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

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 3, Issue 2, February 2014*

# A Study of Methods Involved In Voice Emotion Recognition

**P. Bhardwaj[1], S. Debbarma[2]**

M. Tech Scholar, Computer Science and Engineering, NIT Agartala, Agartala, India [1]

Assistant Professor, Computer Science and Engineering, NIT Agartala, Agartala, India [2]

**Abstract:** The analysis of speech has been a very interesting research area. Now the physiologists have dived in this area to take the research to the new level such that the voice can tell many important physical, mental and physiological aspects of human being. Though there are many individual efforts to recognize different aspect of human being, but there is a need to make a combined effect to get better results taking into considerations of age, gender etc.

**Keywords**: MFCC; emotion detection; big5

## I. INTRODUCTION

Human voices can convey a considerable amount of information to a listener and can be used to assess a number of qualities about a person. For instance, several studies have shown that listeners can determine a variety of physical attributes of speakers such as gender, race, height, weight, and other body dimensions by simply hearing their voice. Speech is a signal that can provide information about the permanent personal traits as well as temporal emotional traits of the speaker, through the careful processing of acoustics. In day-to-day life we perceive hundreds of person's voice and act accordingly. It is very easy to understand the emotions of our known ones because we are accustomed to the habits and activities of them, but when we interact with a stranger, our mind reads their voice and predict their emotion by matching the acoustic patterns of voice with previously encountered voice patterns. Similarly if a robot needs to interact with the humans, they should be able to read the emotions of people interacting with them. Section II throws light on the various features of voice that are used for voice processing techniques.

Section III describes some classification methods used for speech emotion recognition. Section IV presents some developments in emotion detection area.

## II. FEATURE OF SPEECH AND EXTRACTION

Feature extraction is the process of calculating the speech signal features which are relevant for speech processing. Since the computer has no sense of hearing and perception like humans, they have to be fed with these features of speech which become a determining factor after classification. Feature extraction involves analysis of speech signal. The researchers have used various features such as pitch, loudness, MFCC, LPC etc for extracting emotion. The number of features range from 39 extracted from mfcc to few hundreds including formants, maximum, minimum, standard deviation and so on for improving the correctness of results. The feature extraction techniques are classified as temporal analysis and spectral analysis technique. In temporal analysis, the speech waveform itself is used for analysis. In spectral analysis, the spectral representation of speech signal is used for analysis. Features are primary indicator of speaker's emotional state. A lot of features are extracted from feature extraction process like Mel Frequency Cepstral Coefficient (MFCC), pitch, PLP, RASTA-PLP, loudness etc. Feature extraction process can be divided into two steps: spectral feature extraction and prosodic feature extraction.

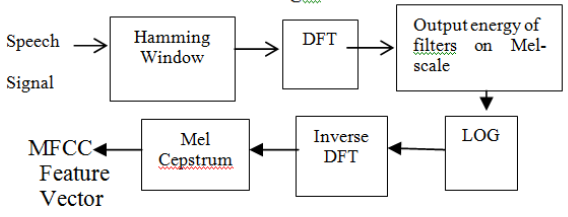
*a. Spectral Feature Extraction*

*1. MFCC*

The MFCC [1] is the most relevant example of a feature set that is extensively used in voice processing. Speech is usually segmented in frames of 20 to 30 ms, and the window analysis is shifted by 10 ms. Each frame is transformed to 12 MFCCs plus a normalized energy parameter. The first and second derivatives (D's and DD's) of MFCCs and energy are estimated, resulting in 39 numbers that represent each frame. Assuming sample rate of 8 kHz, for each 10 ms the feature extraction module delivers 39 numbers to the modelling stage. This operation with overlap among frames is equivalent to taking 80 speech samples without overlap and representing them by 39 numbers. In fact, assuming that each speech sample is represented by one byte and each feature is represented by four bytes (float number), one can see that the parametric representation increases the number of bytes to represent 80 bytes of speech (to 136 bytes). If a sample rate of 16 kHz is assumed, the 39 parameters would represent 160 samples. For higher sample rates, it is intuitive that 39 parameters do not allow reconstructing the speech samples back. Anyway, one should notice that goal here is not speech compression but using features suitable for speech recognition.

