Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI:10.21474/IJAR01/4803
DOI URL: http://dx.doi.org/10.21474/IJAR01/4803
ISSN NO. 2320-5407

RESEARCH ARTICLE

A CASE STUDY OF SPEECH EMOTION RECOGNITION.

Prof. Anni. U. Gupta, Prof. M. S. Wagh and Prof. P. A. Patil.
Assistant Professor, E&TC DeptMET's IOE, Adgaon,Nashik.

Manuscript Info

Manuscript History

Received: 10 May 2017
Final Accepted: 12 June 2017
Published: July 2017

Key words:-
Classifier, LPC, MFCC, SER, TEO.

Abstract

The speech emotion recognition technology has a potential to provide considerable benefits to the national, international industry and society in general. Speech Emotion Recognition is a vital part of efficient human interaction and has become a new challenge to speech processing. In this paper speech emotion recognition system has been reviewed. Also Features and classifier for recognition of speech has been discussed.

Introduction:-

In Human interaction the information can be exchanged through speech, body language, facial expression etc. Speech is the most well-known way of communication between individuals. Speech is produced from a time varying vocal tract system excited by a time varying excitation source. This enables faster sharing, spreading and exchange of messages, ideas and inventions.

Conversion between humans is in reality not just what humans say, but also how they say it. Furthermore, facial expressions, as a part of non-verbal communication, are responsible for about 55%, voice intonation for about 38% and actual words for 7% of the message perception [1]. Expressions are not enough to correctly understand the frame of mind and objective of a speaker and thus the introduction of human social skills to human-machine communication is of supreme importance. This can be achieved by the researching and creating methods of speech modeling and analysis that embrace the signal, linguistic and emotional aspects of communication. The Emotion can be expressed through social behaviours and classified as shown in Figure 1,

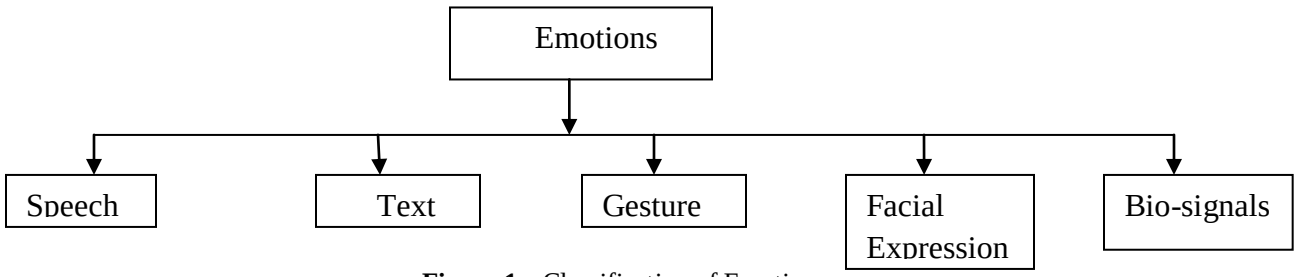


Figure 1:- Classification of Emotions

Corresponding Author:-Prof. Anni. U. Gupta.
Address:-Assistant Professor, E&TC DeptMET's IOE, Adgaon,Nashik.

Out of this Speech is one of the fastest & most efficient methods of communication between humans. People can exchange information and emotions rapidly through speech signals. Speech is a composite signal containing information about individual emotions, language, tone, pitch and message to be conveyed. Emotions are an intrinsic part of speech which determines the actual state of mind of an individual. Emotions play an important role in social, interaction, human intelligence, perception etc. One important aspect of human-computer interaction is to teach computer to understand human's emotion through voice, so that it can give different response. There are different types of emotions in speech are as follows:-

- Fear → frightened
- Anger → annoyed
- Sadness → gloomy
- Joy → pleased
- Disgust → erroneous
- Surprise → unsuspected
- Trust → an optimistic feeling

Studies have found that some emotions, such as fear, joy and anger, are portrayed at a higher frequency than emotions such as sadness.

Speech Emotion Recognition system:-

Speech Emotion Recognition system (SER) is the inter-disciplinary sub-field of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers.

