A BRIEF STUDY ON SPEECH EMOTION RECOGNITION

Akalpita Das
Department of Instrumentation & USIC, Gauhati University
Guwahati, India, dasakalpita@gmail.com;

Laba Kr. Thakuria
Department of Instrumentation & USIC, Gauhati University
Guwahati, India, thakurialaba@gmail.com

Purnendu Acharjee
Department of Instrumentation & USIC, Gauhati University
Guwahati, India
pbacharyaa@gmail.com

Prof. P.H. Talukdar
Department of Instrumentation & USIC, Gauhati University
Guwahati, India
phtassam@gmail.com

Abstract*—Speech Emotion Recognition is a current research topic because of its wide range of applications and it became a challenge in the field of speech processing too. In this paper, we have carried out a brief study on Speech Emotion Analysis along with Emotion Recognition. This paper includes the study of different types of emotions, features to identify those emotions and various classifiers to classify them properly. The first part of the paper is enriched with an introductory description. Second part covers the different features along with some popular extraction method. Third part includes various classifiers used in SER and finally the conclusion part puts an end to this paper.

Key Terms: - Ser; mfcc; lpcc; svm; gmm; hmm; knn; adaboost algorithm.

I.INTRODUCTION

Speech Emotion Analysis refers to the use of various methods to analyze vocal behaviour as a marker of affect (e.g., emotions, moods, and stress), focusing on the nonverbal aspects of speech. The basic assumption is that there is a set of objectively measurable voice parameters that reflect the affective state a person is currently experiencing (or expressing for strategic purposes in social interaction). This assumption appears reasonable given that most affective states involve physiological reactions (e.g., changes in the autonomic and somatic nervous systems), which in turn modify different aspects of the voice production process. For example, the sympathetic arousal associated with an anger state often produce changes in respiration and an increase in muscle tension, which influence the vibration of the vocal folds and vocal tract shape, affecting the acoustic characteristics of the speech, which in turn can be used by the listener to infer the respective state [19].

Usually human beings can easily recognize various kinds of emotions. This can be achieved by the human mind through years of practice and observation. The human mind captures all kinds of emotions since childhood and is taught to differentiate between the emotions based on its observations. For instance, when a person is angry, his tone raises, his expression becomes stern and the content of his speech no longer remains pleasant [3]. Similarly, when a person is happy, he speaks in a musical tone, there is a look of glee on his face and the content of his speech is rather pleasant and joyous. Based on these observations, a person can quickly identify the state of the speaker – whether he is happy, sad, angry, depressed, disgusted etc. Speech Emotion Recognition deals with this part of research in which machine is able to recognize emotions from speech like human. Emotions are expressed in the voice can be analyzed at three different levels:

A) The physiological level (e.g., describing nerve impulses or muscle innervations patterns of the major structures involved in the voice-production process).
B) The phonatory-articulatory level (e.g., describing the position or movement of the major structures such as the vocal folds).
C) The acoustic level (e.g., describing characteristics of the speech wave form emanating from the mouth).

The general architecture for Speech Emotion Recognition (SER) system has three steps shown in Figure 1.

a) A speech processing system extracts some appropriate quantities from signal, such as pitch or energy etc.
b) These quantities are summarized into reduced set of features with the help of feature extractor.
c) A classifier learns in a supervised manner with example data how to associate the features to the emotions.

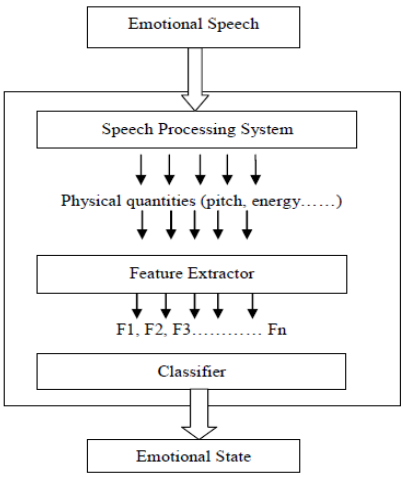

Fig1:Architecture for Speech Emotion Recognition (SER)

Both spectral and prosodic features can be used for speech emotion recognition because both of these features
contain the emotional information. The potential features are extracted from each utterance for the computational mapping between emotions and speech patterns. The selected features are then used for training and testing by using any classifier method to recognize the emotions.

