ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 3, Issue 1, pp: (82-89), Month: April 2016 – September 2016, Available at: www.paperpublications.org

Voice Emotion Recognition

Ankita Kirti, Nishant Anand

Abstract: **Speech technology and systems in human computer interaction have witnessed a stable and remarkable advancement over the last two decades. Today, speech technologies are commercially available for an unlimited but interesting range of tasks. These technologies enable machines to respond correctly and reliably to human voices, and provide useful and valuable services. This thesis presents the characteristics of emotion in voice and on that basis propose a new method to detecting emotion in a simplified way by using a prosodic features and spectral from speech. We classify seven emotions: happy, anger, fear, disgust, sadness and neutral inner state. This thesis discusses the method to extract features from a recorded speech sample, and using those features, to detect the emotion of the subject. Every emotion comprises different vocal parameters exhibiting diverse characteristics of speech, which is used for preliminary classification. Then Mel-Frequency Cepstrum Coefficient (MFCC) method was used to extract spectral features. The MFCC coefficients were again trained by Artificial Neural Network (ANN) which then classifies the input in particular emotional class.**

Keywords: **Speech technology, MFCC, Artificial Neural Network.**

1. INTRODUCTION

Automatic Emotion Recognition is a recent research topic which is primarily formulated for the Human Computer Interaction (HCI) field. As computers have become an integral part of our lives, the need has risen for more natural communication interface between human beings. To make HCI more natural, it would be favourable if modelled systems have the ability to recognize emotional situations the same way as humans do. 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.

1.1 Literature Survey

Recent research concentrates on developing systems that would be much more robust against variability in environment, speaker and language.

This thesis discusses approaches to recognize the emotional user state by analyzing spoken utterances on both, the semantic and the signal level. We classify seven emotions: happy, anger, surprise, fear, disgust, sadness and neutral inner state.

Human Machine Interface (HMI) recognition systems incorporate the principles of corporal interaction that deduce perfunctory characteristic extraction methods. The speech characteristics include pitch, formant, prosody and timbre. The emotion verification task designed for such recognition systems uses a-priori information to determine whether the outcome of a speech sample is efficiently construed in a manner in which the sentence is spoken. In practice, a-priori information would normally be available in a real system, instinctively captured when candidate users are registered with that system. Within such constraints, there are two further main branches to this research area; one in which the material being spoken is fixed and the other in which the material being spoken is unrestricted. In the unrestricted case the problem is more difficult, and accuracy may be more closely related to the amount of captured data that can be analysed than upon the accuracy of the system employed[1].

The first book on expression of emotions in animals and humans was written by Charles Darwin in the nineteenth century [1]. After this milestone work psychologists have gradually

accumulated knowledge in the field. A new wave of interest has recently risen attracting both psychologists and artificial intelligence specialists. There are several reasons for this renaissance such as: technological progress in recording, storing, and processing audio and visual information; the development of non-intrusive sensors; the advent of wearable computers; the urge to enrich human-computer interface from point-and-click to sense-and-feel; and the invasion on our computers of lifelike software agents and in our homes robotic animal-like devices like Tiger's Furbies and Sony's Aibo who supposed to be able express, have and understand emotions. A new field of research in AI known as affective computing has recently been identified [2]. As to research on decoding and portraying emotions in speech, on one hand, psychologists have done many experiments and suggested theories (reviews of about 60 years of research can be found in [3,4]). On the other hand, AI researchers made contributions in the following areas: emotional speech synthesis [5], recognition of emotions [6], and using agents for decoding and expressing emotions [7]. The motivation for our research is to explore the ways how recognition of emotions in speech could be used for business, in particular, in a call center environment. One potential application is the detection of the emotional state in telephone conversations, and providing a feedback to an operator or a supervisor for monitoring purposes. Another application is sorting voice mail messages according to the emotions expressed by the caller. One more challenging problem is to use emotional content of the conversation for the operator performance evaluation.

In the computer speech community, much attention has been given to "what was said" and "who said it", and the associated tasks of speech recognition and speaker identification, whereas "how it was said" has received relatively little.

