IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 7, July 2016

A Survey on Emotion Recognition from Speech Signal

Sandeep Pathak[1], **Vaishali Kolhe**[2]

ME Student, Department of Computer Engineering, DYPCOE, Akurdi, SPPU, Pune, India[1]

Assistant Professor, Department of Computer Engineering, DYPCOE, Akurdi, SPPU, Pune, India[2]

Abstract: Emotion recognition has been studied in recent past with great interest, out of various modalities from which emotions can be extracted, speech is the most natural and fastest of all the modalities. The major challenges for making a Speech emotion recognition system are finding and preparing database, selecting the most suitable features and designing appropriate classification scheme. The common approach is to extract a very large set of features over a generally long analysis time window and perform machine learning methods for classification. Speech emotion recognition increases the naturalness in Human Computer Interaction and can be used in wide variety of application in our day to day life. This paper surveys the three main building blocks of speech emotion recognition system, first part is survey of existing databases, the second part surveys most widely used features and the third part discuss various classification techniques.

Keywords: Emotion recognition, Speech emotion recognition, Statistical classifiers, Emotional speech databases.

I. INTRODUCTION

Various information is contained in speech signal and therefore it is considered as a complex signal, it contains information about the message, speaker, language, emotions etc. A conversation has two parts verbal and nonverbal. During a conversation the nonverbal communication carries out important information like intention of the speaker. Using speech in emotion detection helps in not only utilizing the message but also how the message in conveyed e.g. The word “OKAY” in English can be used to express admiration, disbelief, consent, disinterest or an assertion etc. [8].

Speech emotion recognition is defined as extracting the emotional state of a speaker from his or her speech. It is believed that speech emotion recognition can be used to extract useful semantics from speech, and hence, improves the performance of speech recognition systems [10]. The basic goals of a speech emotion recognition system are a) Understanding emotions present in speech. b) Synthesizing desired emotion in speech according to the intended message [8].

The first challenge in making a Speech Emotion Recognition System (SER) is defining the word emotion. The word emotion is ambiguous and subjective in nature and uncertain in interpretation. Objectively defining the word emotion is difficult, and pose the first challenge in building a SER. To classify all the emotions is a difficult task as 300 emotional states are present in a typical emotion set. “Palette Theory” suggest that an emotion can be decomposed into primary emotions similar to a color which is a combination of two or more than two basic colours. The primary emotions are anger, fear, sadness, joy, disgust and surprise [1].

SER system have five main modules emotional speech input, feature extraction, feature selection, classification, and recognized emotional output [2] as shown in Figure 1.

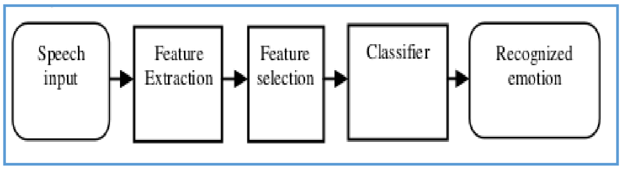

Figure 1 Block Diagram of SER

Feature selection is optional step, one more step is involved which is equally important is building of a database or selecting a existing a database which is discussed in section I. The motivation behind developing a SER decides the method of collecting or choosing a database [8]. Feature extraction step extracts the parameters contained by the speech and change in these parameters results in corresponding change in emotions. Hence extracting these features helps in narrowing down the classification problem and help classifiers to label the emotions more accurately.

There are various application of Speech emotion recognition in our day to day life. It can improvise naturalness in speech based human computer interaction. Accidents due to stressed mental state of driver can be avoided by alerting him/her while driving [2]. Analysis of call center conversation can be performed to study the behaviour of call attendant with the customer to increase the productivity [4]. It can be used in speech to speech translation of languages; source speech emotions are recognized and synthesized into the target speech [10].

DOI 10.17148/IJARCCE.2016.5787

Few challenges that lie in speech emotion recognition are:

1. It is difficult to define emotion. The word emotion is ambiguous and is subjective and uncertain in terms of interpretation. So objective definition is hard to form [12].
2. The SER system should be speaker and language independent but various features vary from language to language and speaker to speaker.
3. There are no standard speech corpora for comparing performance of research approaches used to recognize emotions [8].

