Speech Emotion Recognition using Non-linear Classifier-A Review

Somi Kolita
Research Scholar,
Department of IT, School of Computing Sciences,
The Assam Kaziranga University

Dr. Purnendu Bikash Acharjee
Assistant Professor,
School of Computing Sciences,
The Assam Kaziranga University

Abstract: In the field of Human Computer Interaction(HCI), identifying emotion from speech is a very significant topic and it has been achieving progressive interest in current research area. Therefore, Different researchers have been introduced and established many systems and methods to recognize the emotion from human speech. In this paper, we have discussed about some earlier implementation and advance made of emotion recognition. Also reviewed various feature extraction methods through which to extract feature vector and classifiers for emotion recognition. The classifiers are able to distinguish the six primary types of emotion. Different types of classifier performance for speech emotion recognition are also discussed.

Keywords: HCI, Emotion recognition, Prosody, Classifier.

1. INTRODUCTION:

Speech is vocalized form of communication between two individuals. It is a ability to express our ideas, feelings in front of other human beings. Speech is a signal of complexity which contains information about speaker, language, message etc. Many of speech processing system process studio recording, neutral speech effectively, their performance are very poor in the case of emotional speech, because of difficulty in modelling and characterization of emotions present in speech [1]. Presence of emotion makes speech more real. Emotion recognition from speech plays an important role in recent research world. It is a most crucial topic in the field of human computer interaction(HCI) and commonly used to developed many different applications in e-learning field as for example, determining students emotion timely and making proper treatment can enhance the teaching quality. Using HCI, we can develop the interaction between computer and user by making computer more responsible with users needs. Now a days, HCI system has been designed to identify "Who is speaking" or "What is speaking". Now recent years developments, computers are given ability to detect emotion of human then they can know "How human is speaking" and can behave accurately and naturally.

Importance of emotion recognition system is that in case of the absence of a person the system can identify his or her emotional state through speech. It is not necessary for the person to get face to face with the system. There are some barriers which increase the difficulties in order to get the more accurate output from emotional speech input. If it is not sure which speech features are need to be taken to distinguish between various emotional states then getting the exact output may be difficult. Speech features directly gets affected by speaker, speaking style, speaking rate, language, sentences. Changing of speaker and their environment and culture is also a big challenge in speech emotion recognition. With the changing culture and environment, the speaking style, speaking rate, etc., may also get changed [1, 2].Pronunciation variance in emotional speech also matters in order to detect the underlying emotions in speech. There are many factors which create an effect on word pronunciation, for instance, the gender of the speaker, speaker age, word position within the utterance and dialect [4].The application area of speech emotion recognition is very vast, few of its important applications are: psychiatric diagnosis, lie detection, intelligent toys, conversation with robots, identifying the emotional state of customer may help to enhance quality of service in call centres [1, 2, 5]

In this paper, we are going to discuss some fundamental things about the speech emotion recognition system. Sect.2 of the paper consists of basic working procedure of speech emotion recognition system. Section 3 describes the categories of dataset. In Sect. 4, various feature extraction techniques have been discussed. Section 5 describes classifiers; Sect. 6 contains discussion about the review.

2. EMOTION RECOGNITION SYSTEM

Speech emotion recognition system is a typical type of pattern recognition system which aim is to identify the emotional state of human beings automatically from his or her voice [6]. Appropriate feature extraction and proper classifier selection from the sample emotional speech is very important in order to get the appropriate emotional output. There are five main modules avalible in Speech emotion recognition system ,which are emotional speech input, feature extraction, feature selection, classification, and recognized emotional output. The basic structure of the speech emotion recognition is as shown in Figure below.

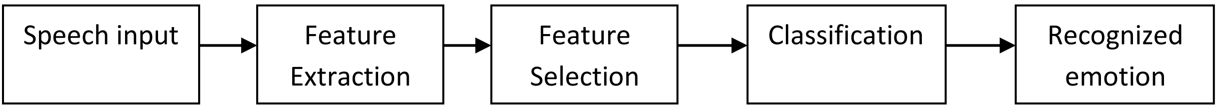

Figure. Structure of the Speech Emotion Recognition System.

