IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. IV (Jan.-Feb. 2017), PP 103-109
www.iosrjournals.org

Auto Regressive Moving Average Model Base Speech Synthesis for Phoneme Transitions

H.M.L.N.K Herath[1], J.V. Wijayakulasooriya[2]
[1]*Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka*
[2]*Department of Electronic and Electrical EngineeringFaculty of Engineering,University of Peradeniya, Peradeniya, Sri Lanka*

Abstract: *Speech synthesizers based on paramedic methods, still have not achieved the expected naturalness. This is due to less consideration on linear time variant nature between the neighbor phonemes. This paper presents a study to model the phoneme transitions between neighbor phonemes with lesser number of parameters using Auto Regressive Moving Average (ARMA) model, where Steiglitz-McBride algorithm is used to estimate the zeros and poles of the system. The results are compared with an Auto Regressive (AR) model, which show that the correlation between the source signal and the reconstructed signal in ARMA model is higher than that of the AR model.*

Keywords: *Auto Regressive (AR) model, Auto Regressive Moving Average (ARMA) model, Correlation Coefficient, phoneme transition, Speech synthesis,*

I. Introduction

Synthetic or artificial speech has been developed progressively during the last decades. In the present day, speech synthesizers are diagnosed with several limitations during synthetic speech production like speech naturalness and personality. However, intangibility has already reached a high level, which makes it possible to use synthesizer in certain applications. The Formant synthesis [1] and Concatenative synthesis [1] methods are the most commonly used in present synthesizers. The formant synthesis was dominant for a long time. But the concatenative method is becoming more and more popular at present as it provides high quality, more natural synthetic speech than other methods. But the main drawback of this method is it needs huge capacity to store the prerecorded speech units.

One of the recent approaches in speech synthesis is to find a way to represent speech sounds in lesser number of parameters while maintaining the naturalness. To represent the speech information in lesser number of parameters the most appropriate approach is to represent them using a combination of mathematical functions or parametric form. When it moves from prerecorded speech samples to parametric model the capacity to store the speech information gets reduced but naturalness of the synthetic speech tends to decreased. In addition to that the posody, style of speech, number of voices are some of the limitation that cannot be achieved in synthetic speech and which leads to the unnaturalness of the output. In most of the paramedic methods the discontinuity of phoneme boundaries is one reason which contributes to this unnaturalness. This discontinuity arises while connecting speech phonemes or segments to form words. In Formant synthesis and Concatenative synthesis models, speech segments or phonemes are synthesized separately and concatenated to form words, phrases and sentences. In this process the segments or phonemes do not mapped with each other at the boundaries, more often than not. PSOLA (Pitch Synchronous Overlap Add) method [1][2][3]is a way of reducing discontinuities arises in phoneme boundaries. This is mostly used in Concatenative speech synthesis as well as formant synthesis method.

The formant synthesis, which is also based on resonant behavior of vibrating structures, consists in letting the resonant behavior be parametrically modeled by means of resonant filters (all-pole or pole-zero) excited by a source signal. For short duration excitation signals and filters parameterized by a few coefficients, such a source-filter model implies a compact representation of sound sources. The problems involved in source-filter approaches can be roughly divided into two sub-problems: the estimation of the filter parameters and the choice or design of suitable excitation. As regards the filter parameter estimation, standard techniques for estimation of AR and ARMA processes can be used.

AR model, Linear predictive coding (LPC) is the stepping stone towards in formant synthesis. The LPC filter gives the synthetic speech, the desired spectral envelop, matching the formants without explicit formant identification. This is enough to create intelligible speech, but fails to produce natural sounding speech because of simplistic excitation model. However, the LPC synthesis fails to capture characteristics of a speaker such as user dependent speech parameters and control of the amplitude which is the main drawback of this method. It can be shown that the amplitude and the phase relationships of the first few harmonics contain crucial information on speaker identity [3]. Therefore modeling speech harmonics directly using a sinusoidal speech

representation seems to be a more appropriate approach towards meeting the transparency requirement. To improve the naturalness of the synthetic speech in linear time variant nature, it is tried to model the transition regions between neighbor phonemes. In this approach two standard techniques, which are AR and ARMA were used to estimate the filter parameters and model the transition regions using sinusoidal noise model. This is in contrast to the existing approaches, which try to model the phonemes, not the phoneme transitions using AR and ARMA models.

II. Methodology

2.1 Word Selection Criteria

In English language there are nearly about 44 phonemes. Those phonemes are classified in terms of vowels, consonants, diphthongs and semi-vowels. According to the articulatory configuration vowels are categories as front, mid, back vowels and consonant as nasals, stops, fricatives, whisper and affricates consonants . Among the vowel phonemes words which include short /a/ phoneme were considered for the study. It is infeasible to carry out the experiment for all those words, thus sample set of words were selected by considering the phoneme classification.