II. DATABASE

Developing or selecting an existing database is prerequisite for building emotion recognition or synthesis system. Selecting and developing a database is highly dependent on the purpose of the emotion recognition system one wants to develop. The speech corpora can be divided into three categories a) Actor (Simulated) b) Elicited (Induced) c) Natural emotional speech databases [8].

Simulated speech databases are collected by recording artists or actors expressing linguistic neutral sentences in different emotions. The major advantage of these databases is they are standardized and result can be easily compared. The major disadvantage is, expression of emotion can be different in real world as opposed in acting e.g. LDC speech corpus [5] Emo-DB [7]. It is one of the easiest and reliable method of collecting database for speech related research and nearly 60% of the databases involved in speech related research are collected through this method. These are typically termed as full blown emotions and are more expressive than the real ones [8]. Elicited database are recorded by artificially creating emotional situations without the knowledge of the speaker. The anchor involves speaker in emotional conversation and speaker's reactions are recorded. Elicited databases are more natural expression of the emotions although all the emotions cannot be recorded which is a drawback of elicited databases is e.g. Wizard of Oz databases, ORESTEIA [13].

Natural databases are created by collecting the real world conversations such as Call centre conversation, conversation between patient and doctors etc. These databases are completely natural and can be used for real world emotion modelling [8]. The disadvantage of natural databases is they may not contain all the emotions and have copy right and privacy issues e.g. Call centre conversations [4].

The problem that researcher face in using existing databases are they do not simulate the emotions well enough, in some cases the human recognition accuracy for emotions was 65% [10]. There is no known standardize speech emotion database. Expression of emotion can be effected by culture and place where speaker live, so there can be same emotion but expressed differently in different parts of the world. While creating the speech corpora, labelling of soft emotions should be done carefully and after discussing with various experts. As soft emotions are highly subjective. Size of the corpora plays an important role for speech emotion recognition and it helps in deciding properties such as scalability and reliability of the Developed system. Most of the existing emotional speech databases used for developing emotion systems is too small in size [8].

III. FEATURE EXTRACTION

Selecting and extracting suitable features are one of the most crucial part in the pattern recognition system. The features are chosen for representing intended information. Since pattern recognition techniques are mostly dependent on problem domain selecting suitable features. One of the issues in feature extraction is selecting region, one method of selecting region is to divide the signal into small intervals, this is called frames and from each frames features are extracted, features obtained from such method are local features. Other method extracts global statics from whole speech utterance. There has been disagreement on superiority of global feature over local feature but majority of researcher have agreed to global being more better option for emotion recognition. The disadvantages of global features include loss of temporal information in the signal and limitation in classifying emotions with similar arousal e.g. Anger versus Joy.

A. Categories of Features

Features are chosen to represent intended information. Different features represent different speech information in highly overlapped manner. Speech features are divided in four categories: Prosody features, Vocal Tract Features and Excitation Source Features [8].

B. Prosody Features:

The most popular prosody features includes, Fundamental Frequency (F$_0$), Energy, Duration, Formants. Human beings impose duration, intonation, and intensity patterns on the sequence of sound units, while producing speech. These prosody constraints make human speech natural. It can be viewed as speech features related with syllables, words, phrases and sentences. Prosody can be considered as super segmental information. Energy, intonation and pattern of duration are acoustically represented by prosody [8].

C. Vocal Tract Features

Vocal Tract Features are obtained by analysing the characteristics of Vocal tract, which are well reflected in frequency domain analysis of speech signal. Fourier transformation of speech frame provides short time spectrum. Cepstrum is obtained by performing Fourier transform on log magnitude spectrum. MFCCs (Mel frequency cepstral coefficients) and LPCCs (Linear prediction cepstral coefficients) are some features that are derived from the cepstral domain and provides vocal tract information. The emotion specific information present in the sequence of shapes of vocal tract may be responsible for producing different sound units in different emotions [8].

Characteristic of a particular channel of a individual personal is represented by LPCCs and the characteristic get changed according to emotions, using this property emotions can be extracted. Advantage of using LPCCs is, less computation is required and can describe vowels in better manner. MFCCs is popularly used in emotion recognition and provide good accuracy when used for emotion recognition. In low frequency region better frequency resolution and robustness to noise could be achieved with the help of MFCC rather than that for high frequency region [14] [1].

