Journal of Engineering and Applied Sciences 13 (13): 4917-4923, 2018
ISSN: 1816-949X

Basic Speech Feature Based Emotional Speech Analysis for Indian Native Language

[1]K.M. Shiva Prasad and [2]G.N. Kodanda Ramaiah
[1]Department of Electronics Engineering, Jain University, Bangalore, India
[2]Department of ECE, Kuppam Engineering College (KEC), Kuppam, India

Abstract: Speech is known to be the most convenient form of communication between individuals. This study deals with the modus operandi for estimation of basics acoustic features as also of the function of the vocal tract area, direct from the acoustic of speech wave form. Linear Predictive Coding (LPC) is used for estimation of the function of vocal tract area. LPC is based on autoregressive modelling of the speech signals. The method is an effective technique for speech analysis. It is predominant in speech research work for locating and estimating basic acoustic features and vocal tract area function and also to represent the speech for low bit rate transmission and application in storage area.The principal objective of emotional speech analysis/recognition is to identify the different basic emotional states (primary emotions) and to categorize them under positive (non-negative) and negative emotions. Emotion analysis finds use as a tool for improvement of the quality of service found in many speech supported application like call centres. It also helps in interfaces for human computer application and others.

Key words: Emotions, LPC, MATLAB, speech analysis, Telugu, emotional speech, supported application

INTRODUCTION

Speech which is the most convenient form of communication between individuals is not just a sequence of some sounds in a steady state with abrupt changes from one to another or some signals to be ignored after listening. It actually conveys information of both linguistic and non-linguistic types and also messages of sources of multiple levels of knowledge. Speech signals are illustrative of information bearing ones that emerge as functions of a single independent variable like chronological. Speech does not imply just information signals but more than that it is wave known for complexity as the acoustic output flowing from the speaker's effort.

Sound considered typical during the articulation of a phoneme is referred to as phone. Most known languages have anything between 20-40 phonemes. They provide a range of alphabets which help us to get acquainted with many words, words have syllables as their components. Syllables are sequences of phonemes (Prasad *et al*., 2015a, b). Sounds coming from speech are produced when air is exhaled from the lungs. They are modulated and get shaped by the vibrations of glottal cords and the resonance of the vocal tract during the process of air being pushed through the lips and the nose. Speech analysis is referred to as feature extraction of speech (He *et al*., 2011).

Feature extraction is the process in which speech wave gets converted into a known form of parametric representation which is meant for further speech processing which in turn finds applications in many areas. Parameters so obtained from significant cues in acoustics. Raw speech data is unintelligible. It has to be reduced to manageable quantity. Information has then to extract. This process is of vital importance in understanding and interpreting the speech signal. A through and meaningful analysis of speech features and the linking there of to perception still remain challenging tasks (Kumar *et al*., 2015a-c).

Speech is known for its information galore activity. It exploits frequency modulated amplitude. Such frequency and time modulated carriers convey information on words, identity of the speaker, his accent and style of speech as also the emotion with which he is changed. Examples of such carriers include resonance movements, harmonics, noise, pitch of voice, intonation and duration. The entire gamut of information is conveyed within the traditional telephone bandwidth of 4 kHz. Anything above this indicates audio quality and sensation (Prasad *et al*., 2015a, b).

Scope of the work: Languages of a large number are in use the world over. Each language is identified through speech. A speech can get recognition through speech but identification of his speech on the basis of known and

Corresponding Author: K.M. Shiva Prasad, Department of Electronics Engineering, Jain University, Bangalore, India

accepted parameters still remains a complex issue with a number of digital speech processing techniques involved. Speech is caused by excitation of many organs. It consists of several complex and also simple resonant frequencies called formants. Measurements of formant frequencies is a sin qua non for locating speech utterances and quality of voice speech signals help identification of acoustic features and emotion which determine speaker recognition. The spectrogram technique is implemented and extraction of acoustic features is done for speeches in Telugu language samples of which have been provided by IIT-Kgp (Altun and Polat, 2009; Yu *et al*., 2001).

