TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 2166 - 2171

Indian Vowels based Evaluation of First Fundamental Frequency and its Variant Bandwidths

Praveen. K.B[1], B. Siva Kumar[2]

[1]Research Scholar, Dept of Electronics Engg, PRIST University, Thanjavur
[2]Professor, Dept of TCE, Dr. Ambedkar Institute of Technology, Bangalore
[1]prvn.guru@gmail.com, [2]sivab881@gmail.com

Article Info
Volume 82
Page Number: 2166 - 2171
Publication Issue:
January-February 2020

Abstract

The study presents analysis of Indian English vowels based on fundamental frequency along with the bandwidths of the four fundamental frequencies for an untrained database for the speakers age ranging between 16 to 21 years. An autoregressive modelling is incorporated in addition to Linear Prediction Coding (LPC) for the analysis and estimation of fundamental frequency, bandwidth are calculated for various vowel recording speech samples. These parameters are considered as the tool for the phonetic distinction, speaker unique among the range of individual speakers under consideration. The research concern is to employ and trace the speakers with the ingrained speech parameters in it. The researchers propose is to use this autoregressive model for utterances made by speakers from south India taken on different vowels and SWIPE algorithm for the fundamental frequency estimation. The speech samples are recorded for the neutral condition of the speaker. The frequency components obtained are comparatively plots against the various individual bandwidths along the fundamental frequency of the every speaker which are uttered over an span of time for thirty individual times by the speaker, the mean values are taken into account for the comparative analysis for the investigation of the vowel uniqueness and its variability as a parameter for speech recognition criteria, the entire demonstration is done using MATLAB.

Keywords: *Auto-aggressive modelling, LPC, Matlab, vowels, speech signal, spectrogram.*

Article History
Article Received: *14 March 2019*
Revised: *27 May 2019*
Accepted: *16 October 2019*
Publication: *12 January 2020*

1. Introduction

Speech is one of the most important aspects if communication in human beings, Speech is the basic and most commonly used means of human communication. It is effective in the transmission of words, feelings, emotions, explanations, threats, affection and messages [1]. Speech is often successful in the perception of expression, accent, emotions, gender and many more human attributes. Fundamentally speech is considered to communicate with each other (transmission of information from source to destination and vice versa). Speech can be considered as auditory information including multi-layered spectral variations which help to convey intentions and identity. This moving air is modulated by the glottis and is often affected by the resonance of the vocal tract and nasal cavity, tongue and the opening/closing of the mouth.

Speech sounds are produced by air pressure vibrations generated by pushing inhaled air from the lungs through the vibrating vocal cords and vocal tract and out from the lips and nose airways. The air is modulated and shaped by the vibrations of the glottal cords, the resonance of the vocal tract and nasal cavities, the position of the tongue and the openings and closings of the mouth. [2]

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 2166 - 2171

It is a known fact that a given message described as sequences of symbols in discrete format can always be quantified with respect to the contents in terms of bits while the rate of such transmission is quantified in terms of bits per second (bps). Such information if transmitted at a bandwidth of 4 KHz. In general, audio qualities as well as sensation are conveyed at an energy level higher than 4 KHz. Such information is encoded appropriately at the source and then transmitted to be either recorded or manipulated to be finally perceived at the destination. Speech is considered to be a sequence of PHONEMES (acoustic sounds / symbols which are used to convey the spoken format of a given language). For instance, English language consists of 40-60 phonemes. The phones are such that each of the generated phonemes is affected by the neighbouring ones. Phonemes form the fundamental aspect in case of speech processing and are an integral part of modern communication systems as well. [3]

2. Speech Modelling

Figure 1 shows the complete process of producing and perceiving speech from the formulation of a message in the brain of a talker, to the creation of the speech signal, and finally to the understanding of the message by a listener. In their classic introduction to speech science, Dense and Pinson aptly referred to this process as the "speech chain". The process starts in the upper left as a message represented somehow in the brain of the speaker. The message information can be thought of as having a number of different representations during the process of speech.