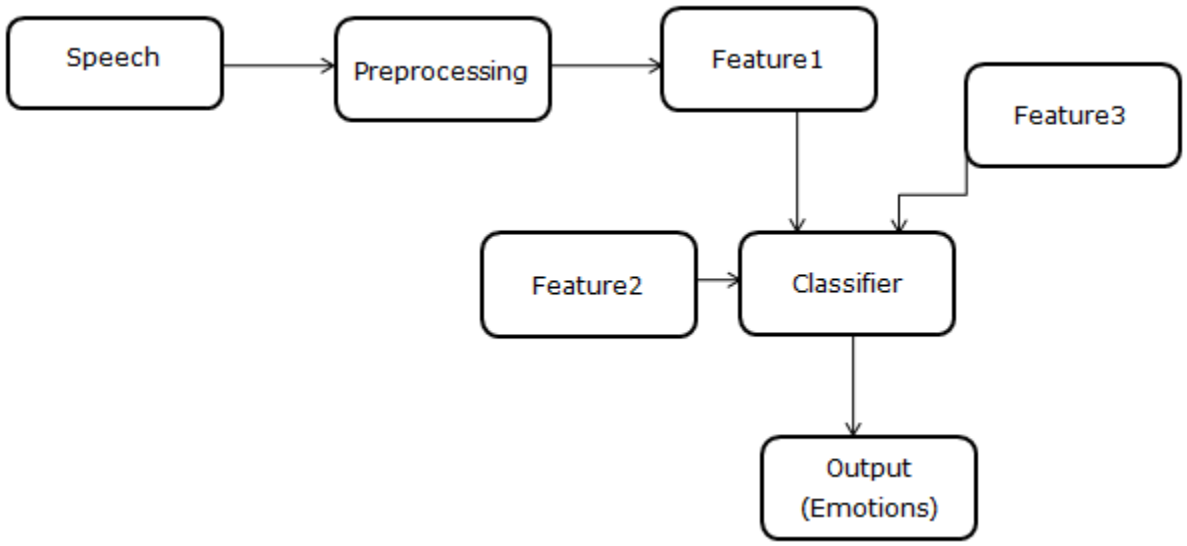


Figure 2:- Basic Flow of Speech Emotion Recognition.

The Basic flow of SER system is as shown in Figure 2. The system flow as follows, [2]

1. Input values are represented by audio signals from created database and used for training and testing [3]. The audio signals can from various sources irrespective of gender. This input is given to the preprocessing block for further use.
2. Preprocess mainly includes sampling, normalization and segmentation.
   a. Speech voice is analog signal and it needs to be converted into digital signal to process in computer. Sampling theory is used to reconstruct the original analog signal.
   b. The goal of normalization is to compensate for different recording settings among the databases. Normalization process use the signal sequence divided by maximum value of the signal to make each sentence has a comparable volume level.

c. Speech is a random signal and its characteristic is changing with time, but this change is not instant. Therefore segmentation process divides the signal sequence into many frames with overlap. Overlapping is used to avoid loss of data due to aliasing
d. Thus the preprocessing system extracts some appropriate quantities from audio signal, such as pitch, energy etc. by reducing the inconsistency.

3. These quantities are summarized into reduced set of features. The feature extraction is to some extent invariant to the changes in the speaker. So it involves analysis of speech signal [4].
4. A classifier learns in a supervised manner with example data how to associate the features to the emotions.

Feature Extraction:-

Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm [5]. Therefore Feature extraction is based on partitioning speech into small intervals known as frames [3]. The speech features are classified as in Figure 3,