Objective of the study: To generate the spectrogram and shape of vocal tract for different vowels that find place in a speakers speech through samples are taken. The researchers use the mat lab software for analysis of the formant frequencies and also samples of vocal tract shape speech. The spectrogram is used for extracting the formant frequencies and the pitch (Sheikhan *et al*., 2013).

MATERIALS AND METHODS

Speech database: IIT-Kgp SESC has been selected as the database for the purpose of analysis. It helps the study on speech emotion recognition. The proposed speech data base is the pioneering effort in the direction of telugu an Indian language. It is directed for analysis of emotions common in conversations of day-to-day occurrence. analysis of changed emotions in areas of text, session variability and gender is possible due to the corpus being extensive. The corpus or database known as simulated emotional speech corpus has been developed by the Indian Institute of Technology (Kharagpur) (Sheikhan *et al*., 2013).

Ten professional artistes (5 males and 5 females) from All India Radio, Vijayawada form the population for the study. They were chosen for the recording of the corpus. The artistes are persons with rich experience known for their ability to freely express emotions from neutral sentences. They were in the age group 25-40 with professional experience of 8-12 years. These are significant details of the artistes. The 15 sentences in 8 varied emotions were delivered to each artiste in one session. 10 such sessions formed the core. Thus, the total number of utterances worked out to 12,000 (15 sentences×8 emotions ×10 artistes×10 sessions) with 1500 utterances for each emotion.

Each signal was sampled at 16 kHz and represented as a 16 bit number the recording was done on alternate days to ensure the variability that is inevitable. Each artiste was asked to speak at a stratch all the sentences in a specific emotion. These ensured coherence among the artistes for each type of emotion. A single microphone was used for the entire database at the same location (He *et al*., 2011).

Problem statement: The corpus relating to speeches in Telugu language considered in this study is combination of vowels, yogavahakas and consonants. Consisting of voiced and unvoiced sounds. Different sentences in the language were used vis-a-vis basic acoustic features (formant frequencies and pitch more specifically) (Pao *et al*., 2008; Schuller *et al*., 2004) (Fig. 1).

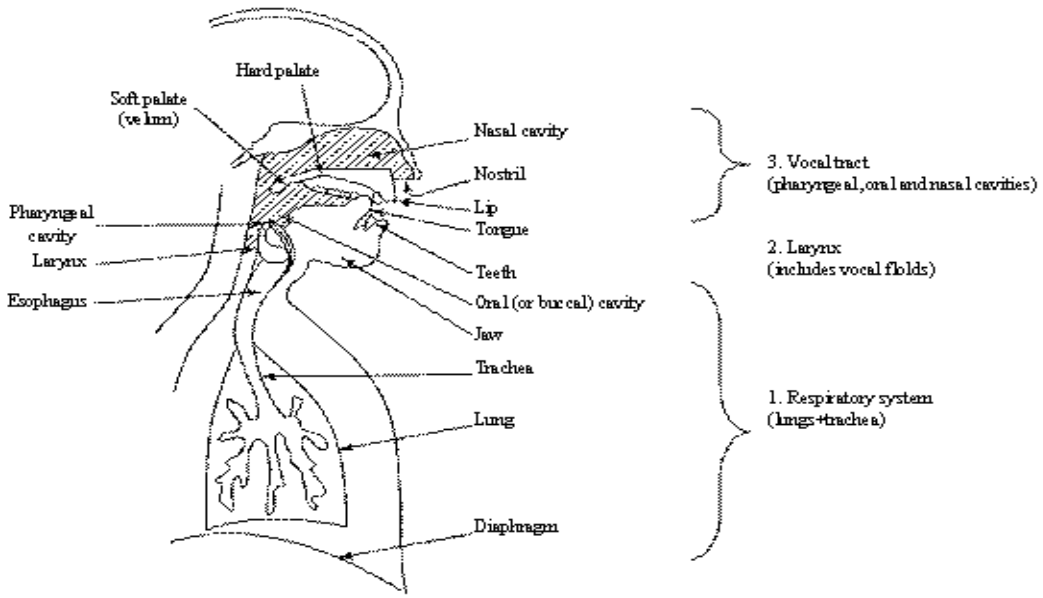

Fig. 1: Anatomy of speech production

Human-speech production mechanism: Figure 1 shows the anatomy of the human speech production systems, composed of three regions viz., larynx, vocal tract and respiratory system. This is capable of producing a variety of vibrations and spectral temporal composition which forms different speech sounds. Kumar *et al.* (2015a-c) the act of speech production begins with exhaling the inhaled air from the lungs (glottal volume). The air will sound like a random noise without any information in the absence of subsequent modulations. These are done through the manner and frequency of closing and opening of the glottal folds. Speech production is fused directly with the combination of volume and sinus cavity ending with the opening of the mouth (Prasad *et al.*, 2015a, b)