The following figure shows steps involved in MFCC feature extraction.

Fig:1




MFCCs are the most widely used spectral representation of speech in many applications, including speech emotion

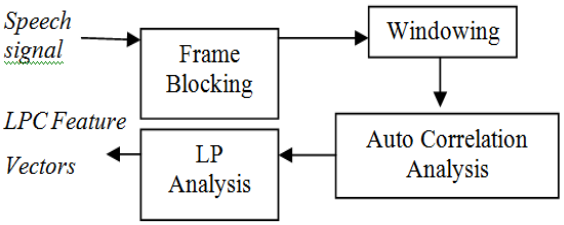
ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

recognition because statistics relating to MFCCs also carry emotional information.

*2. LPC*

It is one of the powerful signal analysis techniques is the method of linear prediction. Linear predictive coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using information of a linear predictive model [2]. It provides an accurate estimate of the speech parameters and it is also an efficient computational model of speech. The idea behind LPC is that a speech sample can be approximated as a linear combination of past speech samples. Through minimizing the amount of squared differences (over a finite interval) between the actual speech samples and predicted values, a unique set of parameters, the predictor coefficients can be determined. These coefficients form the basis of LPC of speech [3]. The analysis provides the capability for computing the linear prediction model of speech over time. Predictor coefficients are therefore transformed to a robust set of parameters known as cepstral coefficients. Figure 2 shows the steps involved in LPC feature extraction.



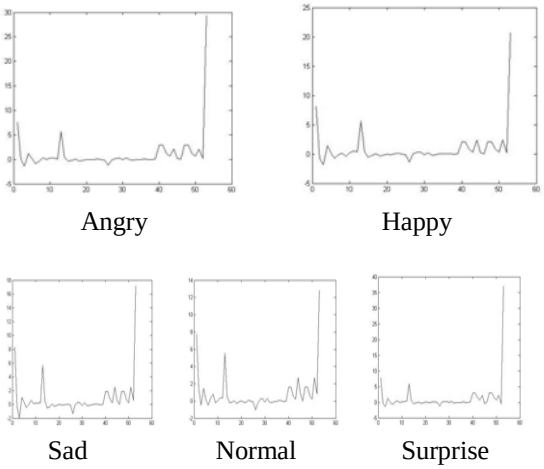

*b. Prosodic feature extraction*

*1. Pitch*

Statistics related to pitch [13] conveys considerable information about emotional status. For this project, pitch is extracted from the speech waveform using a modified version of the RAPT algorithm for pitch tracking implemented in the VOICEBOX toolbox. Using a frame length of 50ms, the pitch for each frame was calculated and placed in a vector to correspond to that frame. The various statistical features are extracted from the pitch tracked from the samples. We use minimum value, maximum value, range and the moments- mean, variance, skewness and kurtosis. We hence get a 7 dimensional feature vector which is appended to the end of the 39 dimensional supervector obtained from the GMM.

*2. Loudness*

Loudness [14] is extracted from the samples using DIN45631 implementation of loudness model in MATLAB. The function loudness() returns loudness for each frame length of 50ms and also one single specific loudness value. Now the same minimum value, maximum value, range and the moments- mean, variance, skewness and kurtosis statistical features are used to model the loudness vector. Hence we get an 8 dimensional feature vector which is appended to the already obtained 46 dimensional feature vector to obtain the final 54 dimensional feature vector. This vector can now be given as input to the SVM.




