Journal of Engineering and Applied Sciences 13 (13): 4932-4938, 2018
ISSN: 1816-949X

Analysis of Vocal Tract Shape Variability for Different Conditions in Indian English Vowels

[1]Anil Kumar Chandrashekar and [2]M.B. Manjunatha
[1]Department of Electronics Engineering, Jain University, Bangalore, India
[2]Akshaya Institute of Technology (AIT), Tumkur, India

Abstract: The study presents vocal tract shape estimation for vowels. Autoregressive model of vocal tract has been used herein for different recording conditions. Vocal tract shapes are known for phonetic distinction and speaker exclusiveness among vowels found by individual speakers ingrained in them. The researchers propose to use this autoregressive model for utterances made by speakers from south India taken on different occasions and in various recording conditions. Samples of speeches have been recorded assuming ice cold water and having an instant check of its effect and again later after 5 min. Vocal tract shapes so obtained are compared to those recorded under normal conditions. Vowel utterances made by speakers have been recorded 30 times, variations in the resulting vocal tract for different conditions on intra speaker basis. The researchers propose to investigate those and also, the variability as a parameter for the speech recognition process. The entire process has been done using MATLAB.

Key words: Auto regressive model, digital signal processing, dynamic model, LPC, vocal tract, process

INTRODUCTION

That speech is the most accepted and convenient means of communication is well known and recognized. The narrow concept of speech is that it is just a sequence of sounds punctuated by abrupt changes happening from one to another or some signals that are ignored and go into oblivion soon after utterance. It is much more than that. It is a unique signal that conveys information of linguistic and non linguistic type. Such information conveyed by speech is for knowledge of multiple levels speech signals typify (typify) information bearing that come up as a function of a single independent variable like time. Speech is not just an information signal. It is something beyond that it is actually a complex wave and acoustic output arising as a result of the speaker's effort (Kumar *et al*., 2015a-c).

Speech analysis is synonymous with feature extraction of speech. Speech sounds are sensations of air pressure variations produced by exhaled air and later modulated and shaped by vibration of glottal cords and the resonance of vocal tract during the time air is pushed out through the lips and the nose. Speech is signal with information galore exploring frequency modulated amplitude modulated and time-modulated carriers (example: resonance movements, harmonics, noise, pitch, power, duration). The objective is to convey information on words, speaker identity, accent, speech style and emotion. The entire gamut of information is basically conveyed in the large of the traditional telephone bandwidth of 4 kHz. Speech energy 4 kHz reflects audio quality and sensation (Prasad *et al*., 2015a, b).

Speech production: Speech is meant for communication. It has the distinct feature as a signal that carries a message or information. It is known to be an acoustic waveform that carries a message or information. It is known to be an acoustic waveform that carries temporal information from the speaker to the listener. Efficiency underlies acoustic transmission and reception of any speech. But this is applicable only for transmission over a short distance. There is a spread of radiated acoustic energy at frequencies that are used by the vocal tract and ear. But this gets reduced in intensity rapidly. Even on occasions when the source is able to produce substantial volume of acoustic power there is a support of only a fraction there of by the medium without any distortion while the rest of the it gets squandered in air dust particles molecular disturbance. It is also, resulting in getting over aero molecular viscosity. The ambient

Corresponding Author: Anil Kumar Chandrashekar, Department of Electronics Engineering, Jain University, Bangalore, India

acoustic noise places a restriction or limit on the sensitivity of the ear. Physiological noses to play this role in and around the ear drum. Voluntary, formalized motions of the respiratory and masticators apparatus have the speech as the acoustic end product. The closed loop has the ability to develop, control, maintain and correct it. Acoustic feedback of the hearing mechanisms and the kinesthate feedback of the speech musculature too have a role here. The central neurons system organizes and coordinates information from the senses which is then used for directing the function as also for delivering the descend, linguistically dependant, vocal articulator motion and acoustic speed.

Problem statement: The vowel speech database considered in our study is by considering all five vowels of English at three different conditions. Database is created by us at normal recording room environment, vowels (a, e, i, o, u) to be analyzed with respect to vocal tract shape variability.