Table 1. Selected phoneme transition sounds and words

Starting Phoneme	Phoneme Category	Word List
B	Stops Voiced consonant	Bad, Bag, Ban, Bat Back, Band, Bank etc
T	Stops Unvoiced consonant	Tab, Tan, Tad, Tag, Tap, Tax, Tang, etc
S	Fricatives unvoiced consonant	Sam, Sat, Sag, Sad, Sap, Sand, Sang .etc
M	Nasals consonant	Man, Mat, Mag, Mad, Map,Mam etc
H	Whisper consonant	Ham, Has Had, Hat, Hag, Hack, Hang etc

Transition regions were detected by hearing voice components and they were segmented manually (Fig 1) The speaker of all of the utterances was a male speaker. 44100 Hz was selected as the sampling rate.

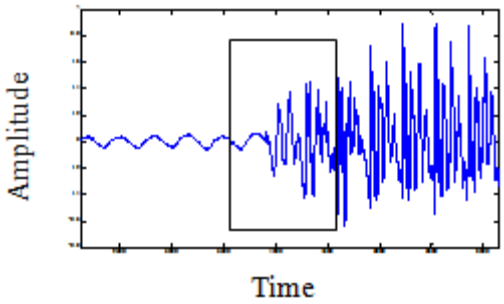

Fig 1: 'Ba' Transition Region

The amplitude, phase, frequency and exponential decay (speech parameters) values were estimated by considering the dominant poles of the ARMA model. The basic analysis process was explained in Fig 2.The most suitable filter coefficients of ARMA model (IIR filter) were estimated by comparing Pearson's Correlation values between the source and the synthesized signal by changing the number of filter coefficients in the algorithm.All the parameters were stored in a database.

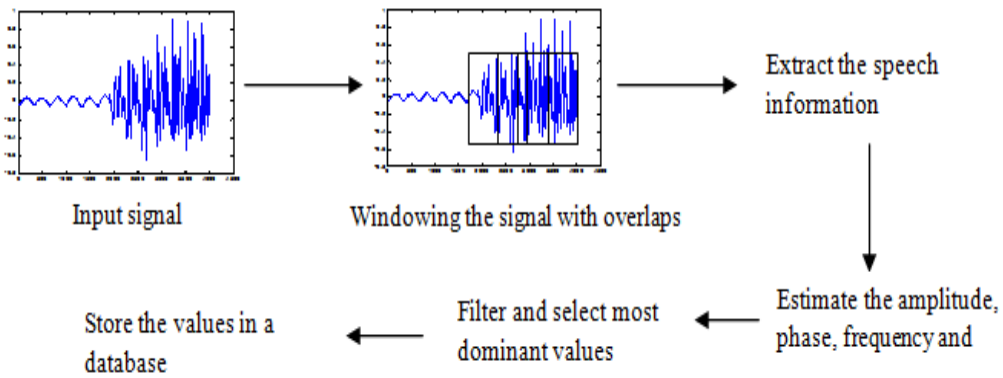

Fig 2. Basic Analysis Process

1.2 Estimating Speech Parameters

2.2.1 Auto Regressive Moving Average Model (Steiglitz-McBride Algorithm) and Auto Regressive model (Linear Predictive Coding Algorithm)

Speech parameters frequency, phase, amplitude and exponential decay derived according to the (1) given in AR model and (2) given in ARMA model.

$H(z) = \frac{1}{1-\sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)}$ ……(1)

$H(z) = \frac{\sum_{k=0}^{q} b_k\ z^{-k}}{1-\sum_{k=1}^{p} a_k\ z^{-k}}$……(2)

The partial fraction representation H(z) express as,

$H(z) = \frac{B(z)}{A(z)} = \frac{r_m}{s-p_m} + \frac{r_{m-1}}{s-p_{m-1}} + \cdots + \frac{r_0}{s-p_0}$+k(z)……(3)

Where, the values $r_m \ldots r_0$represents the residues, the values $p_m \ldots p_0$ are poles and *k(z)* is a polynomial in z, which is usually 0 or constant[44]. The real and imaginary parts of the complex transform of residues r_m are used to estimate the amplitude A_n and the phase ϕ_n

$A_n = |r_m|$.……(4)

$\phi_n = \tan^{-1}\left(\frac{r_{im_n}}{r_{Re_n}}\right)$……..(5)

Pole locations p_m used to calculate the frequency and attenuation coefficient r_n

$f_n = \tan^{-1}\left(\frac{p_{im_n}}{pl_{Re_n}}\right) \times ((Fs/2)/\pi)$……..(6)

$r_n = |p_m|$……(7)

Where, *fs* sampling frequency, *n* designate the frequency increment (*n*= 0, 1,…,N) and *Re* an *Im* are the real and the imaginary parts of the $r_m \ldots r_0$and $p_m \ldots p_0$ transform.