Previous research on emotions both in psychology and speech tell us that we can find information associated with emotions from a combination of prosodic, tonal and spectral information; speaking rate and stress distribution also provide some clues about emotions [2]

1.2 Motivation

Speech is one of the most natural communication forms between human beings. Humans also express their emotion via written and spoken language. Enabling systems to interpret user utterances for a more intuitive human machine interaction therefore suggests also understanding transmitted emotional aspects. The actual user emotion may help system track the user's behaviour by adapting to his inner mental state. Generally recognition of emotions is in the scope of research in the human-machine-interaction. Among other modalities like mimic speech is one of the most promising and established modalities for the recognition [1][2][3]. There are several emotional hints carried within the speech signal. Nowadays attempts in detecting emotional speech analyze in general signal characteristics like pitch, energy, duration or spectral distortions [4]. However, on semantically higher levels emotional clues can also be found. In literature one can even rely almost only on such semantic hints besides spare graphical attempts to capture prosodic elements like in bold or italic characters typed phrases. Therefore we aim to also spot emotional keyphrases, analyze the dialogue history and the degree of verbosity in the communication between man and machine. This is realized through a parallel analysis of spoken utterances in view of general system announcements, command interpretation and detection of emotional aspects. However, the semantic means introduced could as well be used for analysis of nonspoken language.

Emotions are fundamental for humans, impacting perception and everyday activities such as communication, learning and decision-making. They are expressed through speech, facial expressions, gestures and other non-verbal clues. Speech emotion detection refers to analysing vocal behaviour as a marker of affect, with focus on the nonverbal aspects of speech. Its basic assumption is that there is a set of objectively measurable parameters in voice the affective state a person is currently expressing. This assumption is supported by the fact that most affective states involve physiological reactions which in turn modify the process by which voice is produced. For example, anger often produces changes in respiration and increases muscle tension, inuencing the vibration of the vocal folds and vocal tract shape and affecting the acoustic characteristics of the speech [25]. So far, vocal emotion expression has received less attention than the facial equivalent, mirroring the relative emphasis by pioneers such as Charles Darwin.

In the past, emotions were considered to be hard to measure and were consequently not studied by computer scientists. Although the field has recently received an increase in contributions, it remains a new area of study with a number of potential applications. These include emotional hearing aids for people with autism; detection of an angry caller at an

automated call centre to transfer to a human; or presentation style adjustment of a computerised e-learning tutor if the student is bored. A new application of emotion detection proposed in this dissertation is speech tutoring. Especially in persuasive communication, special attention is required to what non-verbal clues the speaker conveys. Untrained speakers often come across as bland, lifeless and colourless. Precisely measuring and analysing the voice is a difficult task and has in the past been entirely subjective. By using a similar approach as for detecting emotions, this report shows that such judgements can be made objective.

1.3 Challenges

This section describes some of the expected challenges in implementing a realtime speech emotion detector. Firstly, discovering which features are indicative of emotion classes is a difficult task. The key challenge, in emotion detection and in pattern recognition in general, is to maximise the between-class variability whilst minimising the within class variability so that classes are well separated. However, features indicating different emotional states may be overlapping, and there may be multiple ways of expressing the same emotional state. One strategy is to compute as many features as possible. Optimisation algorithms can then be applied to select the features contributing most to the discrimination while ignoring others, creating a compact emotion code that can be used for classiffication. This avoids making difficult a priori assumptions about which features may be relevant. Secondly, previous studies indicate that several emotions can occur simultaneously [14]. For example, co-occurring emotions could include being happy at the same time as being tired, or feeling touched, surprised and excited when hearing good news. This requires a classiffier that can infer multiple temporally co-occurring emotions. Thirdly, real-time classiffication will require choosing and implementing efficient algorithms and data structures. Despite there existing some working systems, implementations are still seen as challenging and are generally expected to be imperfect and imprecise.

2. EMOTIONS AND THEIR ACOUSTIC FEATURES

There has not been any considerable published matter about the properties of the emotional states of speech orated. The different emotions are characterised by specific properties which vary from person to person, but on an average the properties of every emotion can be distinguished from other.