For example the message could be represented initially as English text. In order to "speak" the message, the talker implicitly converts the text into a symbolic representation of the sequence of sounds corresponding to the spoken version of the text. This step, called the language code generator referred linguistic level as in Figure 1, converts text symbols to phonetic symbols (along with stress and durational information) that describe the basic sounds of a spoken version of the message and the manner (i.e., the speed and emphasis) in which the sounds are intended to be produced. The third step in the speech production process is the conversion to "neuro-muscular controls," i.e., the set of control signals that direct the neuro-muscular system to move the speech articulators, namely the tongue, lips, teeth, jaw and velum, in a manner that is consistent with the sounds of the desired spoken message and with the desired degree of emphasis. [4][5]

The end result of the neuro-muscular controls step is a set of articulator motions (continuous control) that cause the vocal tract articulators to move in a prescribed manner in order to create the desired sounds. Finally the last step in the Speech Production process is the "vocal tract system" that physically creates the necessary sound sources and the appropriate vocal tract shapes over time so as to create an acoustic waveform. As we move from text to speech waveform through the speech chain, the result is an encoding of the message that can be effectively transmitted by acoustic wave propagation and robustly decoded by the hearing mechanism of a listener. [6]

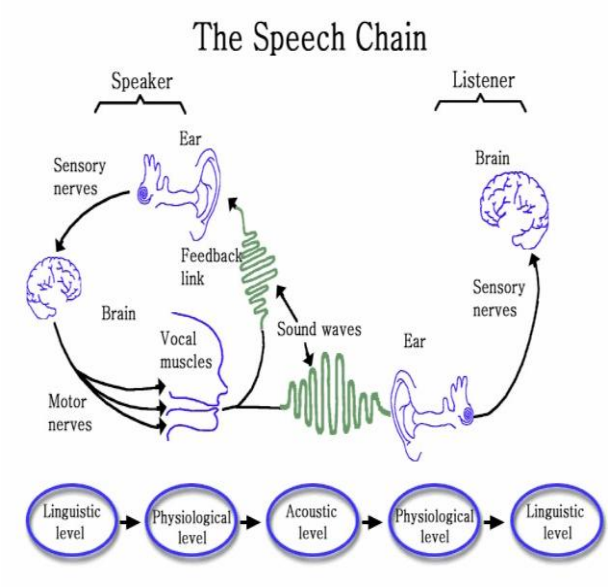


Figure 1: The Speech Chain: From message, to speech signal, to understanding

The complete speech chain consists of a linguistic level, physiological level and Acoustic level, as shown progressing to the left in the bottom half of Figure 1.The speech perception model shows the series of steps from capturing speech at the ear to understanding the message encoded in the speech signal.

3. Autoregressive modelling for Bandwidth Estimation.

The proposed algorithm can be summarized as different steps as represented in figure 2. In which it describes the necessary steps for frequency component estimation as follows:

a. Input speech signal.
b. Pre-emphasis
c. Windowing (Hamming window)
d. Autoregressive modelling
e. PARCOR coefficients extraction

Pre-emphasis

The area function obtained utilizing reflection coefficients are considered as the area function of the human vocal tract [52]. In the event of the utilization of pre-emphasis prior to linear predictive analysis to expel the impacts due to the glottal pulse smoothing and lip radiation impacts, the subsequent range capacities were found to be the same as vocal tract setups that would be utilized as a part of human speech. Pre-emphasis was completed at 6 dB for every octave rate, bringing about the adequate increment of 6 dB/octave.

January - February 2020
ISSN: 0193 - 4120 Page No. 2166 - 2171

Window Analysis

The Hamming window is sufficient for approximating the exactness for approximating the transfer function of the vocal tract. In the process of extracting reflection coefficients for quantization purposes, Hamming windows of length 30 ms have been utilized, with a cover of 10 ms joined, for obtaining smooth appraisals [7].