The speech communication pathway: Figure 1 gives a simplified view of speech communication route starting from the speech and reaching the listener. An idea or a concept originates at the linguistic level of communication in the speakers mind. It is the stimulated longitudinal acoustic wave propagation in air starting from the speaker and terminating with the listener.

Pressure changes within the vocal tract are caused by the vocal tract and vocal cord movement. These are seem more specifically at the lips, initiating the sound wave that is known for propagation in space, this propagation activity occurs through space as a sequence of compassion and save faction of air dust molecules. As a result, temporal pressure variations are noticed at the listener's exterior ear which is funnel shaped collecting this acoustic energy efficiently. Later it manages to carry the media vibration ultimately to the final vole ration sensor, the ear drum set in the interior ear (Prasad *et al*., 2015a, b).

Variation in pressure experienced at the speakers lips, speakers lips causes sound. This sound propagates with channel losses, resulting in pressure variations at the listeners outer ear. The eventual vibrations in the ear-drum induces electric signals which move along the sensory noses to the brain. To the extent of the listeners perception the brain decodes these electrical signals known for the sensitivity to language. Later, it filters these signals in a recognized pattern which becomes known as a language speech perception and hearing.

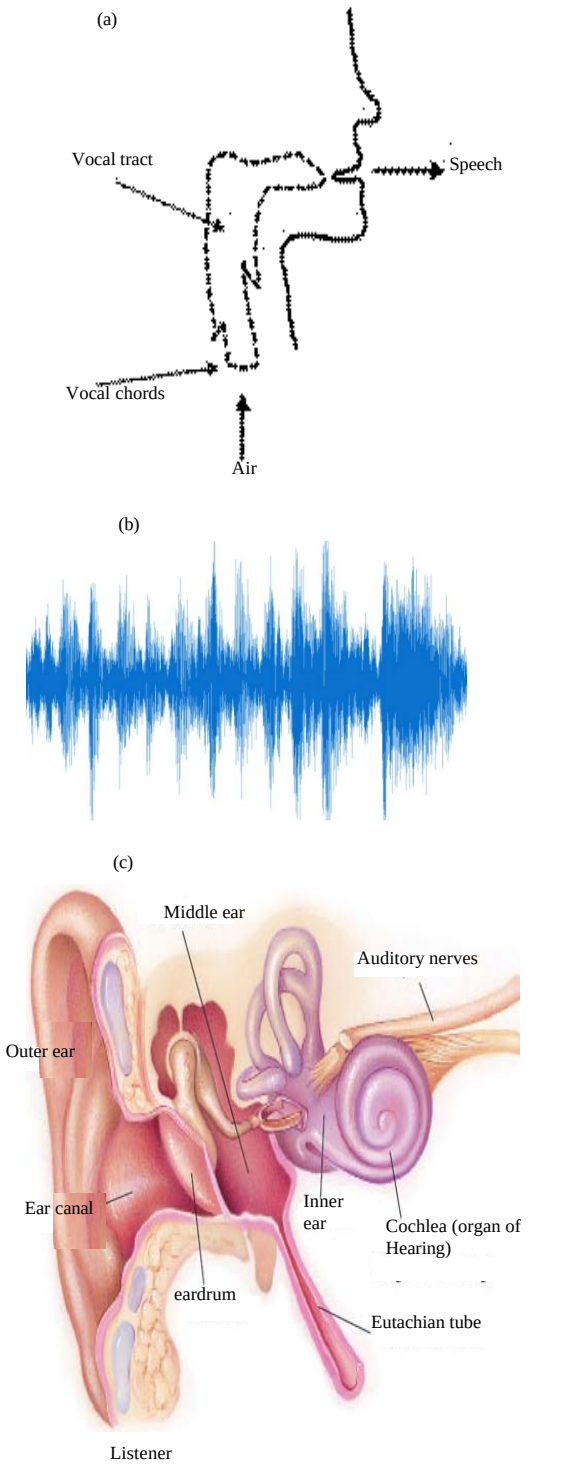

Fig. 1: Speech communication pathway: a) Speaker; b) Speach signal and c) Listener