*3. Formant*

Formants are the distinguishing or meaningful frequency components of human speech and of singing. By definition, the information that a human requires to distinguish between vowels can be represented purely quantitatively by the frequency content of the vowel sounds. In speech, these are characteristic partials that identify vowels to the listener. The formant with lowest frequency is called f1, the second lowest called f2, and the third f3. Most often the first two formants, f1 and f2, are enough to disambiguate a vowel. These two formants determine quality of vowels in terms of the open/close and front/back dimensions (which have traditionally, though not accurately, been associated with position of the tongue). Thus first formant f1 has a higher frequency for an open vowel (such as [a]) and a lower frequency for a close vowel (such as [i] or [u]); and the second formant f2 has a higher frequency for a front vowel (such as [i]) and a lower frequency for a back vowel (such as [u]).[15][16] Vowels will almost always have four or more distinguishable formants; sometimes there are more than six. However, the first two formants are the most important in determining vowel quality, and this is displayed in terms of a plot of the first formant against the second formant,[17] though this is not sufficient to capture some aspects of vowel quality, such as rounding.[18]

Nasals usually have an additional formant around 2500 Hz. The liquid [l] usually has an extra formant at 1500 Hz, while the English "r" sound ([ɹ]) is distinguished by virtue of a very low third formant (well below 2000 Hz).

Plosives (and, to some degree, fricatives) modify the placement of formants in the surrounding vowels. Bilabial sounds (such as /b/ and /p/ in "ball" or "sap") cause a lowering of the formants; velar sounds (/k/ and /g/ in English) almost always show f2 and f3 coming together in a 'velar pinch' before the velar and separating from the same 'pinch' as the velar is released; alveolar sounds (English /t/ and /d/) cause less systematic changes in neighboring vowel formants, depending partially on exactly which vowel is present. The time-course of the changes in vowel formant frequencies are referred to as 'formant transitions'.

If the fundamental frequency of the underlying vibration is higher than a resonance frequency of system, then the

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

formant usually imparted by that resonance will be mostly lost. This is most apparent in example of soprano opera singers, who sing high enough that their vowels seem to be very hard to distinguish.

## III. CLASSIFICATION

*1. Gaussian Mixture Model*

A Gaussian Mixture Model (GMM) [4] is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as parametric model of the probability distribution of continuous measurements or features in a biometric system, such as the vocal-tract related spectral features in a speaker recognition system. GMM parameters are estimated from the training data using the iterative Expectation- Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior model. The density function of a GMM is defined as:

$$p(x) = \sum_{i=1}^{n} w_i N(x; \mu_i; \sum_i)$$

For each emotional utterance, a GMM is trained with the extracted spectral features, and the corresponding super vector is obtained. Instead of training the GMM via EM algorithm, we adapt the GMM from a universal background model which is widely used in speaker recognition. In this paper, the UBM is a GMM trained via EM algorithm using the training data of the experiments. The adaptation of each emotional utterance's GMM is performed with maximum a posteriori (MAP) algorithm. The GMM super vector is formed by concatenating the normalized means of the Gaussian component.

The GMM super vector can be considered as a mapping from the spectral features to a high-dimensional feature vector, which has a fixed dimension for all the emotional utterances. Therefore, we the GMM super vectors can be used as input for SVM training. Thus we have a 39 dimensional feature vector as a 1x39 matrix.

*2. Dynamic Time Warping (DTW)*

Dynamic Time Warping is an algorithm for measuring similarity between two sequences which may vary in time or speed [5]. A well-known application has been ASR, to cope with the different speaking speeds. In general, it is a method which allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions, i.e. the sequences are "warped" non-linearly to match each other. The sequence alignment method is often used in the context of HMM. In general, DTW allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions. This technique is quite efficient for the isolated word recognition and can be modified to recognize connected word also [5].

*3. Hidden Markov Model*

In this approach, variations in speech are modeled statistically (e.g., HMM), using automatic learning procedures. This approach represents the current state of art. Modern general-purpose speech recognition systems are based on statistical acoustic and language models. Effective acoustic and language model for ASR in unrestricted domain require large amount of acoustic and linguistic data for parameter estimation.