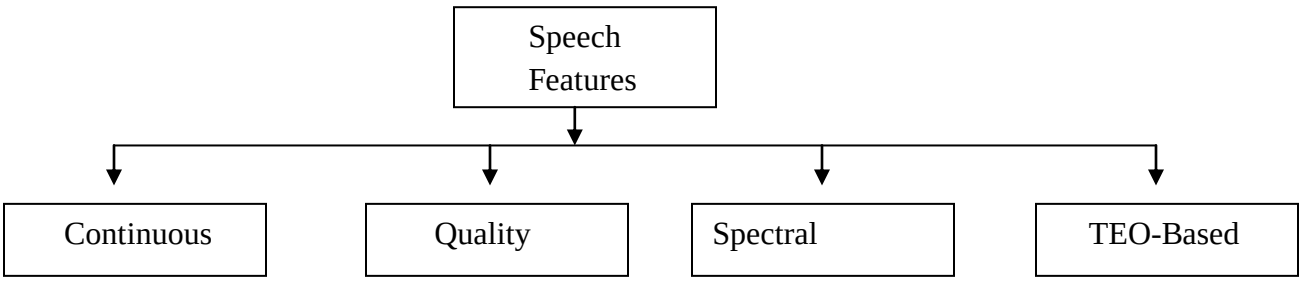


Figure 3:- Types of speech features

Continuous Features:-

Pitch: Pitch is an important attribute of voiced speech. The Pitch is the fundamental frequency of the vocal cords vibration (also called F0) followed by 4-5 Formants (F1-F5) at higher frequencies. Typical pitch values for male and female are 85-155 Hz and 165-255 Hz respectively. But the singer's vocal range is from bass to soprano i.e. 80 Hz-1100 Hz. Three mostly used methods for estimation of pitch include, autocorrelation of speech, cestrum pitch determination and SIFT pitch estimation.

Energy: Energy is the basic and most important feature in speech signal. To obtain the statistics of energy feature, we use short-term function to extract the value of energy in each speech frame. Then we can obtain the statistics of energy in the whole speech sample by calculating the energy, such as mean value, max value, variance, variation range, contour of energy [6].

Formants: The relation between a glottal airflow velocity input and vocal tract airflow velocity output can be approximated by a linear filter with resonances called Formants (like resonances of organ pipes and wind instruments). Generally, the frequencies of the formants decrease as the vocal tract length increases. Therefore, a male speaker tends to have lower formants than a female, and a female has lower formants than a child.

Qualitative Features:-

Voice quality: It refer to any of the suprasegmental properties of speech that result from how vocal apparatus is configured. Example: nasality. Usually voice quality is referring specifically to the properties of speech affected by stuff inside larynx. **Vocal tension** and **harshness** indicate that phonation is being produced with excessive strain in the laryngeal musculature. This may result in over adduction of vocal folds which creates disturbances in the vibratory pattern of the folds.**Harsh voice** is due to the very strong tension of the vocal folds (especially medial compression and adductive tension), which results in an excessive approximation of the vocal folds. Harsh phonation is therefore irregular in both cycle duration and amplitude. The characteristic fundamental frequency is above 100 Hz.Breathy voice is normally regarded as a compound phonation type (voiceless + modal).

Spectral Features:-
The analysis of the speech signal is always the foundation of related processing techniques. So spectral features of speech signals is important which includes methods like LPC, MFCC, and PLP.

LPC (Linear Predictive Code): It is desirable to compress signal for efficient transmission and storage. Digital signal is compressed before transmission for efficient utilization of channels on wireless media. For medium or low bit rate coder, LPC is most widely used [7]. The LPC calculates a power spectrum of the signal. It is used for formant analysis [8]. LPC is one of the most powerful speech analysis techniques and it has gained popularity as a formant estimation technique [9].

MFCC (Mel-Frequency Cepstrum Coefficients): MFCC is based on the characteristics of the human ear's hearing, which uses a nonlinear frequency unit to simulate the human auditory system. Mel frequency scale is the most widely used feature of the speech, with a simple calculation, good ability of the distinction, anti-noise and other advantages [10].

PLP(Perceptual Linear prediction): PLP models the human speech based onthe concept of psychophysics of hearing. PLP discards irrelevant information of the speech and thusimprovesspeech recognition rate. PLP is identical to LPC except that its spectral characteristics have been transformed to match characteristics of human auditory system.

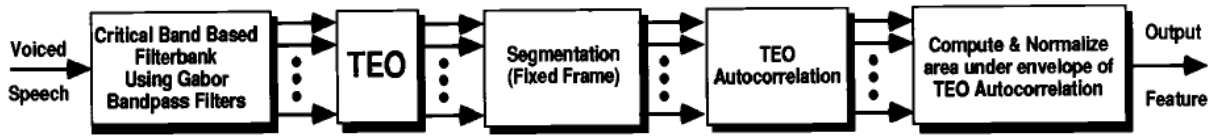
The main difference between PLP and LPC analysis techniquesis that the LP model assumes the all-pole transfer function of the vocal tract with a specified number of resonances within the analysis band. The LP all-pole model approximates power distribution equally well at all frequencies of the analysis band [11].