The whole speech emotion recognition system may be speaker dependent and speaker independent. A system is said to be speaker dependent if it is developed to work for a specific speaker, and on the other hand, a system is said to be speaker independent if it is designed to work for any speaker. The analysis of the speech emotion recognition system is based on the naturalness of the database which is used as an input to the speech emotion recognition system [7].

3. CATEGORIES OF DATABASE

Database is very important part of the speech emotion recognition System. There are 300 emotional states are included in typical dataset, so it is very difficult to analysis and identify emotions from these huge number of emotional dataset, for this condition emotions are classified six basic types namely neutral, Anger, happy, fear, surprise and sadness.

There are three types of emotional database used in emotion recognition system, i.e acted or simulated emotional data, elicited emotional data and natural emotional data.In case of actor-based emotional speech dataset, data are collected by asking any trained actor or professional to speak with a specific type of emotion. This type of emotion is also known as full-blown emotion. There are many actor-based speech emotional dataset available but some are most commonly used and publicly available like Berlin Emotional Speech Dataset, Danish Emotional Speech Dataset, and Electromagnetic articulography (EMA) dataset. Elicited emotional speech dataset is collected by creating an artificial emotional situation, without speaker information. Since collecting this type of data is very much exhausting so only few numbers of elicited speech datasets are available. Natural emotional speech datasets are taken from real-life scenario like call centre conversation, cockpit recordings, etc.

4. FEATURE EXTRACTION

Feature extraction is the main part of emotion recognition system. It is a method of extracting those feature from input signals that help the system to identify the appropriate output. In the field of speech emotion recognition, choice of appropriate feature vector is very important which is able to recognize the exact emotion type. Feature vector can be categorized into two types: long-time feature vector and short-time feature vector. Long-time feature vector is calculated based on the whole length of utterance, as opposed to short-time feature vector is calculated based on window which is generally less than 100 ms.[1-3].

It depends on the partition of speech into small intervals called frames over which the statistical properties of the speech waveform are observed to be relatively constant. Features extracted from these frames are referred to as ‘segmental’ features or typically spectral features. Spectral features refer to short time spectral representation of speech signal. Distribution of the spectral energy over certain frequency range has impact on the emotional content of the utterance[4] Some of the spectral features those are Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), Log Frequency Power Coefficients (LFPC), Mel Energy Spectrum Dynamic Coefficients (MEDC), Zero Crossing Rate (ZCR) etc.

In [8], LFPC was used to represent the speech signal into respective feature vectors and fundamental frequency; it was used as a representative spectral feature in [5]. MFCC was initially used in [9] and finds detailed description in [12] .As a spectral feature, MFCC has also been extracted in [4] and [8]. The delta and double delta MFCC features also help in improving the accuracy since these considers the temporal variations of the spoken signal. In this papers, MFCC feature vector gave the best performance accuracy then the other feature set.

Prosodic features are also called as acoustic features, which is the primary indicator of the speaker’s emotional state. Prosody is referred as the supra-segmental phonology which can be viewed as speech features associated with larger unit such as syllables, words, phrases and sentences [7]. Emotional prosody features used to encode information at least from two sources, i.e., emotion and linguistics [6]. Research in the field of psycholinguistics shows that prosodic information like speaking rate and pitch is very much significant in human identification of underlying emotions in speech signal.The contours of prosodic derived in the research of ASER include in general minimum, maximum, median and interquartile range [6].Some commonly used prosodic features are pitch, energy, intensity, speaking rate of speech signals etc. Pitch, which is referred as fundamental frequency, better referred to the vibration rate of the vocal folds. Basically, anger got the variance of pitch, high-frequency formats, mean value of energy, and higher mean value. On the other hand, happy state has an improvement in variance of pitch, variation range, mean value, and mean value of energy. Sadness has low mean value, variation range and variance of pitch, energy is weak, speak rate is slow and spectrum of high-frequency components reduces. Fear has a high mean value, higher energy, and high variation range of pitch. So the statics of pitch, energy, formats and some important spectrum feature can be extracted in order to identify the emotion from speech signal[6].In [9],extracting the pitch, energy and formant features from their input signals and mentioned that energy plays a vital role in emotion recognition, e.g speech signals corresponding to angry and happiness has much higher energy than those of sadness. Formant frequency are defined as resonance in vocal tact and they determine characteristic timbre of vowel. The peaks of the frequency response from a linear prediction filter are the formant [11].

