Speech Emotion Recognition: A Review

Surabhi Vaishnav[1]**, Saurabh Mitra**[2]

[1]*M.Tech Scholar, Dept. of Electronics & Communication, Dr. C. V. Raman University, Bilaspur, Chhattisgarh, India*
[2]*H.O.D. at Dept. of Electronics & Communication, Dr. C. V. Raman University, Bilaspur, Chhattisgarh, India*

---***---

Abstract - *Speech is the most prominent way of communication between individuals. Human brain can do reception & interpretation of the sound wave very easily. For an effective & natural human-machine interaction (HMI), emotion recognition plays a vital role. It adds ornaments to the communication. In virtual world, emotion recognition could help simulate more realistic interaction.*
There are several spectral and temporal features extracted from human speech. The existing methods for emotion detection from voice use mainly MFCC & Energy feature. This paper briefs a review about the existing work on speech emotion detection useful for carrying out further research.

General terms— Speech Emotion Recognition, Feature Extraction, Classifier, MFCC.

Keywords— Human-Machine Interaction, Emotion Recognition, Speech Processing, ANN, SVM.

1. INTRODUCTION

People have been speaking to each other for thousands of years. In recent time, Human-machine interaction (HMI) has become a growing area of innovation in industry as well as academic field [2]. Speech is one of the fundamental ways of communication known to mankind. A speech signal is a logical arrangement of sounds. Our brain performs a complex set of analyses of auditory input (i.e. sounds). It converts the sounds into some conceptual ideas and thoughts which forms the basis of instructions, commands, information & entertainment.

Automatic recognition is often studied in sense of identifying emotion among some fixed set of classes. Speech emotion recognition is a kind of analyzing vocal behavior. The speech processing involves three main steps i.e. preprocessing, feature extraction and pattern recognition. In case of speech signal, vowels carry the most of the informative part. Vowels are mainly voiced part of the spoken words. Therefore it is customary to separate out voiced part from unvoiced part of the information spoken and proceed further with signal processing on only voiced part.

For an effective and natural HMI, emotion recognition plays a vital role. Emotions reflect the mental state of the person through speech, facial expressions, body postures and gestures and also other physical parameters like body temperature, blood pressure, muscle action, etc. The mental state of the person indirectly affects the speech produced by the person. E.g. in human-human interaction, speech rate is faster in case of anger/ joy and pitch range is also wider while in case of sadness, speech is slower with lower pitch range. Therefore, emotion detection in speech is advantageous in various applications.

1.1 Speech Emotion Recognition System

Speech emotion recognition is nothing but an application of the pattern recognition system in which patterns of derived speech features such as Pitch, Energy, MFCC are mapped using classifier like ANN, SVM, HMM etc.

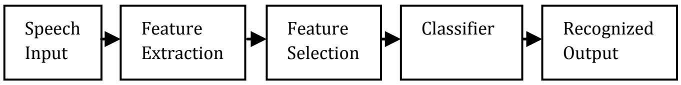

Fig – 1: Structure of Speech Emotion Recognition

The system contains five major modules: speech input database, feature extraction, feature selection, classifier & recognized output as illustrated in figure – 1 above. Overall, the system is based on deep analysis of the generation mechanism of speech signal, extracting some of features which contain information about speaker's emotion & taking appropriate pattern recognition model to identify states of emotion.

Typically, a set of emotion having 300 emotional states [6]. Whenever, signal is passed to the feature extraction & selection process, the extracted speech features are selected in terms of emotion relevance. Allover procedure revolves around the speech signal for extraction to the selection of speech features corresponding to emotions. Forward step is generation of database for training as well as testing of extracted speech features. At the end, detection of emotions has been done using classifier with the usage of pattern recognition algorithm.

The Speech emotion recognition is similar to the speaker recognition system but different types of approach to detect emotions make it secure & intelligent. The evaluation of the system is depending on naturalness of the input database.

1.2 Speech Parameters

Some of the Time domain parameters been applied in recent researches from speech emotion recognition are briefly given.

Energy: In speech processing, Energy refers to the variations in amplitude of some speech signal. The voiced & unvoiced signal contain distinguish energy level so that this could easily identified with the usage of this feature. Also, initials & end of the silent part can be found out.

Pitch: Pitch is frequency of vibration of vocal cords of any voice signal. As during speech production process, our glottis opens and then closes with some period; such period is pitch period. Pitch range is 50-400Hz typically. It can be estimated by quantifying the period or measuring the harmonics. Fundamental frequency is a feature i.e. reciprocal of pitch.

Formant: In speech wave, a concentration of acoustic energy around a particular frequency is formant. The vocal tract behaves like a resonance chamber that always amplifying & attenuating certain frequencies. The spectrum of vocal tract response consists of number of resonant frequencies i.e. Formants. It ranges from 200 to 1000 Hz, 1000 to 2000 Hz and 2000-3500 Hz. This given formants carries informative information so that usually considered whereas the fourth one never considered. Each formant corresponds to a resonance in the vocal tract.