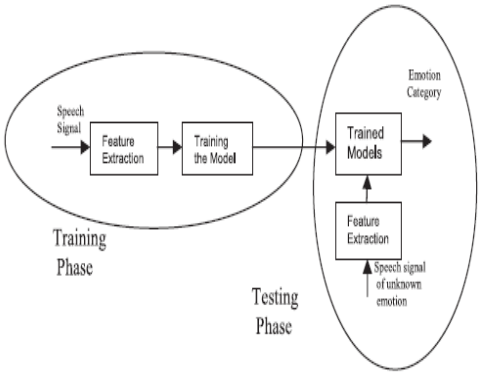

Figure 2: Training and validation of Emotion Recognition Models

Despite all the research efforts, the performances of SER systems designed remain relatively low compared to related fields such as speaker verification. The low recognition rate is reflected through the high level of confusion between emotion classes. The confusion between classes can have several sources. The first raison can be related to the uncertainty that characterize the definition of emotion in the psychology domain. On the other hand, the overlap in the acoustic space is not limited to the neighbouring classes but also between some emotions in a symmetrical position with respect to the active / passive axis of the dimensional model shown in Figure 3. The joy and anger classes are an example of ambiguity case confirmed by several studies [20][21].

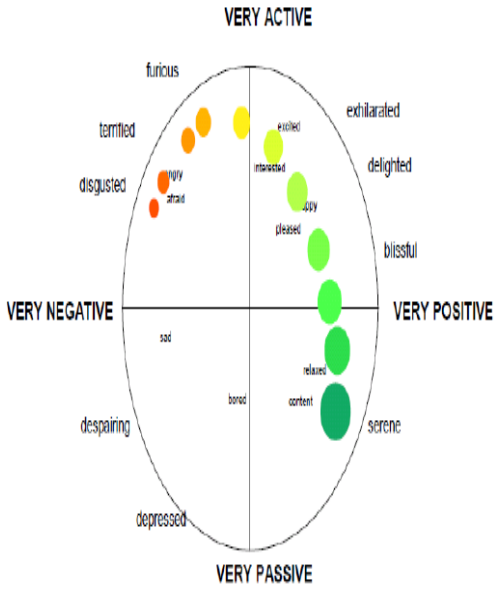

Figure 3: Dimensional model for different emotions

Another important factor which increases the ambiguity is the significant variability between different Individuals in the expression of same emotion class. This wide range of variability can have several origins such as the culture, the age, the gender of the speaker and its spoken language. Finally, the noise contained in the emotion corpus increases the confusion. This noise can be induced by annotators during the labelling operation. This is particularly true for blended emotions such as fear and anger.

II.TYPES OF EMOTIONAL SPEECH

As reported in [19], there are three methods to constitute an emotional corpus: **Natural emotions** are recordings of spontaneous emotional states naturally occurred. It is characterized by a high ecological validity but suffers from the limited number of available speakers and presents difficulties for the annotation. **Simulated emotions** are emotional states portrayed by professional or lay actors according to emotion labels or typical scenarios. Although this method permits easily to constitute an emotion corpus however it has been criticized that this kind of emotion are more exaggerated than natural or induced emotion. **Induced emotions** require speaker's own thinking of the past incident and induce the same emotion in him by remembering the entire situation of the past incident.

III.DATABASES

Generally, there are two types of databases that are used in emotion recognition – acted and real. As the name suggests, in acted emotional speech corpus, a professional actor is asked to speak in a certain emotion. In real databases, speech databases for each emotion are obtained by recording conversations in real-life situations such as call centres and talk shows [6]. But it has been observed that there is a difference in the features of acted and real emotional speeches. This is because acted emotions are not felt while speaking and thus come out more strongly [7].

IV.FEATURES OF SPEECH

Speech signals are produced as a result of excitation in the vocal tract by the source signal. Speech features can here fore be found both in vocal tract as well as the excitation source signal. Features that are extracted from the vocal tract system are called system features or spectral features [8]. The most popular spectral features are Mel frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs) and Perceptual linear prediction coefficients (PLPCs). The features extracted from the excitation source signal are called source features. Linear prediction (LP) and glottal volume velocity (GVV) are some source features. Prosodic features are those features which are extracted from long segments of speech such as sentences, words and syllables. They are also known as supra-segmental features [9]. They contain speech properties

such as rhythm, intonation, stress, volume and duration. The acoustic properties of the prosodic features are pitch, energy, duration and their derivatives. The pitch signal is produced when vocal folds vibrate [10]. Pitch frequency and glottal air velocity are the features related to pitch signal. Speech energy is useful because it is related to arousal levels of the emotion. The prosodic features are used to extract emotional expression or excited behaviour of articulators. Glottal activity characteristics are evaluated using source features [11]. Spectral features are used to capture the information regarding the movement of articulators and the shape and size of vocal tract which produces different sounds. Articulator is the part of vocal organs that helps form speech sounds. Active articulators are organs such as pharynx, soft palate, lips and tongue. Upper teeth, alveolar ridge and hard palate are passive articulators [12].