A short description is being given below which is dominant in normal emotional voices:

Anger: Anger is characterised by many high tones in speech, fast rate of speaking with little and negligible pauses.

Happy: Lot much air is also expelled out in happiness as we tend to laugh. The tones are high but not similar to angry speech, resulting in slightly lower mean values.

Fear: The voice becomes shaky with low rate of speech because of inability to think quick, but there is a good combination of quick high and low tone changes.

Sad: Sad voice has the lowest values for all the properties. The rate becomes slow, speaker uses low tones unable to hear and we find seldom use of high tones;

Surprise: The sentences start with high tone out of excitement, but decreases along the length of sentence. The voice possess similar rate of speech as the normal voice.

Disgust: It has much similarity with sad voice, but is has got some uses of high tones at instances to express the disgust when talking about a specific subject. It specially varies when cause is being discussed.

Normal: It enjoys a moderate rate of speech, easily understandable by all. The words are pronounces in a monotonous tone, because of lack of emotions.

A comparative study is shown below on the very basic properties of speech:

Table 1.Comparision of Prosodic features of Emotions

	Anger	Happy	Fear	Sad	Surprise	Disgust	Normal
Mean	More	More	More	Less	More	Less	Less
Std. Deviation	Less	Mid	More	Less	Less	Mid	Less
Rate of Speech	More	More	Less	Less	Mid	Mid	Mid

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 3, Issue 1, pp: (82-89), Month: April 2016 – September 2016, Available at: www.paperpublications.org

These characterstics are present in normal portrayal of emotion, provided any problem in speech of individuals. The challenging problem is mimicked speech which is one of the most promising for the recognition [6].

3. FEATURES EXTRACTION OF SPEECH

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.