Vocal tract shape: It is possible to model vocal tract as an acoustic tube resonance referred to as formants and anti-resonances and as a chain of cylinders of cross sectional area with varying features. Alteration in the form of the acoustic tube is caused by movement of articulators of vocal tract resulting in changes in frequency response. Poles are caused by resonances while anti-resonances arise from zeroes of frequency response. Kumar *et al*. (2015a-c) articulators are structures in the vocal tract which moves in the production of sounds of varying decibels. The important articulators are the lips and the tongue. They can be classified as direct and indirect the later are the jaw and mandible helping the positioning of the tongue and the lips for a verity of sounds such positioning of the tongue creates a number of vocal tract shapes that are needed for the production of speech sounds. The tract length for an adult male is generally 17 cm running from glottis to lips, the cross sectional area being up to 20 cm^2 (Prasad *et al*., 2015a, b).

Voiced sound: Vocal fold vibrations along with positioned articulators generates the basic sound which is referred to as "voiced sound". Voiced sound in singing is significantly different from voiced sound in speech.

Resonance: Voiced sound, frequency selected, modified and amplified by the vocal tract resonators (the throat, mouth cavity and nasal passages) and helped by articulators to produce a person's recognizable voice.

Unvoiced: The basic sound generated by vocal fold is apart and not vibrations with positioned articulators are called "Unvoiced sound". In the event of the vocal cords getting tensed up and closed, air flow experiences obstruction and air pressure builds up behind the constriction. This highly compressed air passes through the constriction in the vocal tract, becoming turbulent and producing what is known as unvoiced sound.

Articulation: The vocal tract articulators (the tongue, the jaws, the cheek, soft palate, the lips and the hyoid bone) have the function of modifying the voiced/unvoiced sound, they produce recognizable words helped by articulation.

Speech signals: Excitement of a fixed vocal tract produce vowels with quasi-periodic pulses of air, forced through the vibrating vocal cords. A quasi-periodic puff of air flow is the source, acting through vibrating vocal folds at a definite basic frequency. The term "quasi" is used considered with perfect periodicity never being acheived. The term "periodic" is used henceforth.. This shape of vocal-tract length from glottis to lips determines the resonant frequencies of the tract there by defining the produced sound.

MATERIALS AND METHODS

Concatenated tube model: The widely used model for speech production is based on the assumption of the capacity of the vocal tract to represent as a concatenation of small cylindrical tubes (Fig. 2 and 3). Independent variation of the cross sectional area of any tube for stimulating the changing shape of the vocal tract is seen. A change in the length of any tubular segment is effected for the purpose of reflecting the movements of articulators which include the lips, the jaws, the cheeks, the tongue and the hyoid bone. This variations in the shape and length of the tract at different points along its length, ultimately causes production of different sounds (Shah and Pandey, 2006).

Digital Signal Processing (DSP) techniques find use in modeling, use of the speech signal converted into its discrete time sequence on the assumption of all cylindrical segments being small are shown in Fig. 2 as of equal length.

Linear Predictive Coding (LPC): Linear predictive analysis is among the most powerful and widely used speech analysis techniques. This method has emerged as a predominate technique for estimating the basic speech parameters, e.g., pitch, formats, spectra, vocal tract area function. Render this method as highly important ability to provide accurate estimates of the speech parameters and in relative speed of computation. (Kodandaramaiah *et al*., 2010) .The ability of approximation seen by a speech sample as a linear combination of past speech samples is the basic concept that explores linear prediction analysis. The determination of a unique set of predictor coefficients is done by reducing the sum of the squared differences (over a finite