Feature extraction is based on partitioning speech into small intervals known as frames. To select suitable features which are carrying information about emotions from speech signal is an important step in SER system. There are two types of features: prosodic features including energy, pitch and spectral features including MFCC, MEDC, LPCC.

A) ***Energy and related features***

Energy is the basic and most important feature in speech signal. To obtain the statistics of energy feature, we use short-term function to extract the value of energy in each speech frame. Then we can obtain the statistics of energy in the whole speech sample by calculating the energy, such as mean value, max value, variance, variation range, contour of energy [26].

B) ***Pitch and related features***

The vibration rate of vocal is called the fundamental frequency F0 or pitch frequency. The pitch signal has information about emotion, because it depends on the tension of the vocal folds and the sub glottal air pressure, so the mean value of pitch, variance, variation range and the contour is different in seven basic emotional statuses [25]. The following statistics are calculated from the pitch and used in pitch feature vector [27]:

1. Mean, Median, Variance, Maximum, Minimum (for the pitch feature vector and its derivative)
2. Average energies of voiced and unvoiced speech
3. Speaking rate (inverse of the average length of the voiced part of utterance).

C) ***MFCC and MEDC features***

Mel-Frequency Cepstrum coefficients is the most important feature of speech with simple calculation, good ability of distinction, anti-noise. MFCC in the low frequency region has a good frequency resolution, and the robustness to noise is also very good. MEDC extraction process is similar with MFCC. The only one difference in extraction process is that the MEDC is taking logarithmic mean of energies after Mel Filter bank and Frequency wrapping, while the MFCC is taking logarithmic after Mel Filter bank and Frequency wrapping. After that, we also compute 1st and 2^{nd} difference about this feature [25]. Mel frequency cepstrum coefficients (MFCCs) are coefficients of Mel frequency cepstrum (MFC) which is in turn derived from power cepstrum. Cepstrum is derived from the word 'spectrum' by swapping the first half of the word with the second half [13]. A cepstrum is obtained by computing the Fourier Transform of the logarithm of the spectrum of a signal. There are different kinds of cepstrum such as complex cepstrum, real cepstrum, phase cepstrum and power cepstrum. The power cepstrum is used in speech synthesis applications. The cepstrum are linearly spaced frequency bands whereas MFC are equally spaced [14]. Hence, MFCs can provide a better approximation of the speech.

D) ***Linear Prediction Cepstrum Coefficients***

LPCC embodies the characteristics of particular channel of speech, and the same person with different emotional speech will have different channel characteristics, so we can extract these feature coefficients to identify the emotions contained in speech. The computational method of LPCC is usually a recurrence of computing the linear prediction coefficients (LPC) [25].

But before the extraction of features is done, it is necessary that background noises or any other noises are removed. This is because noise or disturbances in the speech interfere with the characteristics of actual speech and the features get altered.

V.THE CLASSIFIER

In real world, human beings can easily detect various kinds of emotions. This can be achieved by the human mind through years of practice and observation. A human machine interface that can process speech having emotional content makes use of a similar concept: training and then testing [4]. In the training phase, the interface is fed with samples of each emotion. The classifier used in the interface extracts features from all the samples and forms a mixture for each emotion. In the testing phase, emotional speech is given as input to the classifier. The classifier extracts the features from the input and compares it to all the mixtures. The input is classified into that emotion to which it is closest. In other words, the input file will be classified into that emotion whose features are the most similar to that of the input file. There are a number of features and classifiers that can be used for the purpose of emotion detection. However, it is difficult to identify the best model among these since the selection of the feature set and the classifier depends on the problem [5].

The various classifiers that are currently being used are Artificial Neural Networks (ANNs), Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), k-nearest neighbours (KNN), Support Vector Machines (SVMs) [15] and AdaBoost Algorithm. The classification techniques can be divided into two categories – those that make use of the timing information and those which do not [16]. Techniques based on HMMs and ANNs retain the timing information whereas classification techniques based on SVMs and Bayes classifier lose the timing information. One positive aspect of the techniques that retain timing information is that they can be used for speech recognition applications in addition to emotion recognition [17]. Different classifiers are discussed below:

A) Support Vector Machine (SVM):

SVM, a binary classifier is a simple and efficient computation of machine learning algorithms, and is widely used for pattern recognition and classification problems, and under the conditions of limited training data, it can have a very good classification performance compared to other classifiers [28]. The idea behind the SVM is to transform the original input set to a high dimensional feature space by using kernel function. Therefore non-linear problems can be solved by doing this transformation.

