Review on the classification of emotions in Speech

Gurpreet Kaur#1, Abhilash Sharma#2

#CSE Dept, RIMT Mandi Gobindgarh

1gurpreet140391@gmail.com

2abhilash583@yahoo.com

Abstract— **Automatic speech emotion recognition is a process of recognizing emotions in speech. This has wide applications in the area of phsycatrics help and in robotics.the human computer interaction the challenging area of research. Any effective HCI system has two sections Training and testing. The techniques used in the system are feature extraction and classification.This paper focuses on the survey of the various classification techniques and feature extraction techniques implemented in the speaker emotion recognition.**

Keywords— **Emotion Recognition, Speech, Classification, Features extraction .**

I. INTRODUCTION

Speaker recognition refers to recognize the person from their speech. The speech signal contains the message being spoken, the emotional state of the speaker and the information of the speaker. So the speech signal can be used for the recognition of the speaker and the emotional state of the speaker .Emotion recognition in speech extracts the speech features in each utterance. It is the process of automatically recognizing who is speaking and in which emotional state the words are spoken on the basis of features present in the speech signal. [1].Speaker recoginition can be text independent and text dependent .The content dependent has more acuuracy. Speaker recognition is helpful in the areas such as voice dialling, banking by telephone, telephone shopping, database access services, information services, and security control for confidential information areas voice mail and remote access to computers [2].The detection of emotions in speech is gaining attention in wide range of application likemostly used to develop wide range of application like application for call centre and learning, gaming software,security applications and machine translation. Influence of emotional state of human speech in speaker recognition is very high [3]. The term “emotion” can refer to an extremely complex state associated with a wide variety of mental, physiological and physical events. Emotional speech database is valuable for this speaker recognition.

In a generalized way, a speech emotion recognition system is an application of speech processing in which the patterns ofderived speech features (MFCC, pitch) are mapped by theclassifier (HMM) during the training and testing session using pattern recognition algorithms to detect the emotions from each of their corresponding patterns. The technique is synonymous to speaker recognition system but its different approach to detect emotions makes it intelligent and adds security to achieve better service in various applications [4].

II. BASIC CONCEPTS

A. DATA MINING

The Mining of the data means searching out a piece of data from a bulk data block. This is also called as the knowledge discovery from data (KDD) [6]. The data mining can be categorized in two subsequent ways. First is called classification and the other is called clustering.. Clustering is the methodology of making groups of a set of data objects into multiple groups or clusters so that the objects within the cluster have high similarity, but are very dissimilar to objects in other cluster. Alternatively, it may serve as the pre processing step for other algorithms, such as characterization, attribute subset selection, and classification, which would then operate on the detected clusters and the selected attribute or feature [5]. Classification is always based on two things a) The area which you choose for the classification *i.e.* the cluster region. b) The kind of dataset which you are going to apply on the selected region. To increase the accuracy of the searching technique, any one would need to focus on two things: a) Whether the data set has been clusterzed in proper manner or not. b) If the clusters are defined, whether they fit into the appropriate classified area or not.

B. NOISE REMOVAL

Wh ile recording the speech signal, it consists ofvarious unwanted piece of signals which are considered to be noise/unvoiced part of speech. [7] This noisy part in the speech may be due to low quality of source, loss of speech segments, channel fading etc .A low-pass filter allows signal frequencies below the low cut-off frequency to pass and stops frequencies above the cut-off frequency. If speaker specific information is available in the higher frequencies, a high pass filter is used to remove all the low frequencies less than some threshold values as unvoiced data.

C. FRAMING

Spectral evaluation are reliable in the case of a stationary signal.Imply that the region is brief enough for the behaviour of (periodicity or noise-like appearance) the signal to be close to constant. In sense, the speech region should be short enough so it will reasonably be assumed to be stationary.. Frame period ranges area unit between ten ~ twenty five ms within the case of speech processing. Each speech signal is split into frame segments of size 30ms (milli second) approx . Each segment is extracted at every 50ms interval. This implies that the overlap between segments is 10ms.