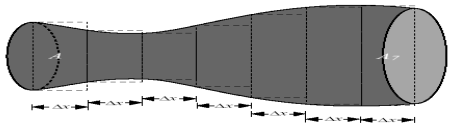

Fig. 2: Concatenated Tube Model

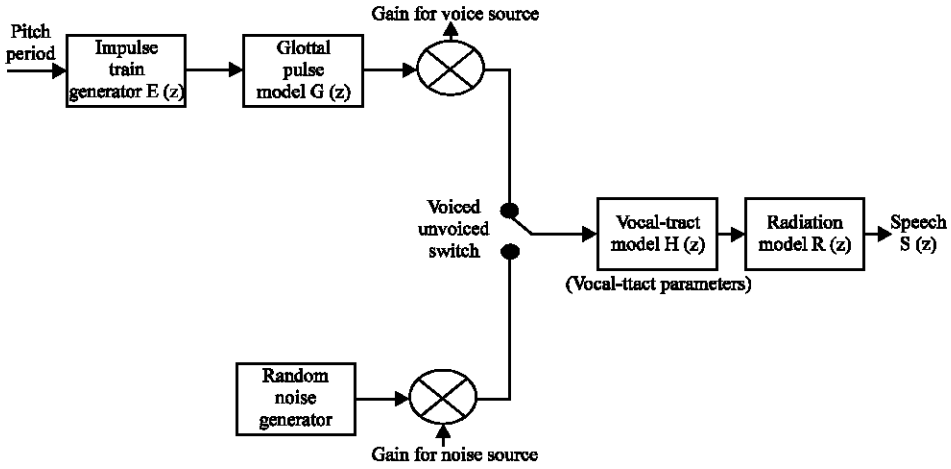

Fig. 3: Simplified model for speech production

interval) between the actual speech and the linearly predicted ones to the minimum. The predictor coefficients are the weighted coefficients used in the linear combination (Prasad *et al*., 2015a, b; Kodandaramaiah *et al*., 2010). Figure 3 shows the simplified model of speech production for providing idea of linear prediction is intimately linked to the basic speech synthesis model in which the sampled speech signal was shown as modelled as the output of a linear, slowly time-varying system excited by either quasi-periodic impulses (during voiced speech) or random noise (during unvoiced speech). The linear prediction method provides a robust reliable and accurate method for estimating the parameters that characterize the linear, time-varying system an all pole system function describes the linear system over intervals of short duration. A time varying digital filter with known steady state system function represents the composite spectrum effects of radiation, vocal tract and glottal excitation.

Speech database: The 30 samples of 30 subjects (male speakers aged about 18-25) at different times and at different conditions, i.e., normal sample by the instant of consuming ice cold water and with time lapse of 5 min were recorded using table top mic make of i-ball Model No. M27 with sensitivity -58±3 dB, frequency response of 100-16 kHz with sampling frequency of 22 and 100 Hz in MATLAB and the samples were normalized to have amplitude normalization by considering English vowels by Indian English speakers (a, e, i, o, u) (Kumar *et al*., 2015a-c).

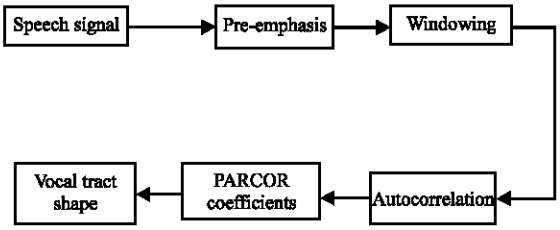

Fig. 4: Block diagram for vocal tract shape calculation

Block diagram for vocal tract shape calculation: Figure 4 shows the various steps involved in the estimation of the vocal tract shape by providing speech signal as input and with various co-efficient calculation and its description as follows.

Pre-emphasis: The area function obtained through use of reflection coefficients cannot be the area function of the human vocal trust. In the event of the pre-emphasis being used prior to linear predictive analysis for removal of defects arising out of glottal pulse and radiation, the area function that emerge as results are often similar to vocal tract configurations that find use in human speech. The speech production model discloses the speech undergoing a special tilt of -6 dB/octave. The use of pre-emphasis in displaying the spectrogram if speech signals is rather common (Khodai-Joopaari *et al*., 2004).