The extraction of all basic speech features for emotion recognition may not be necessary. After the extraction of feature vectors when we give the entire extracted feature as input to the classifier, it gives no guarantee that the system will give exact performance. It is very much important to extract a significant feature vector which is able to give large emotional information about the speech signal. Forward selection (FS) method could be used for selecting the significant feature subset. In the first step of the forward selection method, it initializes with the single best feature out of the whole features and for classification validity the other feature can be added in future [1–3].

5. CLASSIFIERS

Selection of proper classifier is very important in order to get the exact emotional output. So after selecting the feature vector, the most important task is to select the appropriate classifier. No fixed standard is there for classifier selection; it depends on the geometry of the input vector. There are different types of classifier for speech emotion recognition system. Gaussian Mixtures Model (GMM), Support Vector Machine (SVM), Hidden Markov Model (HMM), K-Nearest Neighbours (KNN), Artificial Neural Network (ANN), etc., are some of the widely used classifiers in emotion recognition system. Each classifier has its own advantages and disadvantages over others.

In the field of emotion recognition according to the various researchers, Only when the global features are extracted from the training utterances, Gaussian Mixture Model is more suitable for speech emotion recognition. For this case, maximum accuracy of 78.77% could be achieved using GMM. In speaker independent recognition typical performance obtained of 75%, and that of 89.12% for speaker dependent recognition using GMM[12,13]. The accuracy rate of HMM in case of speaker-dependent classification is 76.12% and in case of speaker-independent classification, it is 64.77. ANN classifier accuracy is 52.87% in case of speaker-independent classification and in case of speaker-dependent classification accuracy rate is 51.19% [1–3, 5]. KNN has the accuracy rate of 64% for four emotional states by using the feature vector like energy contours, pitch[1, 2].For SVM, the accuracy rate is obtained for the speaker independent and dependent classification are 75% and above 80% respectively in some experiments. Some of the study groups considered a gender independent ASER system while some like in [14] attempted to bifurcate the database into male and female utterances over which the ASER system was worked upon to accordingly achieve an accuracy of 79.9% and 89.02% for female and male utterance respectively. But according to table I, in recent years ANN based emotion recognition got better accuracy than other classifiers. Following table I provide the list of some classifiers used for emotion recognition system.

Table I: Literature review on use of different classifiers in different research

Study group and year	Database	Extracting feature	classifier	Accuracy rate
X. Cheng et al [14](2012)	Mandarin emotional speech database	Pitch, MFCC	GMM	79.9% (female) 89.02% (male)
T. Seehapoch et al [15] 2013	Berlin, Japanese and Thai	F0,Energy, ZCR, MFCC	SVM	89.8%
Li Liu et al[16] (2014)	Chinese tweets from Sina Weibo	MI, CHI, TF-IDF and ECE.	HMM	78% accuracy
Monorama Swain et al[10](2015)	Odiya Speech – sambalpuri, cuttaki	MFCC, Delta MFCC, LFPC	SVM	They got 82.14% accuracy with SVM only using MFCC.
Akash shaw [11] (2016)	Real users engaged with a machine agent.	Energy, Pitch, Formant frequency, MFCC	ANN	They got 85% accuracy
Raviraj Vishwambler[17] (2017)	Marathi Speech Database	Cepstral features,NMF,Pitch	ANN	78%
Vishnu Vidyadhara Raju Vegesna et al[18](2018)	Telegu speech corpus	MFCC,Pitch, Intonation	GMM HMM	75% recognition rate for all observation
S.S Poorna et al[19](2018)	South Indian language	LPC,Energy	ANN, SVM KNN	98.32% 94.84% 81.75%