In this, LP investigation is performed on outlines weighed with the Hamming window. This window, w(n), is picked as it gives a decent harmony between its primary signal width and side projection constriction.

$$w(n) = \begin{cases} 0.54(1 - 0.85\cos\left(\frac{2\pi n}{N}\right)) & ; \ 0 \le n \le N-1 \\ 0 & ; \ \text{otherwise} \end{cases}$$

(1)

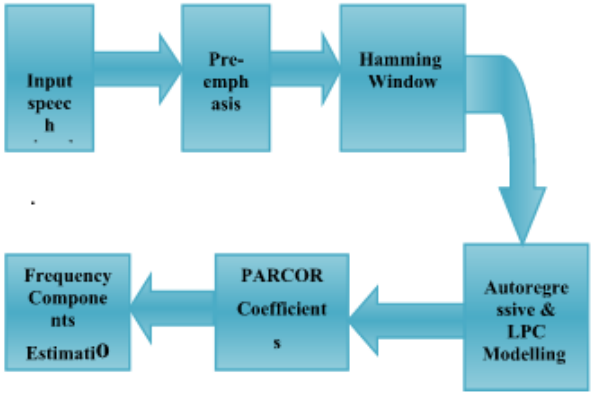


Figure 2: Autoregressive modelling for the Frequency Components

Auto-Correlation Analysis

Auto correlation of each edge of the windowed signal is then done [1]. In the autocorrelation strategy, the investigation fragment $s_n[n]$ is indistinguishably zero beyond the interim $0 \le n \le N-1$, it is communicated as it appeared in the equation 4.2.

$$s_n(n) = s(n+N)\, w(n) \qquad (2)$$

Where w(n) is a limited length window (Hamming window) that is indistinguishably zero outside the interim. The $s_n(n)$ is nonzero for the interim $0 \le n \le N-1$. The relating forecast blunder, $e_n(n)$, for a p^{th} arrange indicator is nonzero over the interim $0 \le n \le N-1+p$.

Partial Auto-Correlation Coefficients (PARCOR) Extraction

Extraction of the parameters involves the setting of the PARCOR coefficients, which are halfway esteems amid the estimation of Levinson-Durbin recursion [8]. The researcher has observed quantizing the middle esteems as hazardous is magnitude than quantizing the indicator coefficients straightforwardly, as the impact of little changes in the indicator coefficients prompts generally expansive changes in the post positions.

The existence of the poles and zeros in the unit hover in the z plane is imperator for the provision of a guarantee for the solidness of the channel coefficients. This is how the high exactness of 8-10 bits for every co-productive is required, the PARCOR coefficients steadiness is the bound of +1 or – 1,[9]

Frequency Components:

The obtained frequency components from the input speech signal are obtained as fundamental frequency usually referred as pitch or f0 and four different bandwidth of the speech samples termed as B_1, B_2, B_3 and B_4 respectively. 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. A formant frequency is the accumulation of acoustic energy in the particular frequency in the input speech wave. Every formant each at a different frequency with the bandwidth of roughly 1000Hz. Every formant corresponds to resonance in turn results in par-cor coefficients in the respective vocal tract. [10][11]

Formant bandwidths are represented by the distance of the prediction polynomial from zero circle. To measure the band width we have to find the peak response of the system and find what limits of frequency are at just 3 db lower than this. The speech signal bandwidth is defined as the width of the spectral band containing significant formant energy. Hence it is required to plot the spectrogram of the speech signal obtained for various analysis. [12]

Input speech signal:

Test samples considered here are listed below for the speaker subjects, subjects as uttered speech samples of Vowels: /a/,/e/,/i/,/o/ and /u/, are recorded using table top mic make of with sensitivity -58dB ±3dB, frequency response of 100Hz to 16kHz, with sampling frequency of 22,100Hz. Each vowel is recorded 30 different times and mean value are taken for analysis. [13][14]

Pitch Estimation:

SWIPE algorithm estimates the pitch of the speech signal as the fundamental frequency of the Saw Tooth Wave form, such that the spectrum of saw tooth wave best matches the input speech signal spectrum. SWIPE computes the pitch by computing the similarity between the square root of the saw tooth wave spectrum and the square root of the speech signal using a pitch dependent optimal window size SWIPE stands for Saw tooth Waveform Inspired Pitch Estimator. It works on the concept that our brain determines the pitch of a speech signal by comparison to have a similarity between our source signal or produced voice signal and the target signal whose pitch is to be determined.[15] [18]

January - February 2020
ISSN: 0193 - 4120 Page No. 2166 - 2171

4. Problem statement

The own speech database relating to speeches in Indian English language considered in this paper is combination of five different vowels i.e., /a/,/e/,/i/,/o/ & /u. The vowels under consideration are recorded by the available speakers of the same age group in normal conditions with a minimum noise in the environment. Different speech samples in the language were used for evaluation of basic acoustic features (formant frequencies and pitch more specifically)[16][17]

5. Results & Discussion

Five various vowels /a/,/e/,/i/,/o/ & /u/ are analyzed as shown in figure 3,4 and 5, which shows the spectrogram of the all the five vowels which are listed above. The spectrogram values are supported by the table 1, which shows the mean value of one of the speaker under consideration. Similarly for all the remaining 29 speakers values are also tabulated and analyzed individually. a and b

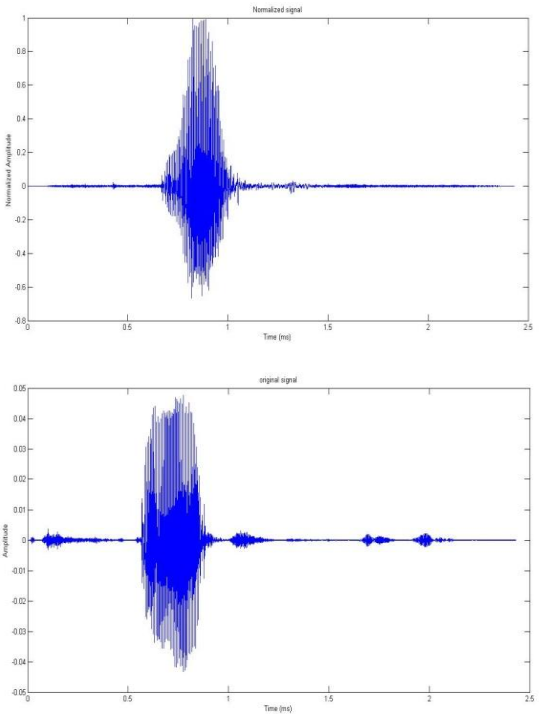
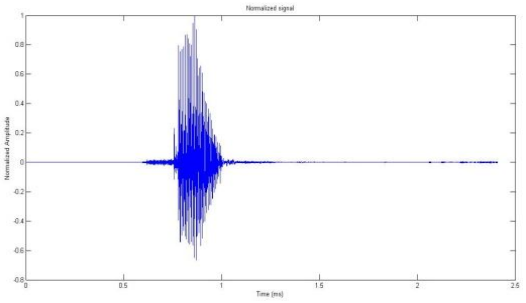


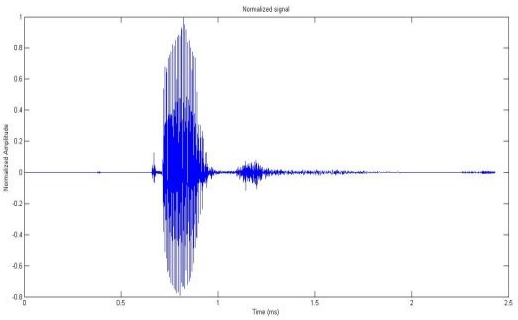
Figure 3: Spectrogram of (a) /a/ (b) /e/