Basic acoustic features of speech: Formant frequency, pitch and intensity form the basic features in speech.

Formants: Formants are normal resonant frequencies (f_1-f_4). They depend on the position and the manner of articulation. These formants are generally considered important but acceptable speech quality requires higher formants. Formant frequencies appear as dark bands in a spectrogram. The mode of execution in the vocal cavity decides the resonating frequencies. Formants find use in speech recognition and speaker verification. They describe the vocal tract resonance. They are quantitative characteristics of the vocal tract, the position of which depends on its shape and physical dimensions. The name has been proposed by researchers in view of the resonant frequencies tending to form the over-all spectrum of the speech signal. Formant frequencies describe the shape of the vocal tract during the production of an emotional speech (Schuller *et al.*, 2004).

Pitch: This is also known as the repetition rate of opening and closing of vocal folds, being the basic frequency of vibration of the vocal folds referred to as fundamental frequency (f_0). Pitch is useful in the classification of speech as voiced and unvoiced. It is low for the former and absent in the latter. Auto correlation is the commonly used method for measuring pitch (Schuller *et al.*, 2004).

RESULTS AND DISCUSSION

LPC based speech analysis: Analysis of speech signals is done through estimation of speech parameters (like formants) and eliminating their effect from speech signal. It estimates the intensity and frequency of the remaining buzz. LPC Model for speech processing is shown in Fig. 2-4. The process of renewal of resonant or formant frequencies is referred to as inverse filtering while

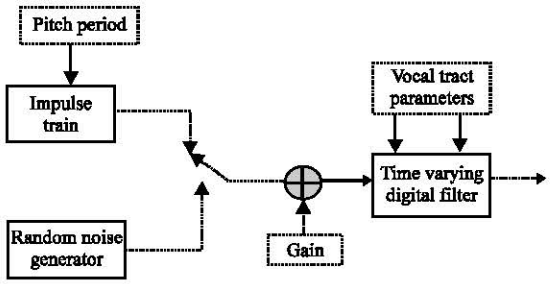

Fig. 2: LPC Model

the signal remaining after elimination is referred to as the residue signal. Speech segments of short duration called frames are used by LPC, the duration being 30 msec. Time variation in the digital system used in LPC is represented by the following transfer function (Kumar *et al.*, 2015a-c)

How emotion is created: Emotion originates from neural organization of the human brain. Generally, human brain processes the human emotion through chemical reaction. Emotion is created through a chemical reaction whose origin is not established in the nervous system due to arised situation. A cognitive state is created by number of situations and by a thinking process which relates those situations and draws some kind of conclusion as emotion (Yang and Lugger, 2010).

Emotional types: Generally their exists large types of emotions in real life namely: hot anger, cold anger, panic, fear, anxiety, despair, sadness, happiness, boredom, shame, pride, disguist and contempt. The commonly considered emotional states are anger, happiness, fear, sadness, disgust, boredom and neutral.

The basic emotions which are more primitive and universal than others are neutral, anger, fear and sadness. Neutral: it is an emotional state where emotionally lacking is noticeable. it is moderately negative, calm and weak (El Ayadi *et al.*, 2011).

Compassion: It is a emotion showing concern for the sufferings of other people (El Ayadi *et al.*, 2011).

Disgust: It is an emotional state or feeling that something is unpleasant, offensive or unacceptable (El Ayadi *et al.*, 2011).