Window technique: The window function w(k-n) is a real window sequence that finds use in isolation of isolate the portion of the input sequence analyzed at a particular time index k. Many window functions have been generated to

improve upon the basic rectangular window design such as hamming, hanning, bartlett, blackman, kaiser, etc. each having different specification with regards to its frequency response. Laprie and Mathieu (1998) In this study, LP analysis was performed on frames weighted with the hamming window. This window, w (n) was chosen as it provides a good balance between its main lobe width and side lobe attenuation.

The hamming window is also deemed to be adequate in determining the accuracy for approximating the transfer function of the vocal tract. This is a crucial aspect when calculating reflection coefficients for quantization purposes (Black, 1998).

Autocorrelation method: The most widely used method to linear predictive analysis is called the autocorrelation method. It is the most popular method of short-term LP analysis. This method provides the most computationally efficient manner in determining the LP parameters with guaranteed stability. It avails of the Toeplitz property possessed by the autocorrelation matrix.

PARCOR coefficients: PARCOR coefficients are bounded ±1. This has been seen earlier. This features has given them the facility of being the attractive parameter for quantization. With a set of PARCOR coefficients, it is first possible to use them at step 2 of the Levinson Durbin algorithm followed by the getting of an algorithm that can help conversion which can be computed from a given set of predictor coefficients by working backward through the Levinson Durbin algorithm (Kuc *et al*., 1985).

Algorithm; Intra speaker vocal tract shape algorithm:

1. Using LPC (AR-Auto Regressive model) forward and backward and speech analysis determining the shape of vocal tract variability of the subject bringing out the vocal tract shape variability of an individual subject
2. Study of variability of the above vocal tract shape among 30 different subjects to high light and identify Intra speaker variability
3. The above identified can be adopted for personal identification similar to signature. It can also be named as vocal tract signature of an individual (Black, 1988)
4. Find the worst (with least variability) and the best maximum variability for each speaker (subject)
5. Finding the averages in Time for best and worst patterns for the ensure of 30 subjects (Paige and Zue, 1970)
6. Plot the resultant worst pattern and resultant best pattern for a subject for the vowel

RESULTS AND DISUSSION

Figure 5 shows the speech signal for vowel /e/ under normal recording condition, Fig. 6 depits the speech signal vowel /e/ in consumption of ice cold water and Fig. 7 represents speech sample vowel /e/ by after

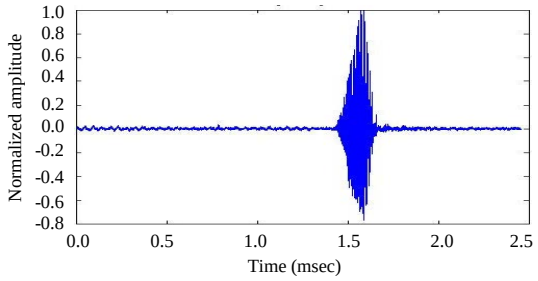

Fig. 5: Speech signal for normal sample /e/; Original signal

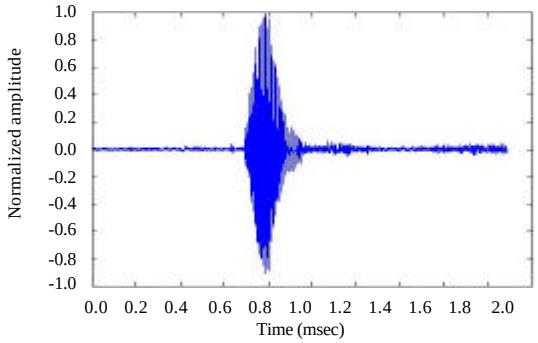

Fig. 6: Speech signal for sample /e/ after consumption of ice water; Original signal

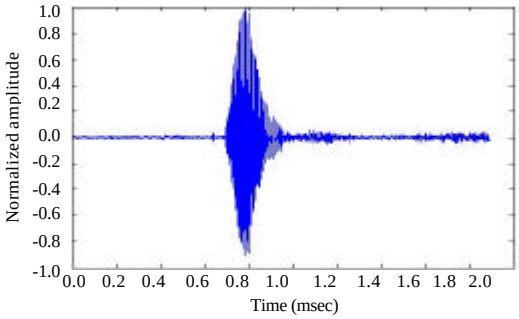

Fig. 7: Speech signal for speech sample /e/ after 5 min of consumption of ice cold water; Original signal

five min of consumption of ice cold water in the similar passion the speech signals are plotted for other vowels like a, i, o and u, under all the three recording conditions as mentioned above and the variations has been observed simultaneously.

