EMOTIONS RECOGNITION FROM SPEECH USING THE TRAINED MODEL

Pramod Mehra[1], Parag Jain[2], Shushil Kumar[3]
[1,2,3]Department of CSE, UTU, Dehradun, Uttarakhand, India

Abstract

In this paper, the trained model is used to recognise the emotions from speech. As Speech has been used as an important mode of communication since the time immemorial. Emotions are an essential part of natural speech communication. Most of the present speech systems can process studio recorded neutral speech with greater accuracy. Therefore, a need is felt to update speech processing systems with the capability to process emotions. The component of emotion processing makes the existing speech systems more realistic and meaningful. In this work, spectral features are extracted from speech to perform emotion classification. Linear prediction cepstral coefficients, mel frequency cepstral coefficients and their derivatives (velocity and acceleration coefficients) are explored as features. Gaussian mixture models are proposed as classifiers. The emotions considered in this study are anger, happiness, neutral, sadness and surprise. The speech emotion database used in this work is semi-natural in nature, which has been collected from the dialogues of actors/actresses in popular Hindi movies.

KEYWORDS: emotion classification, spectral features, GMM, MFCC, LPCC, text dependent emotion recognition, text independent emotion recognition

I. INTRODUCTION

Speech has been used as an important mode of communication since the time immemorial. Emotions are an essential part of natural speech communication. Most of the present speech systems can process studio recorded neutral speech with greater accuracy. Therefore, a need is felt to update speech processing systems with the capability to process emotions. The component of emotion processing makes the existing speech systems more realistic and meaningful. Emotion recognition from speech utterances may be useful in different applications such as call center conversation analysis, entertainment, indexing of audio files based on emotions, development of effective human computer interaction and so on. Speech features may be basically extracted from excitation source, vocal tract or prosodic points of view, to accomplish different speech tasks. This work confines its scope to the use of spectral features for recognising emotions. Spectral features represent vocal tract information such as formant frequencies sequential variation in the shapes/sizes of vocal tracts, spectral bandwidths, spectral roll-off and so on. Generally, most of the speech tasks are accomplished successfully using spectral features. Various spectral features have also been explored for emotion analysis. To distinguish anger from neutral speech in Mandarin language, a combination of MFCCs, LPCCs, Rasta PLP coefficients and log frequency power coefficients (LFPCs) has been used as the features [3]. MFCC features from lower frequency components (20 Hz to 300 Hz) of speech signal have been used to model pitch variations .

Normally, spectral features are extracted through block processing approach. Entire speech signal is processed frame by frame, considering the frame size of around 20 ms, and a shift of 10 ms. It is assumed that with in this frame, speech signal is stationary in nature. Mel frequency cepstral coefficients are extracted as spectral features and used in this work for emotion analysis. In this work, UTU-semi-natural database (UTU-SNESC), collected from Hindi movies has been used.

II.DATABASES

From the available literature, three types of databases are used for analysis of speech emotions. They are simulated, elicited and natural and semi-natural speech databases. Emotional speech extracted from dialogues of actors and actresses of Hindi movies has been used to create semi natural emotion speech corpus known as Uttarakhand Technical University Semi Natural Emotion Speech Corpus (UTU-SNESC). The emotions expressed by actors in Hindi movies are close to real life emotion expression observed in the case of normal Hindi speaking Indian population.