Mel-frequency cepstrum coefficient (MFCC): A unique representation of spectral property of voice signals. These are the best for speaker/speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. The computation of MFCC explained in article by Mirlab[11]. An article about Spectrogram deals that Mel-frequency scale represents subjective pitch i.e. perceived pitch. Also it takes certain properties of the human auditory system [7].

2. LITERATURE

A lot of research works have been performed in identification of emotion through speech processing. Performance of speech recognition systems is usually evaluated in terms of accuracy and speed. Whatever we know about a human speech processing is still very much limited & yet to witness a total unification of speech technology & science behind it. Some of that work has already been done by many people worldwide.

Tin Lay New et. al proposed a text independent method of speech emotion classification, which makes use of short time log frequency power coefficients (LFPC) to represent the speech signals and a Hidden Markov Model (HMM) is used as a classifier. A system for classification of emotional state of utterances is proposed. Six categories of speech emotions are applied in such work. A database of 60 emotional utterances, each from 12 speakers is constructed and is used to train and test the system. Average accuracy of 78% achieved & analyzed performance using LPCC and MFCC. Also goes with the fact that grouping of emotions with same characteristics enhance system performance [3].

Kamran Soltani et al studied the importance of the psychology and linguistics in spoken language man-machine interfaces. Along with the techniques in signal processing and analysis, it also requires psychological and linguistic analysis. The work makes use of six emotions, happiness, sadness, anger, fear, neutral and boredom. It uses pitch i.e. speech fundamental frequency, formant frequencies, energy and voicing rate as features. These speech parameters were used to train neural network classifier and the Berlin Database of Emotional Speech. Average accuracy of 77% is achieved. The work concludes that anger and neutral can be recognized easily while fear the most difficult one [2].

Jana Tuckova et. al performed experimental analysis using parameters like fundamental frequency, formant frequency and statistical analysis was conducted for multi-layer neural network (MLNN). The average accuracy obtained using this technique is 75.93% for multiword sentences while that for one word sentences is 81.67%. Aim was to verify different knowledge from phonetics and neural network. Also classified the speech signal that are been described by musical theory. The result of research work gone mainstream to the area like prosody modelling and for analysis of disordered speech especially in children [8].

Every emotion contains different vocal parameter that exhibits diverse speech characteristics. An MFCC-based vocal emotional recognition [7] performed using ANN in which MFCC features [16] were used as speech parameters and five different emotional states were considered for analysis. Back-Propagation algorithm applied for interpretation of speaker emotion. Also the proposed system for recognition is independent of linguistic background and achieved 60.55% of average accuracy of recognition.

An effective solution to improve human-computer interaction allowing human and computer intelligent interaction was developed [14]. It says, together with MFCC, pitch is the most frequently used parameter in recognizing speaker's gender. Other speech parameters used are formants, bandwidths, source spectral tilt, jitter and shimmer, harmonic to noise ratio. Speech features used for emotion recognition are statistical analyses of amplitude of speech, energy, pitch, formants, 12 MFCC, pitch and amplitude perturbations. The system does two experiments i.e. gender recognition and emotion recognition. Berlin Emotion Speech database is employed in this research work and support vector machine (SVM) supports as classifier.

Sometimes emotions could not be correctly identified in adverse condition like in a noise corrupted telephone channel speech. A research work investigated a filtering technique in automatic detection of emotions from telephonic speech where the MFCC, delta MFCC and delta-delta MFCC features were incorporated with Gaussian mixture model (GMM) as classifier on Berlin database of emotional speech, while autoregressive (AR) model is employed in the proposed filtering method[6].

3. CLASSIFIER

Various machine learning techniques have been employed for classification purpose. Forward to the procedure of selecting features, classification is needed. Aim is to build a classification model with the help of some machine learning algorithm to predict emotional states on the basis of speech parameters [14].

Among all approaches available, mostly applied classification methods are Hidden Markov Model (HMM), Support Vector Machine (SVM), Maximum Likelihood Model (MLB), Artificial Neural Network(ANN), k-Nearest Neighbour (k-NN). Some other classifiers deserves reference here are Decision Tree, Fuzzy Classifier, LDC, GVQ and many more. These entire classification models have their own advantages & drawbacks according to their application area.

Classifier gives a decision based on the patterns of test speech sample and patterns of trained database. GMM is more suitably applied for global features that has been extracted from training utterances and achieves two levels of accuracy. Classifiers like ANN have their own strength in identifying nonlinear boundaries separating the emotional states as well. Out of many, Feed-forward & Multilayer-perceptron layer neural network are frequently used model. ANN exploit concept of acoustical phonetic & pattern recognition. In speech recognition, HMM is used as classifier for classifying sequential data as it represents a set of various states. Also represents probabilities of making a transition from a state to another. HMM has achieved success in modelling of temporal information in speech spectrums. SVM is a new machine learning algorithm introduced by Vladimir N. Vapnik & derived from statistical learning theory in 90s. Belonging to the family of linear classification, SVM interpreted as an extension of the perceptron. SVM minimize empirical classification error & help in maximization of geometric-margin. Well, main thought behind the SVM model is use of kernel functions like linear, polynomial, radial basis function (RBF) for large extent[3][5].