TEO-based:-
TEO is a key feature in recognizing stress. Vortex flow varies accordingly to emotional state of anger or stressed speech because there is fast air flow that causes vortices located near the false vocal folds that gives additional excitation signals other than pitch. To measure this energy which is produced by such a non-linear process, Teager developed an energy operator which is known as Teager energy Operator (TEO) given as follows,

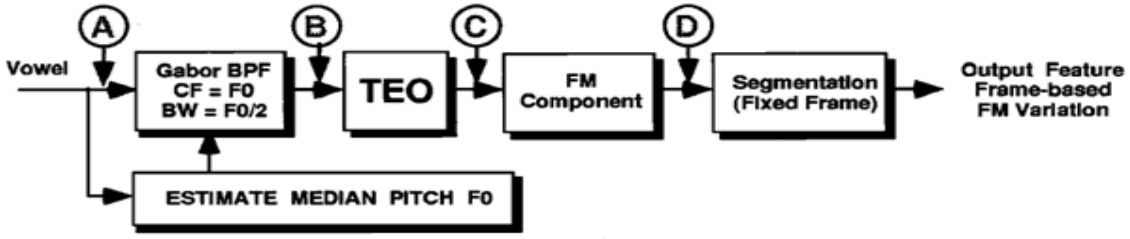
$$\Psi[x(n)] = x2(n) - x(n+1)\,x(n-1)$$

Where, ψ [.] is the TEO and x (n) is the nth speech sample point. if the speech signals broken into smaller bands also called as critical bands(CB) and TEO parameters are calculated for each band, presence or absence of additional harmonic components can be easily obtains which can be further used for processing. For this, smaller bands of speech spectrum are obtained by Gabor filters before calculation of TEO profile for each band.

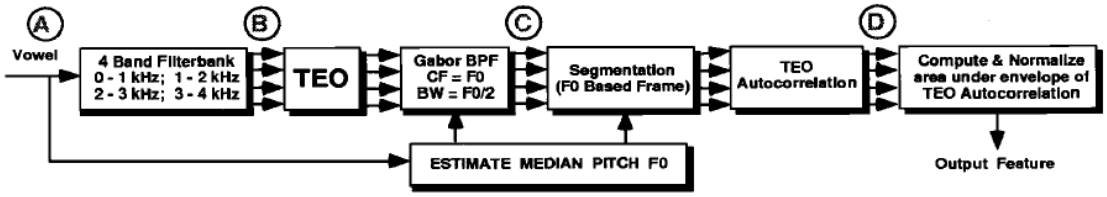
The area under normalized auto correlation environment is then calculated in each critical band to give **TEO-CB-Auto-Env**(Critical Band Based TEO Autocorrelation Envelope) features. Same frequency range was followed for 15 critical bands [10]. TEO-CB-Auto-Env method is used for emotional stress classification [12].



The motivation for the **TEO-FM-Var** (Variation of FM Component) feature is to capture stress dependent information that may be present in changes within the FM component.



The **TEO-Auto-Env** (Normalized TEO Autocorrelation Envelope Area) feature is obtained by passing the raw input speech through a filterbank consisting of 4 bandpass filters (BPF).



TEO-FM-Var and TEO-Auto-Env features are not as effective for stress classification because theydepend on pitch estimation accuracy The TEO-CB-Auto-Env feature, however, is the best feature evaluated for stress classification in terms of both accuracy and reliability.

Classifier:-

Classification algorithm is applied on different set of inputs for each extracted features of emotional speech. 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 [13].

Different classifiers are as discussed below:

Sr. No.	Name of Classifier	Description
1.	Support Vector Machine (SVM)	1.Optimality of the training algorithm 2.Existence of excellent data-dependent generalization bounds 3.Using kernel functions to nonlinearly mapping
2.	Hidden Markov Model (HMM)	1.Easy implementation 2.Solid mathematical basis 3.Small training time 4."Pause" occurrence may confuse state transition
3.	K Nearest Neighbor (KNN)	1. Locating the instance in feature space and comparing it with the k nearest neighbors (training examples) and labeling the unknown feature with the same class label as that of the located (known) neighbor. 2. The majority vote decides the outcome of class labeling.
4.	Gaussian Mixture Models (GMMs)	1.Easy implementation 2.Need to determine the optimal number of Gaussian components 3.Smallest training time 4.Cannot model temporal structure
5.	Linear Discriminant Classifiers (LDC)	1. Feature values to identify which class (or group) it belongs to by making a classification decision based on the value of a linear combination of the feature values. 2. Presented to the system in a vector called a feature vector.
6.	Artificial Neural Network (ANN)	1. More effective in modelling nonlinear mappings 2. Better performance in low relative training examples 3. Training time: Large 4. Classification accuracy is low
7.	Decision tree algorithms	1. based on following a decision tree in which leaves represent the classification outcome, and branches represent the conjunction of

		subsequent features
8.	Boostexter	1. An iterative algorithm 2. Focuses on text categorization tasks. 3. Deal with both continuous valued input (e.g., age) and textual input (e.g., a text string).