6. SUMMARY AND DISCUSSION

Automatic emotion recognitions from the human speech are increasing now a day because it results in the better interactions between human and machine. In this study, we have reviewed some important emotion recognition technique, few commonly used feature extraction approach and classifiers. We also noticed that accuracy and efficiency of emotion recognition system depends on the appropriate feature extraction and classifier selection. The introduced methods gives idea about both capable of a rather reasonable model for the classifiers for speaker independent system is less than that for the speaker dependent system So it is required to improve the classifier performance in case of speaker-independent classification.. On other hand, some of the study consider gender independent ASER model then they got the automatic recognition of human emotions in speech.The method used in various papers defined a reasonable domain for each detected word sample. The important issue in speech emotion recognition system are noticed that the average accuracy of the most of the accuracy of 79.9% and 89.02% for female and male utterance respectively. The difference in accuracy can be supported by the fact that with the features of pitch and MFCC being extracted, a non-linear analysis of speech signal with male pitch being lower than female pitch is occurring. Higher the pitch, greater should be the physical difference between the two tones to be perceived differently and if this is not being achieved, the huge difference between the accuracies with respect to the male and female utterances cannot be accounted for. For pitch and formant, it is not likely to examine the accuracy for each frame, so the reasonable ranges for pitches and formants are defined to filter other error values out. The results obtained in this study demonstrate that more effective feature extraction can give a higher accuracy rate in speech emotion recognition system. Also the combination of various methods will improve the accuracy rate.

REFERENCES

[1] Abhijit Mohanta et al(2015), “Human emotion recognition through speech”. Adv. Comput. Sci. Inf. Technol. (ACSIT) 2(10), 29–32

[2] Akshay S. Utane et al(2013), “Emotion recognition through speech”. Int. J. Appl. Inf. Syst.(IJAIS) 5–8

[3] Dr.Varsha S. Joshi et al(2013), “Speech emotion recognition: a review”IOSR J. Electron. Commun.Eng. (IOSR-JECE) 34–37

[4] Shashidhar G. Koolagudi et al(2012) “Emotion recognition from speech: a review”. © Springer Science + Business Media

[5] I. Luengo and E. Navas, “Feature analysis and evaluation for automatic emotion identification in speech”, IEEE Trans. on Multimedia, vol. 12, no. 6, pp. 267-270, Oct 2010.

[6] Uzzal Sharma et al(2018), “Detection of Human Emotion from Speech—Tools and Techniques”.© Springer Nature Singapore Pte Ltd.

[7] Ashish B. Ingale et al(2012), “Speech Emotion Recognition”, International Journal of Soft Computing and Engineering (IJSCE).

[8] Preeti Suri et al(2014), “Enhanced HMM speech emotion recognition using SVM and neural classifier”. Int. J. Comp. Appl. (0975–8887) 17–20

[9] Bjorn Schuller et al(2003), “Hidden Markov Model based Speech Emotion Recognition”©IEEE2003

[10] Monorama Swain et al(2015), “Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition”© Springer science+Business Media New yark 2015

[11] Akash Shaw et al(2016), “Emotion Recognition and classification in Speech using Artificial Neural Networks”. International Journal of computer Applications.

[12] Aditya Bihar Kandali et al (2009), “Emotion Recognition from Assamese language using MFCC features and GMM classifier”.©IEEE.

[13] Aditya Bihar Kandali et al (2009), “Vocal emotion recognition in five native languages of Assam using new wavelet features” © Springer Science+Business Media.

[14] X. Cheng et al (2012), “Speech emotion recognition using Gaussian
Mixture Model”, 2nd Intl. Conf. on Computer Application and System
Modeling.

[15] T. Seehapoch et al(2013),“Speech emotion recognition using Support Vector Machines”, 5th IEEE Intl. Conf. on Knowledge and Smart Technology (KST).

[16] Li Liu et al (2014),“A Self-Adaptive Hidden Markov Model for Emotion Classification in Chinese Microblogs” Mathematical Problems in Engineering , Article ID 987189, 8 pages.

[17] Raviraj Vishwambhar Darekar et al(2018), “Emotion recognition from Marathi speech database using adaptive artificial neural network” , Biologically Inspired Cognitive Architectures.

[18] Vishnu Vidyadhara Raju Vegesna et al(2018)“Application of Emotion Recognition and Modification for Emotional Telugu Speech Recognition”, © Springer Science+Business Media, LLC, part of Springer Nature.

[19] S. S. Poorna et al(2018)“A Weight Based Approach for Emotion Recognition from Speech: An Analysis Using South Indian Languages”, © Springer Nature Singapore Pte Ltd.