D. WINDOWING

The aim of characteristic extraction is to supply spectral characteristics that can help us to build classifiers As the spectrum changes very rapidly. So there is no need to extract spectral features from the entire utterance .Talk is a non-stationary signal, significance that its statistical properties are not constant over time. rather than extract spectral features from a small window of talk . To distinguishes a particular subphone it is assumed) that the signal is stationary. This can be done by utilizing a window which is non-zero interior and zero elsewhere Extract the waveform interior this window by running this window. The windowing procedure can be distinguished by parameters: how wide are the window (in milliseconds), what the offset between successive windows is, & whatthe shape of the window is[8]. Each talk extracted from each windowis called a frame, the number of milliseconds in the border is called the border dimensions & the number of milli-seconds between the left perimeters of successive windows is called the frame move. The extraction of the pointer takes place by multiplying the worth of the pointer at time n, s[n], with the worth of the window at time n, w[n]: y[n]=w[n]s[n].

Rectangular window

$$W[n]=\ 1;0\leq n\leq L\text{-}1$$

$$0\ ;\text{otherwise}$$

Hamming window

$$W[n]=0.54\text{-}0.46\ \cos(2\prod n/L);\ 0\leq n\leq L\text{-}1$$

$$0;\text{otherwise}$$

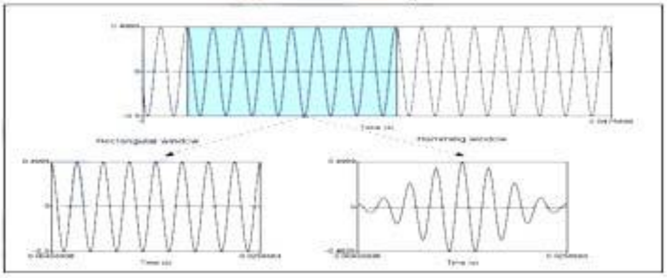

Fig 1:Windowing a portion of a pure sine wave with the rectangular and Hamming Windows[8]

E. FAST FOURIER TRANSFORM (FFT)

The next step after windowing is to extract spectral information for the windowed signal; In this one needs to know how much energy the signal contains at different frequency bands. The tool for extracting spectral information for discrete frequency bands for a discrete-time signal is the Discrete Fourier Transform or DFT[8].A windowed signal x[n]...x[m] is t he input to the DFT and, a complex number X[k] representing the magnitude is the output, for each of N discrete frequency bands.The spectrum can be visualized by plotting the magnitude against the frequency. Hamming windowed portion of a signal and its spectrum as computed by a DFT Fourier analysis in general relies on Euler's formula:

$$\mathbf{e^{ix}=cos(x)+j\ sin(x)}$$

A routinely used algorithm for computing the DFT is the Fast Fourier Transform or FFT [9]. This implementation of DFT is very efficient, but only works for standards of N which are powers of two. The outcomes can also be examined in 2D plots called "Spectrograms".

III. LITERATURE SURVEY

A. Feature Extarction and Matching

Generally the feature extraction techniques are classified as temporal analysis and spectral analysis technique. The speech wave shape is employed for analysis.in temporal analysis In spectral analysis the spectral illustration of speech signal is employed for analysis. Khalid Saeed et al discussed a speech-and-speaker (SAS) identification system based on spoken Arabic digit recognition.[10] The speech signals of the Arabic digits from zero to ten are processed graphically (the signal is treated as an object image for further processing). At the stage of classification, both conventional and neural-networkbased methods are used. The techniques used were FFT, Predictive Coding and PCM. The success rate of the speaker-identifying system obtained in the presented experiments for individually uttered words is excellentand has reached about 98.8% in some cases. The average overall success rate was then 97.45% in recognizing one uttered word and identifying its speaker, and 92.5% in recognizing a three-digit password . Speaker Recognition Using MFCC Front End Analysis and VQ Modeling Technique for Hindi Words using MATLAB was proposed by Nitisha and Ashu Bansal[11]. This paper introduced text dependent systems that have been trained for a particular user.