D. Excitation Source Feature

Excitation source features are obtained from speech signal after suppressing vocal tract (VT) characteristics. To extract excitation source features first VT information is predicted using filter coefficient (linear prediction coefficients (LPCs)) and then separating it by inverse filter formulation the result obtained is called linear prediction residual and contains information mostly about excitation source [9] [8].

The glottal activity characteristics such as closed and open phases of glottis excitation and strength are explored by sub-segmental analysis of speech signal. The glottal activities specific to the emotions can be estimated using excitation source features. Glottal volume velocity is obtained by integrating the LPCs. Excitation source features are not popular in speech emotion recognition. The reasons are 1) Spectral Features are more popular 2) LP features can be confused with errors or noise.

LP residual extraction gives the primary excitation to the vocal tract system, while generating speech signal. Higher order correlation exists among the LP residual samples, these correlations can be utilized to some extent, by source features.

Excitation source features are least used of all features in SER. Although it contains all the information like speaker, language and emotions etc. but it cannot compete with established prosody and vocal tract feature.

IV. CLASSIFICATION

The next part of speech emotion recognition is classifying the emotions in the speech utterance. Different classifier has been used for classifying the emotions by various researchers like Hidden Markov Model (HMM), Gaussian Mixture Model and Support Vector Machine (SVM) [10]. Different Neural Network model have also been like Convolution Neural Network (CNN) [15], Multilayer Perceptron (MLP), and Restricted Boltzmann Machine (RBM) [6] have been used recently for emotion recognition. Although a conclusion cannot be made which classifier is best, each classifier has its own advantage and disadvantage artificial neural network is used for emotion recognition because of its property of finding nonlinear boundaries separating the linear states. ANN reached the accuracy of 51.19% in speaker dependent recognition, and 52.87% for speaker independent [5] [10]. MLP which is a class of neural network provided the accuracy rate of 68.10% for Leave One Text out (LOTO) scheme for testing and 51.65% for Leave One Speaker out (LOSO) scheme; both tests were conducted for speaker independent emotion recognition. MLP is a popular classifier in emotion recognition because it is easy to implement and well defined training algorithm once the architecture of ANN is defined. Similar result were obtained for deep learning method DBN and RBN with accuracy of 69.14% when tested for LOTO and 64.32% for LOSO scheme for speaker independent emotion recognition using the MFCC and prosody features [5]. CNN provided 79% accuracy for speaker independent emotion recognition system [15].

One of the most popular classifier in speech related research is Hidden Markov Model (HMM) because speech signal production mechanism is physically related to it. HMM provides good result in modelling temporal information in speech spectrum. The process of HMM is doubly stochastic and consist of first order markov chain which is hidden from the observer. A random process is associated to each state which generates the observation sequence capturing the temporal structure of the data. HMM is trained for each emotion and an unknown sample is classified according to the model which illustrates derived feature sequence the best. HMM when used as classifier for emotion recognition provides accuracy of 76.12% for speaker dependent system using spectral features for speaker independent emotion recognition HMM provided the accuracy of 64.77%. The drawback of using HMM classifier is the feature selected should not only contain information about emotion but also fit the HMM structure as well. HMM classifier have lower recall rate when prosody and formant features are used than that of classifiers using spectral features [10] [1].

Kernel functions are used in transforming the original feature set to a high dimensional feature space and it is the main thought behind the working of SVM classifier and leads to get optimum classification in the new feature space. SVM classifiers are widely used in task of emotion recognition and performed better than other classifiers. SVM accuracy for speaker dependent speech emotion recognition system was found to be 80% and 75% for speaker independent system.

When global features are considered Gaussian Mixture Model tends to be more suitable for emotion recognition from speech signals. GMM is a probabilistic model for density estimation using convex combination of multi-

DOI 10.17148/IJARCCE.2016.5787

variate normal densities. GMM can be considered as special case of continuous HMM with only one state. Due to low requirement for their training and testing and efficiency in modelling multi-modal distributions, GMM are very efficient in emotion recognition when global features are extracted from training utterances. GMM provided maximum accuracy of 78.77% using the best features. Accuracy of 75% was achieved for speaker independent recognition and 89.12% for speaker dependent recognition using GMM [10].

