech. LNCS 4738, ACII 2007: Springer-Verlag Berlin Heidelberg, 2007.

[11] "Feature." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 4 Apr. 2017. Web. 2 May. 2017.

[12] Gutierrez-Osuna R., Intelligent Sensor Systems, Course Notes, Department of Computer Science, Wright State University, (http://research.cs.tamu.edu/prism/lectures/iss/iss_l9.pdf).

[13] Nidhi Desai, Prof. Kinnal Dhameliya, "Feature Extraction and Classification Techniques for Speech Recognition: A Review," International Journal of Emerging Technology and Advanced Engineering (IJETAE), Vol. 3, Issue 12, December 2013.

[14] Ranu Dixit, NavdeepKaur, "Speech Recognition Using Stochastic Approach: A Review", International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), Vol.2, Issue 2, February 2013.

[15] Sanjivani S. Bhabad, Gajanan K. Kharate, "Overview of Technical Progress in Speech Recognition", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol.3, Issue 3, March 2013.

[16] Rao, K. Sreenivasa, "Emotion recognition from speech." International Journal of Computer Science and Information Technologies 3.2 (2012): 3603-3607.

[17] Majkowski, Andrzej, "Classification of emotions from speech signal." Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2016. IEEE, 2016.

[18] Bansal, Palak, and Tameem Ahmad, "Methods and Techniques of Intrusion Detection: A Review", International Conference on Smart Trends for Information Technology and Computer Communications. Springer, Singapore, 2016.

[19] Singh, Bhupinder, Vanita Rani, and Namisha Mahajan. "Preprocessing In ASR for Computer Machine Interaction with Humans: A Review." International Journal 2.3 (2012).

[20] Shanthi Therese, S., and Chelpa Lingam. "Review of feature extraction techniques in automatic speech recognition." International Journal of Scientific Engineering and Technology 2.6 (2013): 479-484.

[21] S. Karpagavali, P.V. Sabitha, "Isolated Tamil Words Speech Recognition using Linear Predictive Coding and Networks," International Journal of Computer Science and Management Research, Vol.1, Issue 5, December 2012.

[22] UtpalBhattacharjee, "A Comparative Study of LPCC and MFCC Features for the Recognition of Assamese Phonemes," International Journal of Engineering Research and Technology (IJERT), Vol.2, Issue 1, January 2013.

[23] Guyon I. and Elisseeff A., "An introduction to variable and feature selection", Journal of Machine Learning Research, 3, 1157-1182, 2003.

[24] Madan, Akansha, and Divya Gupta. "Speech Feature Extraction and Classification: A Comparative Review." International Journal of computer applications 90.9 (2014).

[25] Han, Wenjing, "Combining feature selection and representation for speech emotion recognition." Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on. IEEE, 2016.