For the creation of the database, dialogues of the popular actors have been extracted from Hindi movies. This database contains single and multi-speaker emotional utterances for male speakers. The emotions collected for this database are anger, happiness, fear, neutral, sadness and surprise. For single speaker database, video clips of different Hindi movies acted by the same actor are used. Later, audio tracks are separated and concatenated to make a single file. Adobe Audition is used to extract audio with mono channel frequency of 16 KHz and 16 bit resolution. Emotional speech has been extracted carefully, containing no background music and disturbances. Long silence regions have been removed with the help of wavesurfer without affecting the embedded emotions. Fifteen minutes of effective data is collected in this way for each emotion for male speakers. For multi-speaker database, the video clips of different Hindi movies are chosen irrespective of actors, but of the same gender and emotions. The Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus (IITKGP-SESC) has been collected by simulating eight emotions using neutral statements. Professional artists from All India Radio (AIR), Vijayawada, India have recorded this database in Telugu language. As the speakers are professional artists, they can simulate the appropriate emotions close to the real and practical situations. Since, all the artists are from the same organization, it ensures the coherence in the quality of the collected speech data. IITKGP-SESC is the first database developed in an Indian language (Telugu) for analyzing the basic emotions present in speech. This database is sufficiently large to analyse the emotions in view of speaker, gender, text and session variability [3].

III. EXTRACTION OF FEATURES

Proper feature extraction eliminates irrelevant features that hinder the recognition rates; it reduces the input dimensionality (and therefore improves generalization); it saves up the computational resources. Feature vectors can be long-time or short-time in nature. Long-time features are estimated over the entire utterance length. Short-time features are determined in a smaller time window (usually 20 to 30 msec). Contemporary research approach favors the long-time features for analysis of emotions [4], since the long time features correlate emotions better than short time ones.

The extraction of individual features and their use in developing the models has been discussed in the following paragraphs. Mel frequency cepstral coefficients (MFCCs), their derivatives, known as velocity (Δ) and acceleration (Δ - Δ) coefficients are separately used for emotion analysis. 21 MFCCs are extracted from speech signal. The above mentioned number of features used separately for developing emotion recognition models. Δ and Δ-Δ coefficients are used in concatenation with respective basic features to form the feature vectors. Hamming window has been used while framing the speech signal. The general block diagram of development of emotion recognition models (ERMs) is given in Fig. 2.

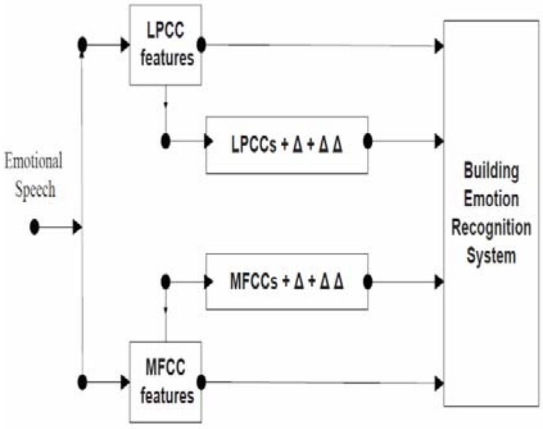

Fig. 1 Scheme of developing emotion recognition models

Human auditory system is assumed to process speech signal in a nonlinear fashion. It is well studied that lower frequency components of speech signalcontain more information. Therefore nonlinear mel scale filter has been designed to emphasize lower frequency components over higher ones. In speech processing mel frequency cepstrum is a representation of the short time power spectrum of a speech frame using linear cosine transform of log power spectrum on a non-linear mel frequency scale. Conversion from normal frequency 'f' to mel frequency 'm' is given by the equation :

$m = 2595 \log10(f/700+ 1)$

The algorithm used in this work for obtaining mel frequency cepstral coefficients (MFCCs) from speech signal is as follows :

1. Obtain Fourier transform of a speech segment to get short time spectrum.

2. Compute powers of the above spectrum within the triangular overlapping

windows placed according to mel scale.

3. Take logs of the power at each of the mel frequencies.

4. Compute the discrete cosine transform (DCT) of the list of mel log powers as if it were a signal. (Note: Basically one of the important applications of DCT is to discard negligible number of high frequency components during compression.)

5. The amplitudes of resulting spectrum give MFCC's.

IV. EMOTION RECOGNITION MODELS (ERMs)

The emotion recognition models are developed using male speaker utterances. 90% of the data is used for training the emotion recognition models and 10% is used for validation. Figure 1 illustrates the overall methodology of an emotion recognition system.