To remove the shortfall of GMM in modelling temporal structure of data Vector autoregressive (VAR) method process was coupled with GMM the combination was named as Gaussian mixture vector autoregressive model (GMVAR). GMVAR was used to classify anger, fear, happiness, boredom, sadness, disgust, and neutral emotions from Berlin Emotional Speech Database. GMVAR provided 90% accuracy in classifying high and low emotion arousal as compared to 86.00% for the HMM technique [11].

V. CONCLUSION

Features and classifiers used in building SER has been studied. Processing emotions from speech and adding it to existing speech system increases naturalness of the system. Important steps in building SER after selecting or developing the database is extracting the features containing the information about emotions and selecting the appropriate classifier for recognizing the emotions from speech. A consensus have not been made on the superior feature set and the best classifiers as each feature and classifier has its own set of advantage and disadvantage. Methodology for building or selecting the database depends upon the purpose of the SER system to be developed. Speaker dependent classification of emotion is generally easier than speaker independent emotion recognition.

Accuracy of SER can be increased by using combination of given methods and also by extracting more effective features. Multi Classifier system (MCS) and combination of different features can be used in order to increase the accuracy and robustness of emotion recognition system.

REFERENCES

[1] Ashish B. Ingale, D. S. Chaudhari, “Speech Emotion Recognition” International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[2] B. Schuller, G. Rigoll, M. Lang, M, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture.” In Proc. IEEE int. conf. acoust., speech, signal processing (pp. 577–580). New York: IEEE Press, 2004.

[3] C. Breazeal, L. Aryananda, “Recognition of affective communicative intent in robot-directed speech”, Autonomous Robots 2, 83–104, 2002.

[4] C. M. Lee, S. S. Narayanan, “Toward detecting emotions in spoken dialogs”. IEEE Transactions on Audio, Speech, and Language Processing, 13, 293–303, 2004.

[5] D. Ververidis, C. Kotropoulos, “A state of the art review on emotional speech databases.”, In Eleventh Australasian international conference on speech science and technology, Auckland, New Zealand, Dec. 2006.

[6] E. M. Albornoz, M. Sánchez-Gutiérrez, F. Martinez-Licona, H. L. Rufiner, J. Goddard, “Spoken Emotion Recognition Using Deep Learning”, Lecture Notes in Computer Science, Vol. 8827, pp. 104- 111, 2014.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, “A database of German emotional speech, in: Proceedings of the Interspeech 2005, Portugal, 2005, pp. 1517–1520.

[8] G. Shashidhar, K. Koolagudi, Sreenivasa Rao, “Emotion Recognition from speech: A Review”, International Journal of Speech Technology, 15, pp. 99–117, 2012.

[9] J. Makhoul, “Linear prediction: A tutorial review.”, Proceedings of the IEEE, 63(4), 561–580, 1975.

[10] M. E. Ayadi, M. S. Kamel, F. Karray, “Survey on Speech Emotion recognition: Features, Classification Schemes, and Databases”, Pattern Recognition 44, PP.572-587, 2011.

[11] M. M. H. El. Ayadi, M. S. Karmel, F. Karray, “Speech Emotion recognition using Gaussian mixture vector autoregressive models”, in ICASSP vol. 4 pp. 957-960, 2007.

[12] M. Schroder, R. Cowie, “Issues in emotion-oriented computing toward a shared understanding”. In Workshop on emotion and computing (HUMAINE), 2006.

[13] McMohan E., Cowie R., Kasederidis S., Taylor J., Kollias S., “What chance that a DC could recognize hazardous mental state from sensor inputs?”, In Tales of dissapearing computer, Santorini, Greece, 2003.

[14] P. Shen, Z. Changjun, X. Chen, “Automatic Speech Emotion Recognition Using Support Vector Machine”, International Conference On Electronic And Mechanical Engineering And Information Technology, 2011.

[15] Zhengwi Huang, Ming Dong, Qirong Mao, Yangzhao Zhan. “Speech Emotion Recognition Using CNN”. In Proceedings of the ACM International Conference on Multimedia, pages 801-804. ACM 2014.

DOI 10.17148/IJARCCE.2016.5787