B) Hidden Markov Model (HMM):

The HMM consists of the first order markov chain whose states are hidden from the observer therefore the internal behaviour of the model remains hidden. The hidden states of the model capture the temporal structure of the data. Hidden Markov Models are statistical models that describe the sequences of events. HMM is having the advantage that the temporal dynamics of the speech features can be trapped due to the presence of the state transition matrix. During clustering, a speech signal is taken and the probability for each speech signal provided to the model is calculated. An output of the classifier is based on the maximum probability that the model has been generated this signal [29]. For the emotion recognition using HMM, first the database is sort out according to the mode of classification and then the features from input waveform are extracted. These features are then added to database. The transition matrix and emission matrix has been made according to the modes, which generates the random sequence of states and emissions from the model. Final is estimating the state sequence probability by using Viterbi algorithm [30].

C) Gaussian Mixture Models (GMMs):

Gaussian Mixture Models (GMMs) are considered good for evaluating density and for performing clustering [18]. The expectation-maximization algorithm is used for this purpose. GMMs are comprised of component functions called Gausses. The number of these Gausses in the mixture model is also referred to as the number of components. The total number of components can be altered based on the count of training data points. However, the model becomes more complex with the increase in the number of components.

D) K Nearest Neighbour (KNN):

A more general version of the nearest neighbour technique bases the classification of an unknown sample on the "votes" of K of its nearest neighbour rather than on only it's on single nearest neighbour. Among the various methods of supervised statistical pattern recognition, the Nearest Neighbour is the most traditional one, it does not consider a priori assumptions about the distributions from which the training examples are drawn. It involves a training set of all cases. A new sample is classified by calculating the distance to the nearest training case, the sign of that point then determines the classification of the sample. Larger K values help reduce the effects of noisy points within the training data set, and the choice of K is often performed through cross validation [31].

E) AdaBoost Algorithm:

AdaBoost algorithm is an adaptive classifier which iteratively builds a strong classifier from a weak classifier. In each iteration, the weak classifier is used to classify the data points of training data set. Initially all the data points are given equal weights, but after each iteration, the weight of incorrectly classified data points increases so that the classifier in next iteration focuses more on them. This results in decrease of the global error of the Classifier and hence builds a stronger classifier. AdaBoost algorithm is also used as a feature selector for training SVMs [32].

VI.CONCLUSION

The recent interest in speech emotion recognition research has seen applications in call centre analytics, human-machine and human robot interfaces, multimedia retrieval, surveillance tasks, behavioural health informatics, and improved speech recognition. In this study, the overview of SER methods are discussed for extracting audio features from speech sample, various classifier algorithms are explained briefly. Speech Emotion Recognition has a promising future and its accuracy depends upon the emotional speech database ,combination of features extracted from those database for training the model, types of classification algorithm used to classify the emotions in appropriate emotion class (e.g. happy, sad, anger, surprise etc.). This study aims to provide a simple guide to the beginner who's carried out their research in the speech emotion recognition.

VII.References

[1]S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, IITKGP-SESC : Speech Database for Emotion Analysis. Communications in Computer and Information Science, JIIT University, Noida, India: Springer, issn: 1865-0929 ed., August 17-19 2009.

[2] S. G. Koolagudi and K. S. Rao, "Exploring speech features for classifying emotions along valence dimension," in The 3rd international
Conference on Pattern Recognition and Machine Intelligence (PReMI- 09), Springer LNCS (S. C. et al., ed.), (IIT Delhi), pp. 537–542,
Springer-verlag, Heidelberg, Germany, December 2009.
[3] S. G. Koolagudi, S. ray, and K. S. Rao, "Emotion classification based on speaking rate," in The 3rd International Conference on Contemporary Computing, (Noida, India), JIIT university and University of Florida, August 2010.
[4] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition. Englewood Cliffs, New Jersy: Prentice-Hall, 1993.

[5] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech Communication, vol. 48, p. 11621181, 2006.
[6] S. R. M. Kodukula, Significance of Eecitation Source Information for Speech Analysis. PhD thesis, Dept. of Computer Science, IIT, Madras, March 2009.

[7] T. L. Pao, Y. T. Chen, J. H. Yeh, and W. Y. Liao, "Combining acoustic features for improved emotion recognition in mandarin speech," in ACII (J. Tao, T. Tan, and R. Picard, eds.), (LNCS 3784), pp. 279–285, Springer-Verlag Berlin Heidelberg, 2005.

[8] 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.