The process is divided into two parts: feature extraction phase and emotion recognition phase. From each speech utterance features (MFCCs) are extracted. Subsequently, feature vectors are formed. These feature vectors are given as input to the emotion recognition development phase. In the training stage, the feature vectors are used to train the GMM models. In the recognition stage, the feature vectors of test utterances are given to already trained models. Validation is done by giving test utterances to already trained models. Emotion Recognition Models are developed using MFCCs and their Δ and Δ - Δ features obtained from the speech signal.

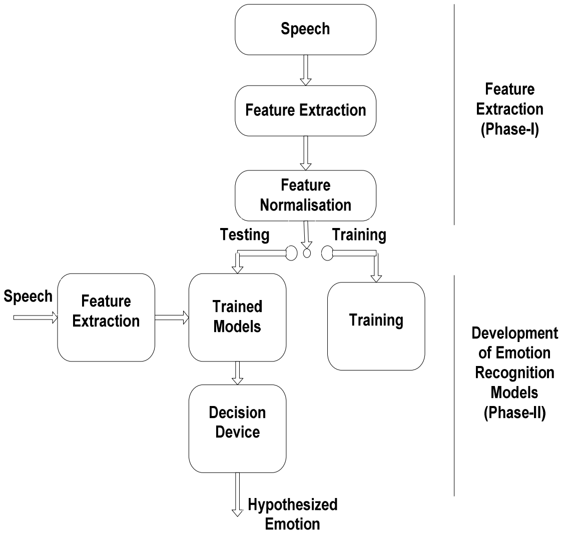

Fig. 2. Development of an emotion recognition system

V. Results and Discussion

While using UTU-SNESC, the text of the dialogues recorded from different movies was not same. Hence, the emotional utterances used for training and testing the GMM models contain different text. This is the reason that the recognition performance is quoted as text independent. In this case, it may be noted that the influence of phonetic information on emotion recognition is least. Table I shows the results of emotion recognition of same UTU – SNESC database using 21 MFCCs in the form of confusion matrix. In the confusion matrix, diagonal elements represent correct classification of emotions whereas other members in the row give us misclassification pattern.

As shown in Table I, 100% of anger utterances are classified as anger. 75% of happy utterances are classified as happiness and 25% are identified as anger. 75% of neutral utterances are classified as neutral and 25% are identified as sad. 100% of sad utterances are classified as sad. 50% of surprise utterances are classified as surprise and 50% are identified as happy.

TABLE I
COFUSION MATRIX % FOR SINGLE MALE SPEAKER USING 21 MFCCS OF SAME UTU - SNESC DATABASE

Emotion	Anger	Happy	Neutral	Sad	Surprise
Anger	100	0	0	0	0
Happy	25	75	0	0	0
Neutral	0	0	75	25	0
Sad	0	0	0	100	0
Surprise	0	50	0	0	50

Table II shows the results of emotion recognition of different UTU – SNESC using 21 MFCCs in the form of confusion matrix. In the confusion matrix, diagonal elements represent correct classification of emotions whereas other members in the row give us misclassification pattern.

As shown in Table 4.2, 100% of anger utterances are classified as anger. 75% of happy utterances are classified as happiness and 25% are identified as anger. 50% of neutral utterances are classified as neutral and 50% are identified as sad. 100% of sad utterances are classified as sad. 50% of surprise utterances are classified as surprise and 50% are identified as happy.

TABLE II

COFUSION MATRIX % FOR SINGLE MALE SPEAKER USING 21 MFCCS OF DIFFERENT UTU-SNESC DATABASE

Emotion	Anger	Happy	Neutral	Sad	Surprise
Anger	100	0	0	0	0
Happy	25	75	0	0	0
Neutral	0	0	50	50	0
Sad	0	0	0	100	0
Surprise	0	50	0	0	50

Table III shows the average emotion classification performance of same UTU – SNESC by which the model is trained and different UTU – SNESC, which is different than UTU – SNESC which is used in trained model.