Application:-

By recognizing the emotional content it can be useful in domains like software engineering, website customization, education, and gaming. Some of the application of the speech emotion recognition system includes [14]:

1. Psychiatric diagnosis, lie detection
2. Call center may be used to examine behavioral study of call attendants with the customers which helps to improve quality of service of a call attendant.
3. Aircraft cockpits, speech recognition systems trained to recognize stressed speech are used for better performance.
4. Emotion analysis of telephone conversation for crime investigation department.
5. Useful for enhancing speech based human machine interaction.
6. Interactive movie, storytelling & E-tutoring applications.
7. Conversation with robotic pets and humanoid associates.

Conclusion:-

This study aims to provide an easy guide to the researchers for carrying out their research in the field of speech emotion recognition. The success of emotion recognition is dependent on appropriate feature extraction as well as proper classifier selection from the sample emotional speech. This survey paper deals with the most commonly methods of feature extraction and classifier.

References:-

1. 'Facial expression recognition: A brief tutorial overview', C. C. Chibelushi F. Bourel ,CVonline: On-Line Compendium of Computer Vision vol. 9, 2003.
2. 'A Study of Speech Emotion Recognition Methods' by Aastha Joshi, RajneetKaur , IJCSMC, Vol. 2, Issue. 4, April 2013.
3. 'Pattern Recognition Methods and Features Selection for Speech Emotion Recognition System' by PavolPartila, MiroslavVoznak, and JaromirTovarek, The Scientific World Journal Volume 2015 (2015), Article ID 573068
4. FEATURE EXTRACTION FOR SPEECH RECOGNITON', Manish P. Kesarkar EE. Dept, IIT Bombay, November 2003.
5. 'Speech Enhancement, Modeling and Recognition- Algorithms and Applications', S. Ramakrishnan ISBN 978-953-51-0291-5, Published14, March, 2012.
6. 'Automatic emotional speech classification', D. Ververidis, C. Kotropoulos, and I. Pitas, in Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol. 1, pp. 593-596, Montreal, May 2004
7. 'Nonlinear Feature Based Classification of Speech Under Stress', Guojun Zhou, Member, IEEE, John H. L. Hansen, Senior Member, IEEE, and James F. Kaiser, Fellow, IEEE, IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 9, NO. 3, MARCH 2001.
8. 'Neural network models of sensory integration for improved vowel recognition', B. P. Yuhas, M. H. Goldstein Jr., T. J. Sejnowski, and R. E. Jenkins Proc. IEEE, vol. 78, Issue 10, pp.1658–1668, Oct. 1990.
9. 'Voice Signal Processing For Speech Synthesis', Ovidiu Buza1, Gavril Toderean1, Alina Nica1, AlexandruCaruntu, IEEE International Conference on Automation, Quality and Testing Robotics, Vol. 2, pp. 360-364, 25-28 May-2006
10. 'Automatic Speech Emotion Recognition Using Support Vector Machine', P.Shen, Z. Changjun, X. Chen, International Conference On Electronic And Mechanical Engineering And Information Technology, 2011.
11. 'A Comparative Study of Audio Features For Audio to Visual Cobversion in MPEG-4 Compliant Facial Animation', Lei Xie, Zhi-Qiang Liu, Proc. of ICMLC, Dalian, 13-16 Aug-2006.
12. 'A Survey: Pre-processing and Feature Extraction Techniques for Depression Analysis Using Speech Signal' , DiptiPatil,ShamlaMantri,RiaAgrawal, ShraddhaBhattad,AnkitPadiya, RakshitRathi,International Journal of Computer Science Trends and Technology (IJCST) –Volume 2 Issue 2, Mar-Apr 2014
13. 'Gender driven emotion recognition through speech signals for ambient intelligence applications' Igor Bisio, Alessandro Delfino, Fabio Lavagetto, Mario Marchese and Andrea Sciarrone. IEEE Transactions on Emerging Topics in Computing, vol. 1, no. 2, 244-257, December 2013.
14. 'Emotion Recognition through Speech', Akshay S. Utane, S.L. Nalbalwar, Ph.D, IJAIS – ISSN : 2249-0868, NCIPET 2013.