Anger: It is a feeling or emotion showing extreme displeasure, very negative, very strong and very excited. it is an emotional state with high arousal levels which is characterised by the tense voice with faster speech rate. utterance duration, shorter inter word silence. There are two types, namely cold anger and hot anger. Hot anger is

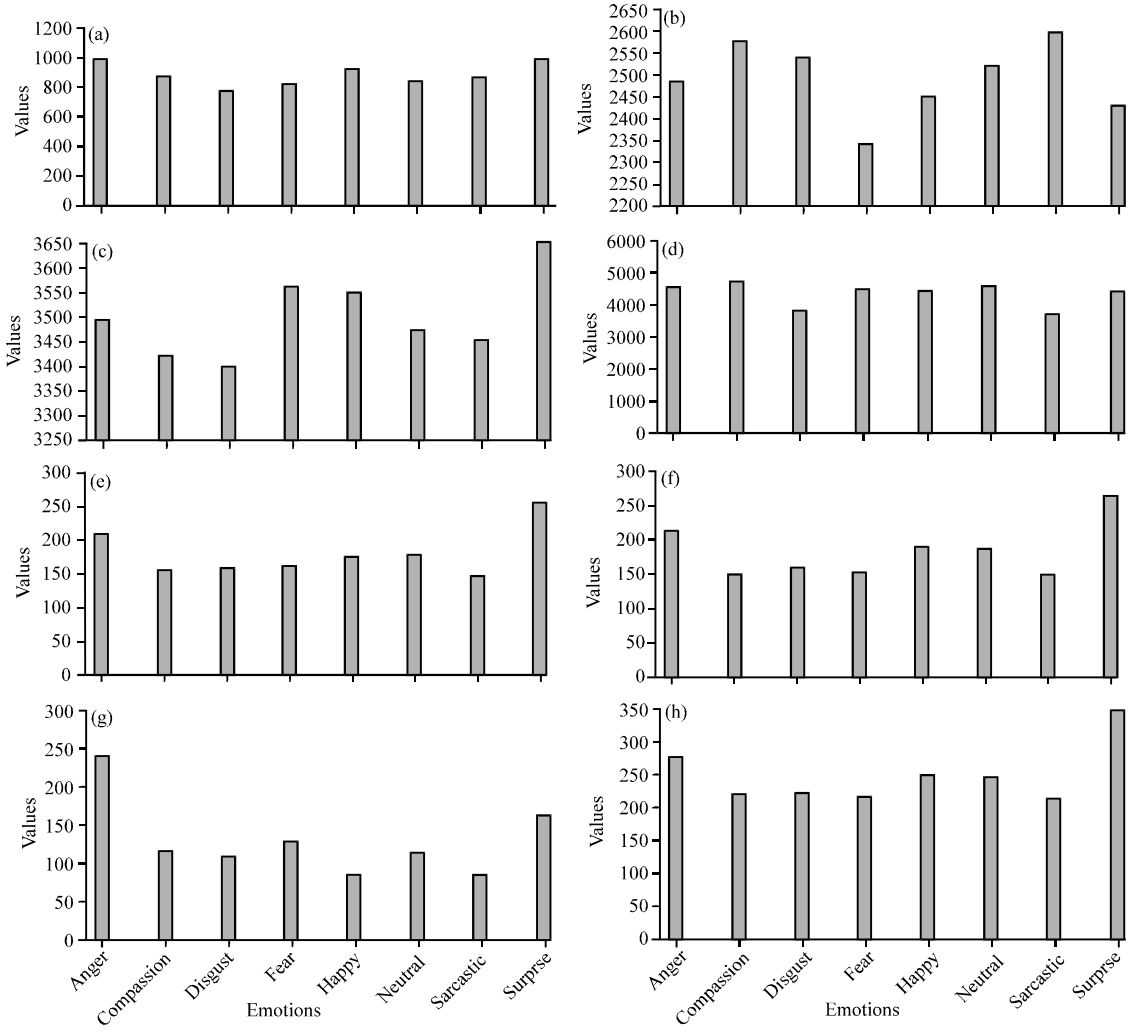

Fig 3: Comparative graph of 8 different emotions for sentence S_1 "Talli tandrulannu gowravinchavalanu": a) F1 (Hz); b) F2 (Hz); c) F3 (Hz); d) F4 (Hz); e) Mean (Hz); f) Median (Hz); g) Minimum (Hz) and h) Maximum (Hz)

more intense in quality, characterised by the increased articulation rate. Cold anger is characterised by increase in mean F_0, increase in high frequency energy and increase in mean intensity (Ayadi *et al.*, 2007)

Fear: It is an unpleasant emotion caused by threat of danger, afraid of others. it is characterised by increase both in mean F_0, F_0 range and increase in energy and also articulation rate (Ayadi *et al.*, 2007).