Few noticeable variations can be observed while interpreting the same in the dynamic modelling of the vocal tract as plotted in the Fig. 8a vocal tract model of /e/ Fig. 8b vocal tract model of /a/ Fig. 8c vocal tract model of /i/ Fig. 8d vocal tract model of /o/ Fig. 8e vocal tract model of /u/.

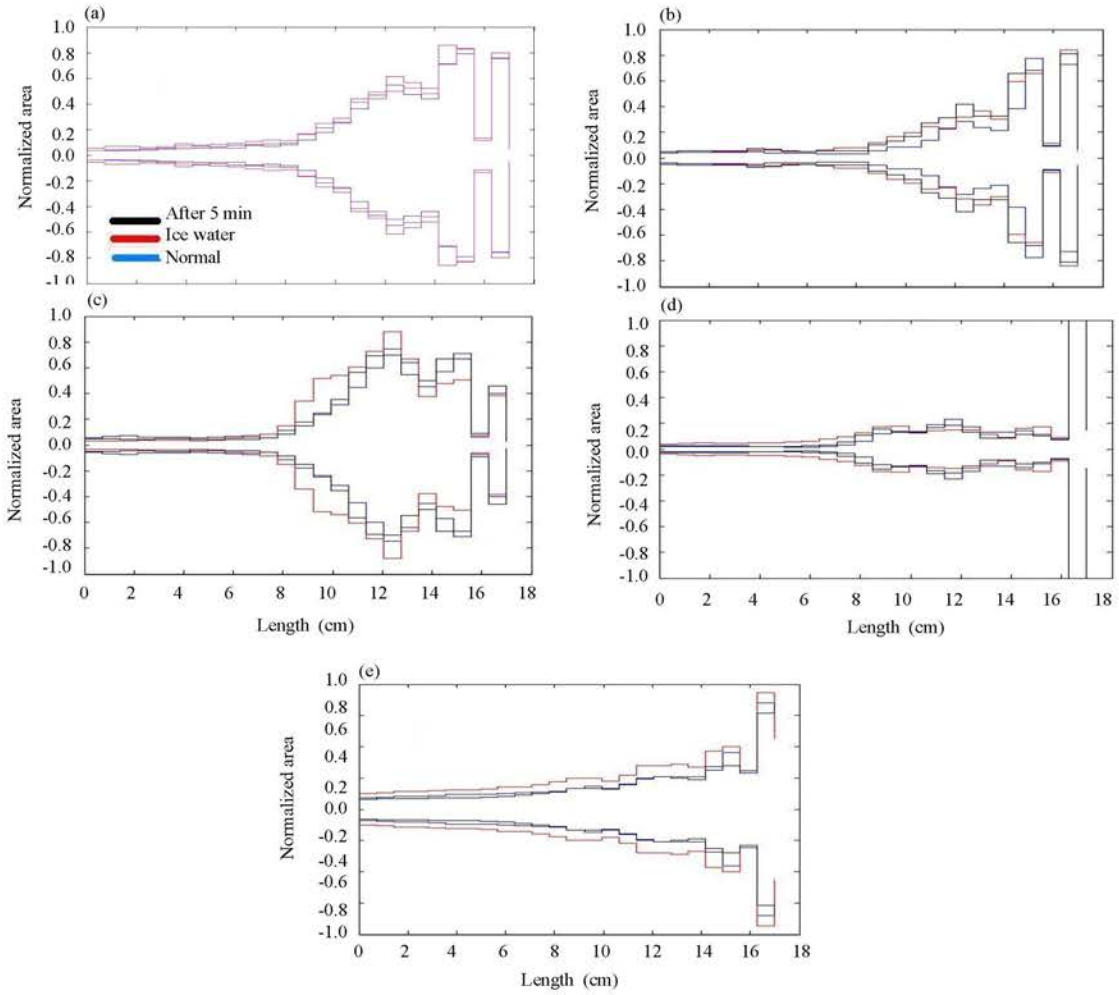

Fig. 8: a-e) Dynamic vocal tract model for vowel /e/; Dynamic vocal tract model for vowel /a/

As observed from the dynamic model in Fig. 8a the significant variability in vocal tract is observed in the region of 14-16 cm for the vowel /e/. From Fig. 8b, vocal tract will swing dominently in the area labelled 8-12 cm for the vowel /a/. In Fig. 8c, model represented for the vowel /i/ shows the clear variations for the length 10-14 cm. In regard of Fig. 8d, dynamic model of vowel /o/ the varibility in the vocal tract shape observed between 4-10 cm and lastly Fig. 8e gives the modeling of vowel /u/ by showing the deviation of model in the full length till 17 cm.

CONCLUSION

Using LPC analysis of speech, reflection coefficients are utilized for the estimation of vocal tract. Vocal tact length values are obtained for the vowels (a, e, i, o, u) for individual speakers. The speech input signal is separated into frames each frame length into 30 msec with an overlapping of 10 msec frame length the order of LPC filter is 25. Sampling rate of the speech signal is 22,100 Hz. Each frame consists of 663 samples. By using LPC method calculating the reflection coefficients. The incidence of the largest reflection of coefficients is seen in areas where the relative changes in vocal tract are the largest. Measurement of the area function of an individual taken as different occasions and in different contexts as also, the variability of the vocal tract shapes. The measurement is done on intra and inter speaker basis model getting vocal tract shape if the vowels for every speech. Vocal tract shape can be used as personal passwords or signature for verification and identification.