B. Speaker emotion recognition

Many researches has been done on extorting the different features of the speech and then classify these featuers with different classifiers.The common used features are MFFC.This Section presents the survey of the various classification techniques in combination with the feature extraction techniques.

Bidondo A. et al[12] use Euclidean distance vector Classifier for classification of the speaker in speaker recognition They use the features Running Autocorrelation function and microscopic parameters. Hasan T. et al [13]proposed the method for speaker recognition evaluation .The features extracted were Mean Hilbert Envelope Coefficients (MHEC), PMVDR Front-End, Rectangular Filter- Bank Cepstral Coefficients (RFCC), MFCC-QCNRASTALP classifier used in this was L2-Regularized Linear Regression (L2LR), UBSSVM Anti-Model (UBS-SVM), Score-Averaged PLDA (PLDA-2).the accuracy reate in this system was found to be 50-60%, Md.

Jahangir Alam et al. presented the system for speech and speaker recognition .They extracted the MFCC features of the speech and low-variance multitaper spectrum estimation methods as the classifier.The results were compared with the

Hamming window technique, the sinusoidal weighted cepstrum estimator, multi-peak, and Thomson multitaper techniques provide a relative improvement of 20.25, 18.73, and 12.83 %, respectively, in equal error rate. M. Afzal Hossan et al[14] proposed a Discrete Cosine Transform (DCT-II) based Mel requency Cepstral Coefficients (MFCC) with Fuzzy vector quantization as classifier.The results were compared to GMM. Taufiq Hasan et al. [15]presented a PPCA for acoustic factor analysis with i-vector system. David A. van Leeuwen et al[17] use 19 MFCC's with PLDA classifier. Balaji Vasan Srinivasan et al[16] extract 57 mel-frequency cepstral coefficients (MFCC) features and classify them with Kernel partial least squares (KPLS) used for Discriminative training in the ivector space.they gain 8.4% performance Improvement (relative) in terms of EER. Akshay S. Utane, et al [18]extract 13 MFCC coefficients with delta and acceleration.clasiify it with HMMs and gain 25.1% equal error rate reduction relative to a GMM baseline system. Martinez, J. et al. [19] extract mel-frequency cepstral coefficients (MFCC) featuresand classify them with Vector quantization Technique. They gain 100% of precision with a database of 10 speakers. Joder, Cyril [20] et al. extract mel-frequency cepstral coefficients (MFCC) features and classify it with Nonnegative Matrix Factorization (NMF). P. M. Ghate, et al present Automatic Speaker Recognition System using Mel Frequency Cepstral Coefficients (MFCC) features and Dynamic Time Warping (DTW)algorithm as classifier. Tobias May[22] et al present Automatic Detection of Speaker Attributes Based in Utterance Text using lexical features as well as features inspired by Linguistic Inquiry linear kernel SVM. Garg. Vipul [21]et al proposed Speech based Emotion Recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers.they gain 83% recognition results for closed set. Shashidhar et al.[23] proposed the method for emotion detection using Prosodic features in isolation and in Combination .they use ANNs, GMMs, and SVMs as classifiers. Jae-bok kim et al[24] used Log energy, 12-dimensional MFCCs, pitch, and their first and second derivatives as a feature vector and Modified maximum likelihood linear regression (mllr) as classifier.Prof Sujata et al[25] proposed Linear prediction coefficients (lpc) and Neural network (nn) based technique for emotion recognition and the Accuracy rate was 46%. Priyanka abhang et al[26]proposed a technique using Electroencephalogram (eeg) brain signal and speech, the melfrequency Cepstral coefficients (MFCC) Higher order crossings (hoc), empirical mode decomposition (emd). N. Murali krishna. et al[27]proposed an Emotion recognition using dynamic time warping technique for isolated words .they use MFCC, delta coefficients (δMFCC) and delta delta coefficients (δδMFCC)and Dynamic time warping (dtw), Svm classifier. Recognition rate came out was above 78%. Krishna mohan kudiri et al [28]used Relative amplitude Rbfc approach for segmentation of speech signal and they achieve recognition rate of 71.2% with backpropagation neural networks 72% with svm classifier. Emily mower et al proposed framework for automatic human emotion classification using emotion profiles.they use Mel filterbank cepstral coefficients (MFCCs) features and Ep-svm for classification Overall accuracy rate was 6 8.2%.