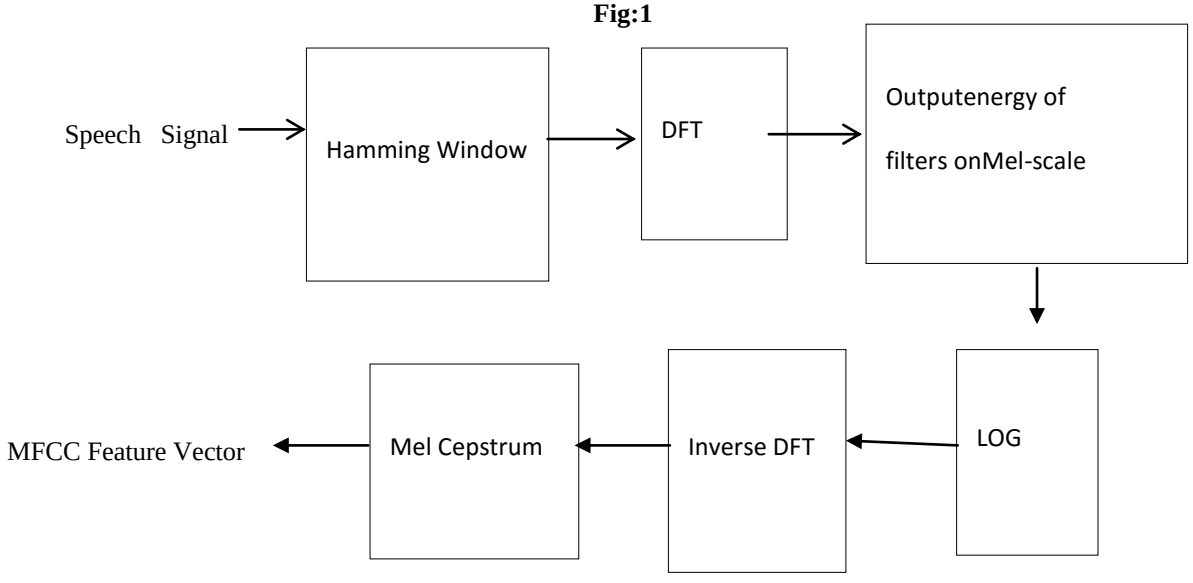
A. Spectral Feature Extraction

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



ISSN 2350-1022

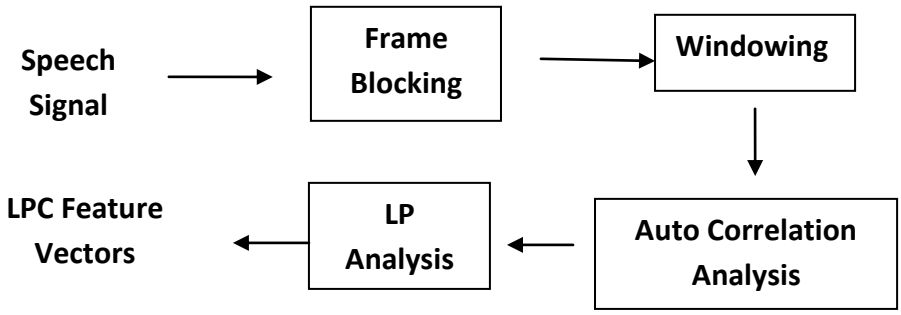
International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 3, Issue 1, pp: (82-89), Month: April 2016 – September 2016, Available at: www.paperpublications.org

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

FIG 2



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 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.

4. IMPLEMENTATION AND RESULTS

The emotions were acted, what surely is a disadvantage since users tend to exaggerate when acting. In an initial phase user statements were not recorded to make the pro-bands familiar with simulating emotions naturally. For the classification of prosodic parameters the system was in advance adapted by training with ten samples for each emotion. However, these results can be seen as upper limit for achievable results.

The confusion between anger fear and surprise is high in comparison with any other pair of emotion. This is due to the fact that the features such as pitch acoustic features of these two emotions are considerably different to other emotions. The back-propagation algorithm proves to be an efficient method for emotion recognition with reference to the graphical result. The detection of anger, fear and surprise is above 80%. The normal, sad and happy voice detection is fairly good. The detection of disgust shows the lowest results.

Table II: Confusion-Matrix obtained for the detection of emotion

Emotional Class	Sad	Anger	Fear	Surprise	Disgust	Happy	Normal
Sad	70.11%	10.48%	6.02%	1.50%	7.50%	0.5%	3.99%
Anger	4.5%	86.55%	2.05%	0%	5.03%	1.47%	0.4%
Fear	5.98%	2.02%	83.67%	0.52%	1.48%	0.00%	6.33%
Surprise	1%	1.67%	1.33%	85.87%	0.23%	7.77%	2.13%
Disgust	26.72%	8.28%	1.44%	0.00%	60.98%	0.56%	2.02%
Happy	0.00%	0.00%	2.06%	4.96%	3.2%	78.34%	11.46%
Normal	5.94%	2.44%	1.5%	1.06%	1.05%	10.89%	77.12%

It could be shown that understanding emotional phrases seems a very promising way. However the combination with prosodic parameters is useful to capture non-verbal expressions. Further semantic features could not be used to satisfyingly detect all accosted emotions, but they also supported robust recognition in the fusion. Finally the fusion was able to resolve ironic phrases by the signal characteristics. Generally the recognition proved rather speaker dependent, but conditioning the system to a new user keeps the system applicable. The concept of integration of models allows the connection of further multimodal input data as general human expressional characteristics like mimic recognition or domain specific data like driving data in a car. The results highly motivate further investigation in this area.

The recognition system strictly adheres to the computed results of the database. The recognition of disgust remains the most demanding problem. Rest emotions are recognised on an average at a rate of 75%.

The future of emotion recognition lies with solving the three low tone emotions that are the signs of Depression and Crime in the society. The major health problems are due to the lack of emotional stability of persons that lead to suicide, cardiac risks, and neurological problems and in a whole disruption in the growth of society and mankind.

5. FUTURE SCOPE

In the interpersonal communication partners adapt in their acoustic parameters to show sympathy for each other. A technical system enabled to talk by speech synthesis therefore needs to know the actual user emotion and the according acoustic parameters to adapt instead of staying neutral all the time. Furthermore the communication channels of a speaker interact with each other. The knowledge of the implicit channel is needed to interpret the explicit channel. Irony might be a good example to demonstrate that prosodic features help understand the explicitly uttered intention. An emotion recognition system might also be called in for an objective judgment in psychiatric studies [5]. Finally there is certainly a funfactor in automatic reaction to user emotions in many applications like video games.