It is an emotional state with high arousal levels which is characterised by the tense voice with faster speechrate, moderately positive, high F_0 and broader pitch range yields higher heart rate and blood pressure (Ayadi *et al.*, 2007).

Sarcastic: It is an emotional state of hurting or mocking someone. (use of words which says the opposite of what you mean) (Ayadi *et al.*, 2007).

Surprise: It is an emotional state describes the feeling of mild astonishment or shock caused by something unexpected. characterized by high values of glottal velocity (Ververidis and Kotropoulos, 2006) (Table 1).

The detailed analysis of speech features and their relationship to human perception is a challenging task in speech processing. Speech conveys both linguistic as well as non-linguistic information, there exist many more methods of speech analysis each having their own merits and demerits for required application. But there is still no single method to be considered distinctly the best method for speech analysis or recognition. Speech analysis or

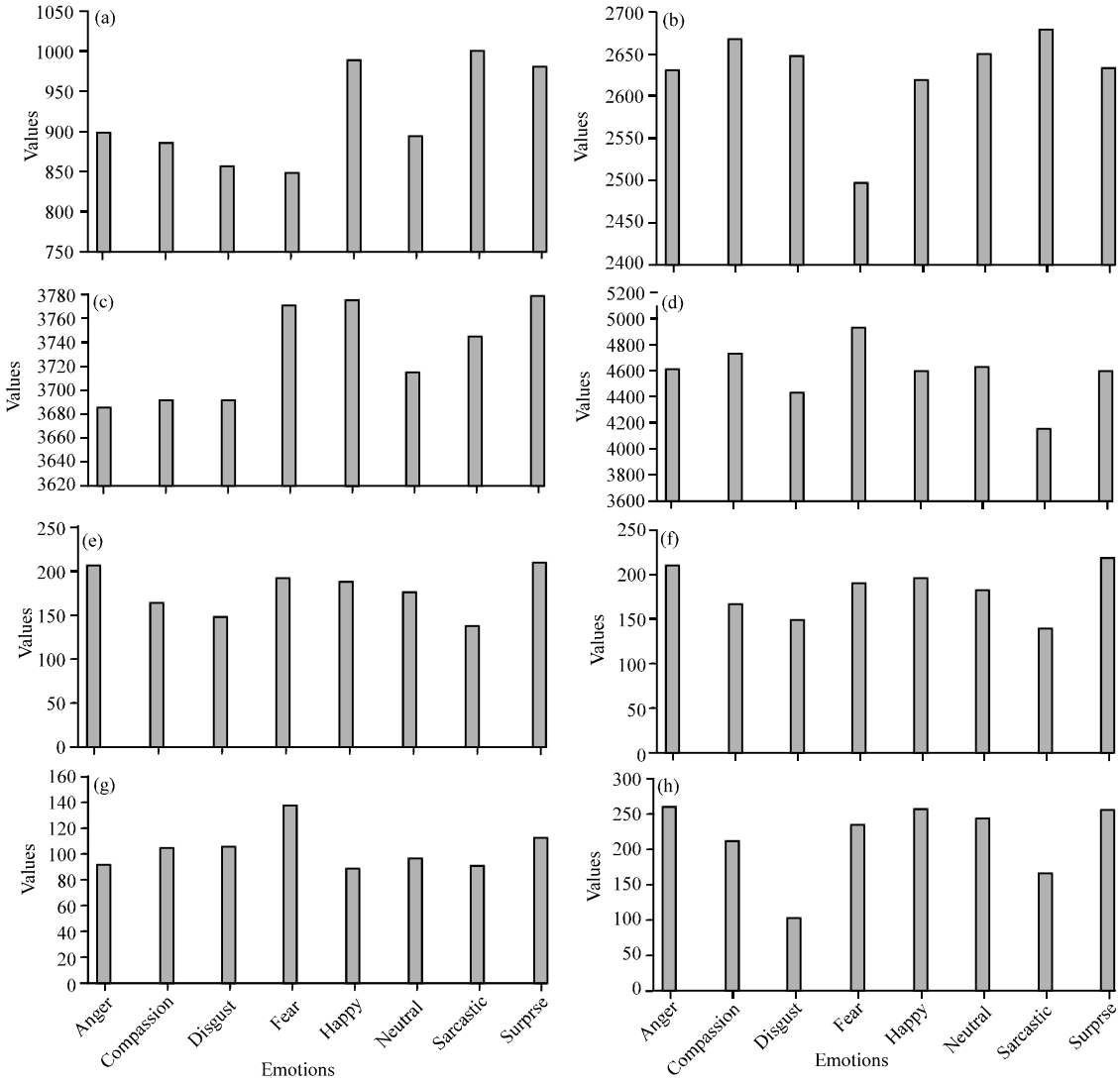

Fig 4: Comparative graph of 8 different emotions for sentence S_2 "meekosam chalachepadnundi chustannam": a); F1 (Hz); b) F2 (Hz); c) F3 (Hz); d) F4 (Hz); e) Mean (Hz); f) Median (Hz); g) Minimum (Hz) and (h) Maximum (Hz)

Table 1: Formants and various pitch listing

Parameters		Formant frequency (Hz)				Pitch (Hz)			
Emotion	Sentence	F_1	F_2	F_3	F_4	Mean	Median	Mini.	Maxi.
Anger	S_1	985.80	2490.98	3481.02	4908.81	210.56	218.890	240.680	262.980
	S_2	898.52	2630.33	3679.18	4670.41	206.26	212.450	92.605	263.400
Compassion	S_1	871.49	2587.55	3412.74	5057.56	156.41	153.780	117.170	210.430