IV. Conclusions

In this we presented the introduction to the speaker emotion recognition. We have also introduced the basic concepts that are necessary in the speech signal processing and the literature survey of the past researches on the emotion recognition in speech has also been presented in this paper . thus we conclude our survey of the research papers such that many feature extarcton techniques have been used by the researchers in combination with many classification techniques ,Still there is lots of work to do in the area of human computer interaction systems

References

[1] J. Sirisha Devi , Y. Srinivas and Siva Prasad Nandyala *Automatic Speech Emotion and Speaker Recognition based on Hybrid GMM and FFBNN" in* International Journal on Computational Sciences & Applications (IJCSA) Vol.4, No.1, February 2014.

[2] M. Kockmann, L. Ferrer, L. Burget, E. Shriberg, and J. H. Cernocký, *"Recent progress in prosodic speaker verification,"* in Proc. IEEE ICASSP, (Prague), pp. 4556--4559, May 2011.

[3] Marius Vasile Ghiurcau , Corneliu Rusu,Jaakko Astola,(2011) *"A Study Of The Effect Of Emotional State Upon Text-Independent Speaker Identification",* Published in ICASSP

[4] Rahul.B.Lanjewar, D.S.Chaudhari,(2013) *"Speech Emotion Recognition:A"Review* International Journal of Innovative Technology and Exploring Engineering, ISSN:2278-3075, Vol.2,Issue-4

[5] K.A. Senthildevi and E. Chandra(2007) *"Speech Data Mining & Document Retrieval,*" publication of the IEEE signal processing.

[6] R. Agrawal, M. Mehta, J. Shafer, and R. Srkant,(1996) "*The Quest Data Mining System,"* Proceedings of International Conference on data mining and knowledge Discovery, vol. 96, pp. 244-249

[7] J. Sirisha Devi, Dr. Srinivas Yarramalle, Siva Prasad Nandyala *"Speaker Emotion Recognition Based on peech Features and Classification Techniques"*in I.J. Computer Network and Information Security, pp 61-77

[8] Prasanna, S. R. M., Reddy, B. V. S., & Krishnamoorthy, P. (2009). *"Vowel onset point detection using source, spectral peaks, and*

modulation spectrum energies”. IEEE Transactions on Audio, Speech, and Language Processing, 17, 556–565.

[9] Lawrence Rabiner, Biing-Hwang Juang and B.Yegnanarayana, ”*Fundamental of Speech Recognition* ”, Prentice-Hall, Englewood Cliffs, 2009.

[10] Khalid Saeed and Mohammad Kheir Nammous, *“A Speech-and-Speaker Identification System: Feature Extracton, Description, and Classification of Speech-Signal Image”,* IEEE Transactions on Industrial Electronics Vol. 54, No.2, April 2007, pp. 887-897.

[11] Nitisha and Ashu Bansal, *“Speaker Recognition Using MFCC Front End Analysis and VQ Modelling Technique for Hindi Words using MATLAB”,* Hindu College of Engineering, Haryana, India.

[12] Alejandro Bidondo, Shin-ichi Sato, Ezequiel Kinigsberg, Adrián Saavedra, Andrés Sabater, Agustín Arias, Mariano Arouxet, and Ariel Groisman (2013) *“Speaker recognition analysis using running autocorrelation function parameters”,* POMA - ICA Montreal Volume 19, pp. 060036

[13] Taufiq Hasan, Seyed Omid Sadjadi, Gang Liu, Navid Shokouhi, Hynek Boˇril, John H.L. Hansen,” *crss systems for 2012 nist speaker recognition evaluation”,* ICASSP 2013.