REFERENCES

Black, N.D., 1988. Application of vocal tract shapes to vowel production. Proceedings of the IEEE Annual International Conference on Engineering in Medicine and Biology Society, November 4-7, 1988, IEEE, New Orleans, Louisiana, pp: 1535-1536.

Khodai-Joopari, M., F. Clermont and M. Barlow, 2004. Speaker variability on a continuum of spectral sub-bands from 297-speakers non-contemporaneous cepstra of Japanese vowels. Proceedings of the 10th Australian International Conference on Speech Science and Technology, December 8-10, 2004, Macquarie University, Sydney, New South Wales, pp: 504-509.

Kodandaramaiah, G.N., M.N. Giriprasad and M. Mukundarao, 2010. Implementation of LPC based vocal tract shape estimation for vowels. Technol. World Q. J., 5: 97-102.

Kuc, R., F. Tuteur and J. Vaisnys, 1985. Determining vocal tract shape by applying dynamic constraints. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'85) Vol. 10, April 26-29, 1 985, IEEE, Tampa, Florida, USA., pp: 1101-1104.

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, 2015c. Basic acoustic features analysis of vowels and C-V-C of Indian english language. ITSI. Trans. Electr. Electron. Eng., 3: 20-23.

Laprie, Y. and B. Mathieu, 1998. A variational approach for estimating vocal tract shapes from the speech signal. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing Vol. 2, May 15, 1998, IEEE, Seattle, Washington, USA., pp: 929-932.

Paige, A. and V. Zue, 1970. Calculation of vocal tract length. IEEE. Trans. Audio Electroacoust., 18: 268-270.

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.

Shah, M.S. and P.C. Pandey, 2006. Estimation of vocal tract shape for VCV syllables for a speech training aid. Proceedings of the IEEE-EMBS 27th Annual International Conference on Engineering in Medicine and Biology Society, January 17-18, 2006, IEEE, Shanghai, China, pp: 6642-6645.