Table 1: summarized values of one of the speaker under consideration

Speaker	B_1 (Hz)	B_2 (Hz)	B_3 (Hz)	B_4 (Hz)	FF (Hz)
/a/	83	87	114	144	216
/e/	97	101	118	157	211
/i/	93	92	118	143	183
/o/	100	83	106	144	202
/u/	81	111	120	168	203



a)



b)

Figure 4: Spectrogram of (a) /i/ (b) /o/

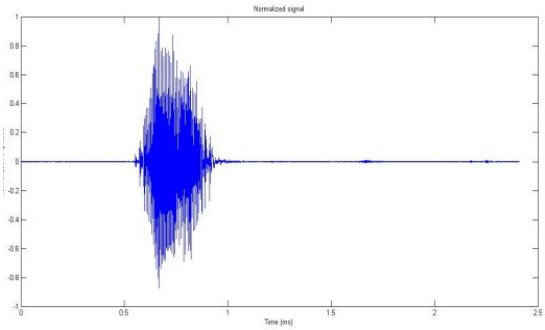


Figure 5: Spectrogram of /u/

The values in the table 1 depicts the mean values of the four different formant bandwidths and fundamental frequency of the single speaker, in the table it is observed that the mean bandwidth values of the first bandwidth B_1 is comparatively less as the initial formants are concentrated with the nearest values, B_2 the second bandwidth provides higher bandwidth than B_1 as the formant frequency shift occurs the spreading of the speaker information takes place as followed by the B_3 and fourth bandwidth is widely spread due to the region of the articulation also the higher formants ensure wide region of operation.

From the figure 6 it is representing the comparative analysis of the frequency components obtained from the autoregressive and LPC modelling by the analysis of the parcor coefficients. The important observation from the comparative graph it is analyzed that all the four bandwidth of formants obtained are less than the mean values of the fundamental frequency of the particular vowel.

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 2166 - 2171

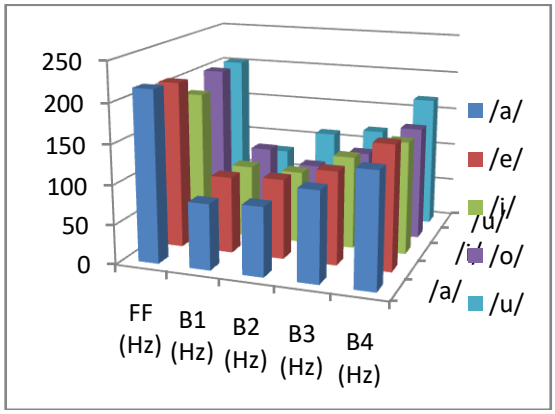


Figure 6: Comparative graph of vowels

6. Conclusion

By using autoregressive modelling and LPC for speech processing the bandwidth of the formants are estimated and fundamental frequency i.e., pitch is calculated using the SWIPE algorithm. The frequency components are extracted by taking PAR-COR coefficients from the input speech signal which is an untrained input database for analysis. The formant bandwidth for Indian English vowels (/a/,/e/,/i/,/o/,/u/) for individual speakers are done by taking thirty different speakers into consideration. The input speech signal considered is divided into frame length of 30msec with an overlapping of 10msec frame length by keeping the sampling rate of 22,100Hz by 663 samples in each frame. It is observed that the significant variation in the bandwidth of the vowels are noted depending on the three different categories namely, front vowel /a/, unrounded vowels /e/, /i/ and rounded vowels /o/, /u/. always the front vowel has the least value of bandwidth and fundamental frequency, unrounded vowels has slightly higher values and rounded vowels has the similar value of the front vowels, which can analyzed that bandwidth values depends on the tongue position during speech production which will less ensure the less amount of energy comparatively is accumulated in the oral cavity. The rounded vowels has the higher values of the frequency component due to the condition that air volume from the lung cavity can freely flows across the oral cavity. These observations will serve as the basic parameter setting for many user/speaker authentication and speech operated devices.