[9] S. Wu, T. H. Falk, and W. Y. Chan, "Automatic recognition of speech emotion using long-term spectro-temporal features," in 16th International Conference on Digital Signal Processing, (Santorini-Hellas), pp. 1–6, IEEE, 5-7 July 2009. DOI: 10.1109/ICDSP.2009.5201047.

[10] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using
gmms," in INTERSPEECH 2006 - ICSLP, (Pittsburgh, Pennsylvania), pp. 809–812, 17-19 September 2006.

[11] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," Speech Communication, 2010. Article in press.

[12] M. Sigmund, "Spectral analysis of speech under stress," IJCSNS International Journal of Computer Science and Network Security, vol. 7, pp. 170–172, April 2007.

[13] R. Nakatsu, J. Nicholson, and N. Tosa, "Emotion recognition and its application to computer agents with spontaneous interactive capabilities," Knowledge-Based Systems, vol. 13, pp. 497– 504, December 2000.

[14] V. A. Petrushin, "Emotion in speech:recognition and application to call centers," Proceedings of the 1999 Conference on Artificial Neural Networks in Engineering (ANNIE 99), 1999.

[15] E. Bozkurt, E. Erzin, C. E. Erdem, and A. T. Erdem, "Improving automatic emotion recognition from speech signals," in 10th Annual
Conference of the International Speech Communication Association (Interspeech 2009), (Brighton, UK), pp. 324–327, 6-10 September 2009
2009.

[16] N. Kamaruddin and A. Wahab, "Features extraction for speech emotion," Journal of Computational Methods in Science and Engineering, vol. 9, no. 9, pp. 1–12, 2009. ISSN:1472-7978 (Print) 1875-8983 (Online).

[17] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, "Approaching automatic recognition of emotion from voice: A rough benchmark," (Belfast), 2000.

[18] F. Dellert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," (Philadelphia, PA, USA), pp. 1970–1973, 4th International Conference on Spoken Language Processing, October 3-6 1996.

[19] Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. Psychological Bulletin, 99, 143-165.

[20] Banse, R., Scherer, K.R.: Acoustic Profiles in Vocal Emotion Expression. Journal of Personality and Social Psychology, 614–636 (1996)

[21] Ververidis, D., Kotropolos, C.: Automatic speech classification to five emotional states based on gender information. In: Proc. of Eusipco, pp. 341–344 (2004)

[22] Zhou y., Sun Y., Zhang J, Yan Y., "Speech Emotion Recognition using Both Spectral and Prosodic Features", IEEE, 23(5), 545-549, 2009.

[23] Yashpalsing Chavhan, M. L. Dhore, Pallavi Yesaware, "Speech Emotion Recognition Using Support Vector Machine", International Journal of Computer Applications, vol.1, pp.6-9,February 2010.

[24] Xia Mao, Lijiang Chen, Liqin Fu, "Multi-level Speech Emotion Recognition Based on HMM and ANN", 2009 WRI World Congress, Computer Science and Information Engineering, pp.225-229, March 2009.

[25] Yixiong Pan, Peipei Shen and Liping Shen, "Speech Emotion Recognition Using Support Vector Machine", International Journal of Smart Home, Vol. 6, No. 2, April, 2012.

[26] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification", in Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol. 1, pp. 593-596, Montreal, May 2004.

[27] F.Yu, E.Chang, Y.Xu, H.Shum, "Emotion detection from speech to enrich multimedia content", Lecture Notes In Computer Science,Vol.2195, 550-557, 2001.

[28] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, P.-J. Li, "Mandarin emotional speech recognition based on SVM and NN", Proceedings of the 18th International Conference on Pattern Recognition (ICPR?06), vol. 1, pp. 1096- 1100,September 2006.

[29] B. Schuller, G. Rigoll, M. Lang, "Hidden Markov model-based speech emotion recognition", Proceedings of the IEEE ICASSP Conference on Acoustics,Speech and Signal Processing, vol.2, pp. 1-4, April 2003.

[30] Ashish B. Ingale, Dr.D.S.Chaudhari, "SPEECH EMOTION RECOGNITION USING HIDDEN MARKOV MODEL AND SUPPORT VECTOR MACHINE".

[31] Muzaffar Khan, Tirupati Goskula, Mohmmed Nasiruddin ,Ruhina Quazi, "Comparison between k-nn and svm method for speech emotion recognition", International Journal on Computer Science and Engineering (IJCSE).

[32] Anurag Kumar, Parul Agarwal, Pranay Dighe1, " Speech Emotion Recognition by AdaBoost Algorithm and Feature Selection for Support Vector Machine".