	S_2	885.24	2667.82	3684.55	4789.48	164.76	168.780	105.590	216.670
Disgust	S_1	779.58	2548.61	3391.28	4094.20	159.24	164.577	109.250	213.020
	S_2	856.75	2647.78	3682.87	4484.87	147.02	150.710	106.000	105.200
Fear	S_1	833.15	2345.75	3543.55	4844.31	161.78	156.040	129.060	206.430
	S_2	847.39	2497.54	3755.75	5007.88	191.11	192.300	138.370	238.045
Happy	S_1	919.49	2456.95	3532.75	4762.24	176.97	193.420	87.140	237.260
	S_2	988.28	2619.61	3758.28	4654.19	187.43	198.370	88.210	260.920
Neutral	S_1	831.92	2527.00	3460.61	4943.05	179.59	191.890	114.150	235.080
	S_2	894.46	2649.90	3703.92	4691.82	175.48	183.730	96.130	246.290
Sarcastic	S_1	860.57	2607.21	3443.53	4009.25	147.69	151.790	84.030	202.830
	S_2	1000.31	2678.59	3731.25	4186.03	137.99	141.350	90.011	167.620
Surprise	S_1	1003.92	2435.33	3631.83	4740.87	255.86	267.560	162.840	331.190
	S_2	979.69	2632.28	3761.98	4652.02	210.21	221.170	112.187	260.175

speech signal front ends helps in extracting the speech features. There are many number of features can be used for purpose of emotion analysis/detection. The selection of feature set are depends on application requirement. it is very difficult to identify the best among available. The analysis/recognition performance of emotional speech using prosodic (Acoustic) features is not found to be appreciable but recognition can be improved by combining prosodic with spectral features. The emotional analysis/recognition can be extended to databases recorded in other native languages having wide range of emotional expression by taking number of speakers, different sex, different age group as subject to find the degree of universality of emotional features (Table 1).