Processing of large amount of training data is a key element in the development of an effective ASR technology nowadays. The main disadvantage of statistical models is that they must make a priori modeling assumptions, which are likely to be inaccurate, handicapping the system's performance. The reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use [6] [7]. HMMs to represent complete words can be easily constructed (using the pronunciation dictionary) from phone HMMs and word sequence probabilities added and complete network searched for best path corresponding to the optimal word sequence. HMMs are simple networks that are capable to generate speech (sequences of cepstral vectors) using a number of states for each model and modeling the short-term spectra associated with each state, usually, mixtures of multivariate Gaussian distributions (the state output distributions). The parameters of model are the state transition probabilities and means, variances and mixture weights that characterize the state output distributions. Each word, each phoneme, will have a different output distribution; a HMM for a sequence of words or phonemes is made by concatenating the individual trained HMM [9] for the separate words and phonemes. The Current HMM-based large vocabulary speech recognition systems are often trained on hundreds of hours of acoustic data.

The word sequence and a pronunciation dictionary and the HMM [8] [9] training process can automatically determine word and phone boundary information during training. This means that it is quite straightforward to use large training corpora. It is the major advantage of HMM which extremely reduces the time and complexity of recognition process for training large vocabulary.

*4. Neural Network-Based Approaches*

Another approach in acoustic modeling is the use of neural networks. They are proficient of solving much more complicated recognition tasks, but do not scale as excellent as the Hidden Markov Model (HMM) when it comes to large vocabularies. Rather than being used in the general-purpose speech recognition applications they can handle low quality, noisy data and speaker independence [10] [11]. Such systems can achieve greater accuracy than HMM based systems, as long as the training data is there and the vocabulary is limited. A more general approach using the neural networks is phoneme recognition. This is an open field of research, but the results are better than HMMs [10] [12]. There are also NN-HMM hybrid systems that use the neural network part for phoneme recognition and the HMM part for language modeling.

## IV. RELATED WORK

A lot of work has been done to study the speech properties and their relationship with personal trait of speaker.

### A. *Eagerness or Intensity of Situation*

Elaleem et. Al. conducted a research based on fuzzy inference system that can identify the mood of the person leaving a message on the answer machine and set a priority of the message. The paper sets the priority on the basis of several basic emotions from human speech including (very serious person (something important is happening like accident ...), regular person, persons in hurry, happy person and sad person). These emotions are classified into three categories according to priority of voice [19].

### B. *Personality of a Person*

Mohammadi et. Al. proposed prediction of trait attribution based on prosodic features. A large number of models (see [20] for an extensive survey)have been proposed to extract the personality information of speaker through voice investigation, but the most common personality representation relies on the Big Five (BF), five broad dimensions that "appear to provide a set of highly replicable dimensions that parsimoniously and comprehensively describe most phenotypic individual differences" [21].

The BF has been identified by applying factor analysis to the large number of words describing personality in everyday language (around 18,000 in English [20]). Despite the wide variety of terms, personality descriptors tend to group into five major clusters corresponding to the BF:

- Extroversion: Active, Assertive, Energetic, etc.
- Agreeableness: Appreciative, Kind, Generous, etc.
- Conscientiousness: Efficient, Organized, Planful, etc.
- Neuroticism: Anxious, Self-pitying, Tense, etc.
- Openness: Artistic, Curious, Imaginative, etc.

### C. *Gender*

The gender of a person can also be determined by using the pitch variation features. Kumar et.al. [21] used acoustic measures from both the voice source and the vocal tract, the fundamental frequency (F0) or pitch and the first formant frequency (F1) respectively. It is well-known fact that F0 values for male speakers are lower due to longer and thicker vocal folds. F0 for adult males' voice is typically around 120 Hz, whereas F0 for adult females is around 200 Hz. Further voice of adult males exhibit lower formant frequencies than adult females due to vocal tract length differences.