In a first approach we used a two-dimensional emotion sphere defined by the axes activeness and positveness [7]. In this plain different areas could be assigned to emotional states. For example a very active and positive user is meant to be joyful, while an as well passive as negative user is associated with sadness. Other approaches introduce even a third dimension [8] with an axis of control level. The basing measurement of the extent of positiveness or activeness however turned out to be over-dimensioned. In a second approach we directly distinguished between seven basic emotional states according to the MPEG-standard [9]: joy, anger, irritation, fear, disgust, sadness and neutral user state. This is also a far spread classification of emotions with more or less states [10]. However, a provided confidence level of an assumed emotion might still also be seen as a measurement of its extent.

The detected emotions recognized by the methods presented in this thesis are used in our man-machine interfaces. We want to recognize errors in the man machine- interaction by a negative user emotion. If a user seems annoyed after a system reaction error-recovery strategies are started. On the other hand a joyful user encourages a system to train user models without supervision. First or higher order user preferences can be trained to constrain the potential intention sphere for erroneously recognition instances like speech or gesture input. To do so a system online needs a reference value like a positive user reaction. Furthermore our system initiatively provides help for a seemingly irritated user. Control or induction of user emotions is another field of application that requires the knowledge of the actual emotion. For example in high risk-tasks it seems useful to calm down a nervous person, do not distract her by shortening dialogues, or keep a tired user awake.

REFERENCES

[1] L. Yang, “The expression of emotions through prosody”, ICSLP 2001, Beijing, China, Proc. Vol. 1, 2000, pp. 74-77.

[2] W. Yoon, K. Park, A study of emotion recognition and its applications, in: Modeling Decisions for Artificial Intelligence, vol. 6417, 2007, pp. 455–462.

[3] http://www.speech-therapy-information-and-resources.com/voice

[4] Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S., 2004. An acoustic study of emotions expressed in speech. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP ’04), Korea, Vol. 1, pp. 2193–2196.

[5] Schu¨ ller, B., Rigoll, G., Lang, M., 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP ’04), Vol. 1, pp. 557–560.

[6] N. Amir, and S. Ron, “Towards an automatic classification of emotion in speech”, in Proc. of ICSLP, Sydney, Dec. 1998, pp. 555–558.

[7] Kwon O., Chan K., Hao J., Lee T. “Emotion Recognition by Speech Signals”, Proc. of Eurospeech. 2003, Genewa, p. 125-128, September 2003.

[8] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall.,1993

[9] L. Rabiner, B. H. Juang, "Prentice Hall", Prentice Hall, 2010.

[10] Banse, R. and Scherer, K.R. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology.70: 614-636, 1996; Scherer, K.R., 2000b. Emotion effects on voice and speech: paradigms and approaches to evaluation. In: Proc. ISCA Workshop on Speech and Emotion, Belfast, invited paper; Scherer, K.R., 2003. Vocal communication of emotion: a review of research paradigms. Speech Comm. 40, 227–256.

[11] Scherer, K.R., R. Banse, H.G. Wallbott, and T. Goldbeck. 1991. Vocal cues in Emotion Encoding and Decoding. MorivarionandEmotion 15:123-148

[12] Goh,A.T.C., "Back-propagation neural networks for modelling complex systems", Artificial Intelligence in Engineering, vol.9, no.3, pp.143-151,1995.

[13] Schuller, G. Rigoll, M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine belief network architecture, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04), vol. 1, IEEE, 2004, pp. 577–580.

[14] M. Gilke, P. Kachare , R. Kothalikar , V. Pius, M. Pednekar, Int. Con. Elec. Eng. Info. 49 (2012), 150-154

[15] V. Arulmozhi, International Journal of Wisdom Based Computing, Vol. 1 (2), August 2011, 59-60.

[16] http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/