CONCLUSION

In this resarch we investigate the acoustic properties of speech corpus associated with 4 primary emotions namely neutral, happy, sad and angry intentionally expressed by an untrained actors.The speech associated with anger and happiness are characterised by longer utterance duration (time), higher pitch and energy values with wide range. However, it is observed that slightly higher pitch in anger, fear compared to neutral speech. It is noticed that we cannot draw any conclusion on the variability of mean formant frequencies as a function of emotion, as they vary depending on which formant is considered. First formant frequencies range is has higher values in angry emotion when compared to other emotions. Angry and happy have higher mean F_0 values and greater variations compared to that of neutral speech. The mean of F_0 for neutral, sad, angry and happy are as listed in the table. It is confirmed that angry speech have higher F_0 values and greater variations compared to that of neutral and happy speech. Individual mean F_0 values for each emotion category is as shown in the table.

REFERENCES

Altun, H. and G. Polat, 2009. Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection. Expert Syst. Appl., 36: 8197-8203.

Ayadi, M.M.E., M.S. Kamel and F. Karray, 2007. Speech emotion recognition using Gaussian mixture vector autoregressive models. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Vol. 4, April 15-20, 2007, IEEE, Honolulu, Hawaii, USA., ISBN: 1-4244-0727-3, pp: IV-957-IV-960.

El Ayadi, M., M.S. Kamel and F. Karray, 2011. Survey on speech emotion recognition: Features, classification schemes and databases. Pattern Recognit., 44: 572-587.

He, L., M. Lech, N.C. Maddage and N.B. Allen, 2011. Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. Biomed. Signal Process. Control, 6: 139-146.

Kumar, A., G.K. Ramaiah and M.B. Manjunatha, 2015a. Speaker based vocal tract shape estimation for kannada vowels. Proceedings of the International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO), January 24-25, 2015, IEEE, Visakhapatnam, India, ISBN:978-1-4799-7676-8, pp: 1-6.

Kumar, A., M.B. Manjunatha and G.K. Ramaiah, 2015b. Vocal tract shape estimation of vowels & CVVC for diversified Indian English speakers. Proceedings of the International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO), January 24-25, 2015, IEEE, Visakhapatnam, India, ISBN:978-1-4799-7676-8, pp: 1-7.

Kumar, C.A., K.M.S. Prasad, M.B. Manjunatha and G.N.K. Ramaiah, 2015. Basic acoustic features analysis of vowels and C-V-C of Indian English language. ITSI. Trans. Electr. Electron. Eng., 3: 20-23.

Pao, T., Y. Chen, J. Yeh and Y. Chang, 2008. Emotion recognition and evaluation of mandarin speech using weighted D-KNN classification. Intl. Innov. Comput. Info. Control, 4: 1695-1709.

Prasad, K.M.S., C.A. Kumar, M.B. Manjunatha and G.N.K. Ramaiah, 2015a. Gender based acoustic features and spectrogram analysis for kannada phonetics. ITSI. Trans. Electr. Electron. Eng., 3: 16-19.

Prasad, S., A. Kumar and K. Ramaiah, 2015b. Various front end tools for digital speech processing. Proceedings of the 2nd International Conference on Computing for Sustainable Global Development (INDIACom), March 11-13, 2015, IEEE, New Delhi, India, ISBN:978-9-3805-4415-1, pp: 905-911.

Schuller, B., G. Rigoll and M. Lang, 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1, May 17-21, 2004, Montreal, Canada, pp: I-577-I-580.

Sheikhan, M., M. Bejani and D. Gharavian, 2013. Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method. Neural Comput. Appl., 23: 215-227.

Ververidis, D. and C. Kotropoulos, 2006. Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections. Proceedings of the 14th European Conference on Signal Processing, September 4-8, 2006, IEEE, Florence, Italy, pp: 1-5.

Yang, B. and M. Lugger, 2010. Emotion recognition from speech signals using new harmony features. Signal Process., 90: 1415-1423.

Yu, F., E. Chang, Y.Q. Xu and H.Y. Shum, 2001. Emotion Detection from Speech to Enrich Multimedia Content. In: Advances in Multimedia Information Processing-PCM 2001, Shum, H.Y., M. Liao and S.F. Chang (Eds.). Springer, Berlin, Germany, ISBN:978-3-540-42680-6, pp: 550-557.