4. APPLICATION

The use of computers and human-machine interfaces is increasing in our day-to-day life. Also the emerging technologies in smartphones have involved very huge human-machine interfaces. The emerging technologies in speech recognition have made this interface to perform more and more accurately and are leading the world now as one can use these technologies even while driving the vehicle as it provides hands-free operation.

The applications include human-machine interaction where humans can interactively communicate with their personal computers so that the computers will understand the human emotions and correspondingly decide the work to be done and also once the input is given to the computer, it can decide what should be the next decision made. The same thing is applicable to smartphones as the hands-free operation will be available along with emotion recognition.

One of the main application of this system can be applied in schools as according to the emotion of the students, a teacher can decide what subjects can be taught i.e. Intelligent Tutoring System.

5. CONCLUSION

A lot of uncertainties are still present for the best algorithm to classify emotions. Different combinations of emotional features give different emotion detection rate. The researchers are still debating for what features influence the recognition of emotion in speech. This paper reviews major highlights in the recent developments and research of speech emotion recognition with different approaches to provide a technological perspective on society.

ACKNOWLEDGEMENT

The authors would like to thank the "Department of ECE" at Dr. C. V. Raman University" for facilitating the development of the paper, making available resources and also for final deployment.

REFERENCES

[1] Mohit Shrivastava and Anupam Agarwal. Classification of emotions from speech using implicit features. In 9th International Conference on Industrial and Information Systems, 1-6, Dec. 2014.

[2] Kamran Soltani and Raja Noor Ainon. Speech emotion detection based on neural networks. In 9th International Symposium on Signal Processing and its Applications, 1 4244-0779-6/07, IEEE, 2007.

[3] Tin Lay New, Say Wei Foo and Liyanage C. De Silva. Speech emotion recognition using Hidden Markov Models. Speech Communications 41, 603-623, 2003.

[4] S.Lalitha, Abhishek Madhavan, Bharath Bhusan, Shrinivas Saketh. Speech emotion recognition. International Conference on Advances in Electronics, Computers and Communications, 978-1-4799-5496-4/14, IEEE, 2014.

[5] Mohamed R. Amar, Behjat Siddiquie, Collen Richey and Ajay Divakaran. Emotion detection in speech using Deep network. International Conference on Acoustic, Speech and Signal Processing, 978-1-4799-2893-4/14, IEEE, 2014.

[6] Jouni Pohjalainen and Paavo Alku. Multi-scale modulation filtering in automatic detection of emotions in telephone speech. International Conference on Acoustic, Speech and Signal Processing, 980-984, 2014.

[7] Mandar Gilke , Pramod Kachare , Rohit Kothalikar , Varun Pius Rodrigues and Madhavi Pednekar. MFCC-based vocal emotion recognition using ANN, International Conference on Electronics Engineering and Informatics, 150-154, 2012.

[8] Jana Tuckova and Martin Sramka. Emotional speech analysis using Artificial Neural Networks. Proceedings of the International Multiconference on Computer Science and Information Technology, 141-147, 2010.

[9] L. Rabiner and B.-H. Juang, Rabiner & Juang - Fundamentals of Speech Recognition, PTR Prentice Hall, 1993. ISBN 0-13-285826-6.

[10] H. Fletcher, Speech and Hearing in Communication. The Bell Telephone Laboratories Series, D. Van Nostrand Company, Inc.

[11] Mirlab Audio Signal Processing tutorials, "Speech feature MFCC Calculation guide", <http://mirlab.org/jang/books/audiosignalprocessing/speechFeatureMfcc.asp> (Browsing Date: 26th February 2016).

[12] Elizabeth D. Casserly and David B. Pisoni, 'Speech perception and production'. Wiley Interdiscip Rev Cogn Sci. Author manuscript; available in PMC 2013 Aug 12. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3740754/, (Browsing Date: 25th October 2015).

[13] Ian Mcloughlin, Applied Speech and Audio Processing With MATLAB Examples. Cambridge University Press 2009. ISBN-13 978-0-521-51954-0/2.

[14] Igor Bisio, Alessandro Delfino, Fabio Lavagetto, Mario Marchese and Andrea Sciarrone. Gender driven emotion recognition through speech signals for ambient intelligence applications. IEEE Transactions on Emerging Topics in Computing, vol. 1, no. 2, 244-257, December 2013.

[15] V.Vapnik, The Nature of Statistical Learning Theory. New York, NY, USA: Springer-Verlag, 1999.

[16] Sirko Molau, Michael Pitz, Ralf Schl¨uter, and Hermann Ney. Computing mel-frequency cepstral coefficients on the power spectrum. IEEE Transaction, 2011.

[17] B. Yegnanarayana, Artificial Neural Networks. Englewood Cliffs, NJ, USA: Prentice-Hall, 2004.