References:

[1] Praveen and Siva kumar, "Analysis of Resonant Peaks and Pitch for English Vowels of Diversified South Indian English Speakers", Jour of Adv Research in Dynamical & Control Systems, Vol. 11, 02-Special Issue, 2019 pp no 811-817.

[2] M. E. Ayadi, M. S.Kamel, and F. Karray, "Survey on speech recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, pp. 572–587.

[3] L.R. Rabner, "Applications of speech recognition in the area of Telecommunications", in proceedings of Automatic Speech Recognition and Understanding, pp.501-510, 1997

[4] Pukhraj P. Ahrishrimal, Ratnadeep R. Deshmukh, Vishal B. Waghmare, " Indian Language Speech Database: A Review", International Journal of Computer Applications, volume 47- No.5, pp. 17-21, June 2012.

[5] Qi Li and Yang Huang, "An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification Under Mismatched Conditions", IEEE Transactions on Audio, Speech and Language Processing, Vol. 19, No. 6, pp. 1791-1801, Aug-2011.

[6] R. Vergin and D. O'Shaughnessy, "Pre-Emphasis and Speech Recognition", pages 1062-1066,IEEE-CCECUCCGEI 95, August 1995

[7] Zoran S. Bojkovic, Bojan M. Bakmaz, Miodrag R. Bakmaz, "Hamming Window to the Digital World", pages 1185-1191, Proc of the IEEE, Vol. 105, No. 6, June 2017.

[8] Mehrdad khodai-joopaari, Frantz Clermont, Michael barlow, speaker variability on a continuum of spectral sub-bands from 297-speakers'non- contemporaneous cepstra of japans vowels, proceeding of the 10th Australian intl conference on speech scicence and technology, 2004.

[9] Gautam Vallabha & Betty Tuller, "CHOICE OF FILTER ORDER IN LPC ANALYSIS OF VOWELS", From Sound to Sense: June 11 – June 13, 2004 at MIT.

[10] K. Sajeer1* and Paul Rodrigues2 "Novel Approach of Implementing Speech Recognition using Neural Networks for Information Retrieval". Indian Journal of Science and Technology, Vol 8(33), December 2015.

[11] M. S. Shah and P. C. Pandey, "Estimation of vocal tract shape for VCV syllables for a speech training aid," in Proc. 27th Int. Conf. IEEE Engg. Med. Biol. Soc., 2005, pp. 6642–6645.

[12] Powen Ru, Taishih Chi, and Shihab Shamma, "The synergy between speech production and perception", J. Acoust. Soc. Am. , January 2003 J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, Vol. 63,pp 561-580,1975.

[13] Anil Kumar Chandrashekar., Manjunatha M.B., "Analysis of vocal tract shape variability based on format frequency ratio a various conditions", Volume 10, Issue 15, April 2017, Indian Journal of Science and Technology.

[14] Dong-Ill Kim1 and Byung-Cheol Kim "Speech Recognition using Hidden Markov Models in Embedded Platform" Indian Journal of Science

January - February 2020
ISSN: 0193 - 4120 Page No. 2166 - 2171

and Technology, Vol 8(34), DOI: 10.17485/ijst/2015/v8i34/85039, December 2015

[15] Anil Kumar C., Manjunatha M.B., Thangadurai.N., "Vocal Tract Shape Synthesis For Various Indian Speaker Under Various Conditions". Journal of Advanced Research in Dynamical and Control Systems, vol 10 9-special issue 2018

[16] R. Renita Rexy1, R. Rosita Rosini2 and G. Mani Sankar2, "A Novel Speech Recognition System using Hidden Markov Model", Indian Journal of Science and Technology, Vol 8(32), , November 2015

[17] Wankhede N.S, Shah M.S, "investigation on optimum parameters for LPC based vocal tract shape estimation", IEEE-C2SPCA, 2013, page 1-6

[18] Michael Scordillis and John N Gowdy, "Effects of the vocal tract shape on the spectral tilt of the glottal pulse wave form. ", IEEE Trans 1990 pp 86 – 89.