Linear predictive analysis is used to find both the fundamental frequency and the formant frequency of each frame of speech. The mean of all the frames is calculated to obtain the values for each speaker. The Euclidean distance of the mean point is found from the preset means of the male class and the female class. The least value of the two distances determines whether the speaker is male or female.

It was also observed that by increasing the unvoiced part in the speech, like the sound of 's', the value of pitch increases hampering the gender detection in case of Male samples. Likewise by increasing the voiced speech, like the sound of 'a', decreases the value of pitch but the system takes care of such dip in value and results were not affected by the same.

### D. *Age*

In the speech based communications, age is an important factor for everyone, especially in the first meeting, to adapt him or her in appropriate treatment [23]. By extraction these features of every speaker, we can adapt our speaking style to the person whom we are talking to. It is a common application in every day telephone communications for people. Also, some companies need an automatic age estimation system to play adaptable queue music for different age groups of their customers. There are some approaches in ASR to identify dialogues with angry or unsatisfied speakers, but not many researches which use information about speaker characteristics, like age and gender, and behavioral characteristics while there are a lot of useful applications associated with this topic. The considerable examples of such applications are the adaptation of the waiting queue music, the offer of age dependent advertisements to callers in the waiting queue, the statistical information on the age distribution of a caller group, or changing the speaking habits of the text-to speech module of the ASR system [24].The age of a person can also be estimated by using the voice extraction techniques [23, 24]

## V. PROPOSED FUTURE WORK AND SCOPE

There is a lot of work on emotional intelligence, and there are also separate work on extracting other information like age, gender etc. But it has been proved that the voice features keep on changing by age. Similarly for different genders the emotion matching parameters should be different. It can be felt easily that when we hear a sound, first thing comes in our mind whether the speaker is boy or a girl, then we estimate the age of person, then we guess the meaning and emotion flowing through the voice. There are different physiological aspects related to the both gender and similar is the case with the age of person. So the machine needs to be trained to differentiate between the gender as well as the age groups. If a lady shouts, it shows anger of fear, but this the same perception cannot be applied to the shouting baby. There is a lot of scope of using all the works combined to increase the accurateness of the emotion detection by the machine.

The present work is confined to few applications that are specific in nature. But if we use the combination and develop an algorithm for emotion detection, it would find use a broad use in automation industries for rescue operations, household articles, enquiry and support systems etc.

## REFERENCES

[1] Sahidullah, Md.; Saha, Goutam (May 2012). "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition". Speech Communication 54 (4): 543–565. doi:10.1016/j.specom.2011.11.004.