[14] M. Afzal Hossan · Mark A. Gregory*,”Speaker recognition utilizing distributed DCT-II based Mel frequency cepstral coefficients and fuzzy vector quantization”*, Int J Speech Technol (2013), Springer Science+Business Media, LLC 2012.

[15] Taufiq Hasan, John H. L. Hansen,”*Acoustic Factor Analysis for Robust Speaker Verification”,* IEEE Transactions On Audio, Speech, And Language Processing, Vol. 21, No. 4, April 2013.

[16] Balaji Vasan Srinivasan, Yuancheng Luo, Daniel Garcia- Romero, Dmitry N. Zotkin, and Ramani Duraiswami*,”A Symmetric Kernel Partial Least Squares Framework for Speaker Recognition”,* IEEE Transactions On Audio Speech, And Language Processing, Vol. 21, No. 7, July2013

[17] David A. van Leeuwen and Rahim Saeidi*,” Knowing The Non-Target Speakers: The Effect Of The I-Vector Population For Plda Training In Speaker Recognition”,* ICASSP 2013.

[18] Akshay S. Utane, Dr. S. L. Nalbalwar, *“Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine”*, International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013.

[19] Martinez,J. ;Perez, H. ;Escamilla, E. ; Suzuki, M.M.” *Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques”*in Electrical Communications and Computers (CONIELEC OMP), 2012, 22nd International Conference.

[20] Joder, Cyril ;Schuller *"Exploring Nonnegative Matrix Factorization forAudio Classification: Application to Speaker Recognition”*published in SpeechCommunication; 10. ITG Symposium, 26-28 Sept. 2012

[21] Garg, Vipul, Kumar, Harsh; Sinha, Rohit*,”Speech based Emotion Recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers”*in Communications (NCC), 2013 National Conference.

[22] Tobias May, Steven van de Par, and Armin Kohlrausch*,”Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling”,* IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 1, January 2012.

[23] Shashidhar G. Koolagudi · K. Sreenivasa Rao,” *Emotion recognition from speech using source, system,and prosodic features”,* Springer Science+Business Media, LLC 2012.

[24] Jae-Bok Kim, Jeong-Sik Park, Yung Hwan Oh,” *Speaker-characterized Emotion Recognition using Online and Iterative Speaker Adaptation”,* Springer Science+Business Media, LLC 2012

[25] Prof .Sujata Pathak , Prof .Arun Kulkarni ,” *Recognising motions from speech”,* IEEE Transactions On Audio, Speech, And Language Processing , 2011 IEEE.

[26] Priyanka Abhang, Shashibala Rao, Bharti W. Gawali, Pramod Rokade, *“Emotion Recognition using Speech and EEG Signal – A Review”*, International Journal Of Computer Applications (0975 – 8887) Volume 15– No.3, February 2011.

[27] N. Murali Krishna, P.V. Lakshmi, Y. Srinivas J.Sirisha Devi, *“Emotion Recognition using Dynamic Time Warping Technique for Isolated Words”,* IJCSI International Journal Of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011.

[28] Krishna Mohan Kudiri, Gyanendra K Verma and Bakul Gohel*,” Relative amplitude based feature for emotion detection from speech”,* IEEE Transactions On Audio, Speech, And Language Processing , 2010.

[29] Emily Mower, Maja J Mataric, Shrikanth Narayanan,” *A framework for automatic human emotion classification using emotion profiles”,* IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 5, July 2011.

[30] Jagvir Kaur, Abhilash Sharma” *a review of automatic speechemotion recognition”*published in International Journal of Advanced and Innovative Research (2278-7844) / # 308 / Volume 3 Issue 4