As shown in table III the emotion recognition performance of UTU – SNESC is 80%, while the emotion recognition performance of different UTU – SNESC is 75%.

TABLE III
AVERAGE EMOTION CLASSIFICATION PERFORMANCE USING TRAINED MODEL OF SAME UTU – SNESC DATABASE AND DIFFERENT UTU - SNESC. CLASSIFIERS USED IS GMM.

DATABASE	Avg. Emotion recognition performance (%)
UTU - SNESC	80%
DIFFERENT UTU - SNESC	75%

VI. SUMMARY AND CONCLUSIONS OF PRESENT WORK

In this work, semi natural speech corpus, collected from Hindi movies has been used to characterize and classify the emotions. MFCC features along with their velocity and acceleration coefficients are used to capture emotion specific information from vocal tract of the speaker. The purpose of this study is to investigate the spectral features for their containment of emotion specific information. Gaussian mixture models, known to capture the distribution pattern of feature vectors are used as emotion classifiers. From the obtained results, it may be observed that 21 MFCCs, along with Δ and Δ - Δ features have given better results. It indicates that higher order spectral features contain better emotion specific information. The average emotion recognition performance for the same emotions in the case of same UTU – SNESC database is considerably high compared to the results of different UTU - SNESC mostly due to the influence of phonetic information on emotion recognition.

As a continuance of these studies, prosodic features may be used in combination with spectral features to further improve the emotion recognition performance of the models. Emotion recognition in real world scenario expects language independent emotion recognition. Therefore, there is a need to develop language independent emotion recognition systems as well.

REFERENCES

[1] D. Ververidis, C. Kotropoulos, and I. Pitas, “Automatic emotional speech classification,” in ICASSP, pp. I593–I596, IEEE, 2016.

[2] Shashidhar G Koolagudi and K. Sreenivasa Rao, “Emotion Recognition from Speech: A Review”, International Journal of Speech Technology, Volume 15, Issue 2, pp 99-117, Springer, June 2012.

[3] Van Bezooijen, “The Characterisitcs and Recognizability of Vocal Expression of Emotions”, Drodrecht,The Netherlands: Foris, 1994

[4] Hansen, J. H. L., Cairns, D. A. ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environments, Speech Communication 16, pp 391–422, 1995.

[5] D. O’Shaughnessy, Speech Communication Human and Mechine, Addison- Wesley publishing company, 1999.

[6] M. Schroder, “Emotional speech synthesis: A review,” in 7th European Conference on Speech Communication and Technology, (Aalborg, Denmark), Sept. 2001.

[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] Shashidhar G. Koolagudi, Sudhamay Maity, Vuppala Anil Kumar, Saswat Chakrabarti, and K. Sreenivasa Rao, ”IITKGP-SESC: Speech Database for Emotion Analysis” in IC3, CCIS 40, pp. 485-492, Springer-Verlag, Berlin, Heidelberg 2009.

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

[10] Li Y. and Zhao Y. Recognizing emotions in speech using short-term and long-term features. Proc. of the international conference on speech and language processing. pp. 2255-2258, 1998.

[11] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, Emotions of ser. Communications in Computer and Information Science. Springer, August vol. 40 2009, .

[12] K. S. Rao and B. Yegnanarayana, “Duration modification using glottal closure instants and vowel onset points,” Speech Communication, vol. 51, p. 1263-1269, doi:10.1016/j.specom.2009.

[13] 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.

[14] Rahul Chauhan, Jainath Yadav, S. G. Koolagudi and K. Sreenivasa Rao, Text independent emotion recognition using spectral features, Communications in Computer and Information Science (CCIS): Contemporary Computing, Vol. 168, Part-2, pp. 359-370, Springer, 2011.

[15] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition. Englewood Cliffs, New Jersy: Prentice-Hall, 1993.