[2] Deng, Li; Douglas O'Shaughnessy (2003). Speech processing: a dynamic and optimization-oriented approach. Marcel Dekker. pp. 41–48. ISBN 0-8247-4040-8.
[3] N.Uma Maheswari, A.P.Kabilan, R.Venkatesh, "A Hybrid model of Neural Network Approach for Speaker independent Word Recognition", International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010 1793-8201.
[4] Bishop, Christopher (2006). Pattern recognition and machine learning. New York: Springer. ISBN 978-0-387-31073-2.
[5] Santosh K.Gaikwad, Bharti W.Gawali and Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.
[6] M. Chandrasekar, M. Ponnavaikko, "Tamil speech recognition: a complete model", Electronic Journal «Technical Acoustics» 2008, 20.
[7] Ghulam Muhammad, Yousef A. Alotaibi, and Mohammad Nurul Huda , "Automatic Speech Recognition for Bangia Digits", Proceedings of 2009 12th International Conference on Computer and Information Technology (ICCIT2009) 21-23 December, 2009, Dhaka, Bangladesh, 978-1-4244-6284-1/09/$26.00 ©2009 IEEE.
[8] A.P.Henry Charles & G.Devaraj, "Alaigal-A Tamil Speech Recognition", Tamil Internet 2004, Singapore.
[9] Zhao Lishuang , Han Zhiyan, "Speech Recognition System Based on Integrating feature and HMM", 2010 International Conference on Measuring Technology and Mechatronics Automation, 978-0-7695-3962-1/10 $26.00 © 2010 IEEE.
[10] Meysam Mohamad pour, Fardad Farokhi, "An Advanced Method for Speech Recognition", World Academy of Science, Engineering and Technology 49 2009.
[11] Vimal Krishnan V. R, Athulya Jayakumar and Babu Anto.P, "Speech Recognition of Isolated Malayalam Words Using Wavelet Features and Artificial Neural Network", 4th IEEE International Symposium on Electronic Design, Test & Applications, 0-7695-3110-5/08 $25.00 © 2008 IEEE
[12] Raji Sukumar.A, Firoz Shah.A and Babu Anto.P, "Isolated question words recognition from speech queries by Using artificial neural networks", 2010 Second International conference on Computing, Communication and Networking Technologies, 978-1-4244-6589-7/10/$26.00 ©2010 IEEE.
[13] Anssi Klapuri and Manuel Davy (2006). Signal processing methods for music transcription. Springer. p. 8. ISBN 978-0-387-30667-4.
[14] Olson, Harry F. (February 1972). "The Measurement of Loudness". Audio: 18–22.
[15] Ladefoged, Peter (2006) A Course in Phonetics (Fifth Edition), Boston, MA: Thomson Wadsworth, p. 188. ISBN 1-4130-2079-8
[16] Ladefoged, Peter (2001) Vowels and Consonants: An Introduction to the Sounds of Language, Maldern, MA: Blackwell, p. 40. ISBN 0-631-21412-7
[17] Deterding, David (1997) 'The Formants of Monophthong Vowels in Standard Southern British English Pronunciation', Journal of the International Phonetic Association, 27, pp. 47-55.
[18] Hayward, Katrina (2000) Experimental Phonetics, Harlow, UK: Pearson, p. 149. ISBN 0-582-29137-2
[19] Onsy A. Elaleem , Hassan M. Elragal and Heba M. Shehata, "voice message priorities using Fuzzy mood identifier" , The 23rd National Radio Science Conference,2006, pp- 1-6.
[20] G. Saucier, L. Goldberg, "The Language of Personality: Lexical Perspectives on the Five-Factor Model," The Five-Factor Model of Personality, J. Wiggins, ed., Guilford Press, 1996.
[21] P. Jain, "Automatic Human Gender Identification System", M.Tech. Thesis, IIT Kanpur, 2008
[22] K. Rakesh, S. Dutta, K. Shama," IJAET,1,51,(2011). M. Nishimoto, et al., "Subjective age estimation using speech sounds: Comparison with facial images," in Systems, Man and Cybernetics. SMC 2008. IEEE International Conference on, 2008, pp. 1900-1904.
[23] T. Bocklet, et al., "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines," in Acoustics, Speech and Signal Processing. ICASSP 2008. IEEE International Conference on, Las Vegas, NV, 2008, pp. 1605-1608.
[24] S.M. Hughes, B.C.Rhodes, Proc. of 4th Ann. Meet. NE Evol. Psy. Soc.,4,290,(2010).

## BIOGRAPHIES

**Swapan Debbarma** completed his B.Tech. degree in Computer Science & Engineering from NERIST, AP and M.Tech.from TU, Tripura-Agartala, India, in the same branch. Mr. Debbarma is working in NIT, Agartala as an Asst. Prof in CSE Department since 2001 and pursuing his Ph.D from NIT, Agartala. His field of interests are Nanotechnology, Artificial Intelligence and Networking etc.

**Prashant Bhardwaj** obtained B.Tech degree from MIT, Moradabad, UP. After few years of successful teaching career, he is currently pursuing M.Tech in Computer Science and Engineering from NIT, Agartala, India.