2.3 Signal Reconstruction

2.3.1 Sinusoidal Noise Modeling

The sinusoidal noise model is a parametric speech synthesis model, which is originally proposed for speech coding purposes and for the representation of musical signals. The sinusoidal model speech or music signal is represented as sum of sinusoids each with time-varying amplitude, frequency and phase. Sinusoidal modeling works quite well for perfectly periodic signals, but performance degrades in practice since speech is rarely periodic during phoneme transitions. In addition, very little periodic source information is generally found at high frequencies, where the signal is significantly noisier. To address this issue the sinusoidal model was improved as a residual noise model that models the non-sinusoidal part of the signal as a time-varying noise source. These systems are called sinusoids plus noise systems.

Sounds that are produced by auditory systems can be modeled as sum of the deterministic and the stochastic parts, or as a set of sinusoids plus the noise residual [2]. In the standard sinusoidal noise model, the deterministic part is represented as a sum of sinusoidal trajectories with time varying parameters. The trajectory is a sinusoidal component with time-varying frequencies, amplitudes and phases. It appears in a time-frequency spectrogram as a trajectory. The stochastic part is represented by the residual [4].

$x(t) = \sum_{i=0}^{N} A_i\ (t) \cos\left(\theta_i(t)\right) + r(t)$…….. (8)

where, $A_i(t)$and $\theta_i(t)$ areamplitude and phase of sinusoidal *i*at time *t*, and *r*(*t*) is a noise residual, which is represented with a stochastic model. Further it can be represent as,

$x(t) = \sum_{i=0}^{N} A_i\ (t) \cos\left(\omega_i t + \phi_i\right) + r(t)$…… (9)

where, $A_i$denotes the amplitude, ω_i is the frequency in radians/s (radian frequency), $\phi_i$and is the phasein radians of sinusoidal *i* at time *t*. The radian frequency ω_i denote as $2\pi f_i$ and the equation can be written as,

$x(t) = \sum_{i=0}^{N} A_i\ (t) \cos\left(2\pi f_i t + \phi_i\right) + r(t)$…… (10)

where, f_i is the oscillation frequency in i[th] sinusoidal component.

$x(t) = \sum_{i=0}^{N} A_i\ (t)\ e^{-\alpha t} \cos\left(2\pi f_i t + \phi_i\right) + r(t)$……..(11)

(11) represents a decaying sinusoidal. Where, α is the exponential Decay and $e^{-\alpha t}$ is the decay rate.

Since the sinusoidal noise model has the ability to remove irrelevant data and encode signals with lower bit rate, it has also been successfully used in audio and speech coding. The most of the available models based on the sinusoidal model are capable of synthesizing vowels and the phonemes in high quality.

Signals were reconstructed based on the data extracted from the basic analysis model. With the help of calculated parameters, the sinusoid is generated (Fig 3). White Gaussian noise was applied to generate the noise residuals using mean and standard deviation of the noise.

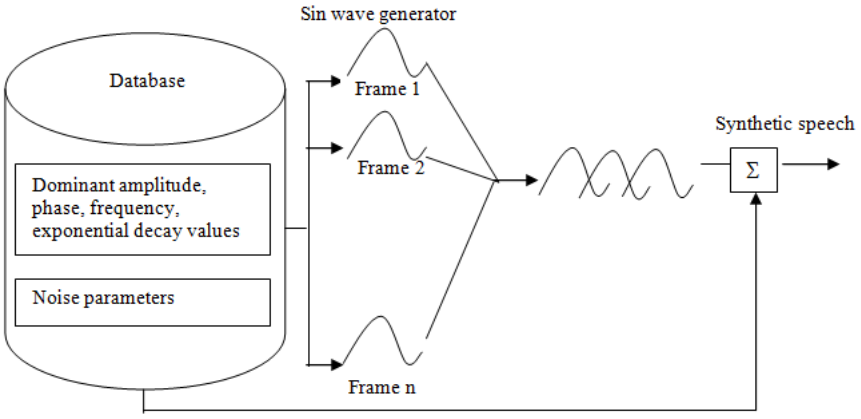

Fig.3. Proposed System

The experiment was carried out changing the number of dominate values from 1 to 5. Same experiment was repeated by changing the frame size and the size of the overlap. The Pearson's Correlation Coefficient between source signal and the synthesized signal were calculated. Next the required capacity to store the source waveform and the proposed method, speech parameters were compared by calculating the capacity ratio. The experiment was repeated by replacing the ARMA data extraction method by AR model.

III. Results And Discussion

Fig 4 shows how the Pearson's Correlation Coefficient changes with the capacity ratio. There the capacity ratio was calculated by considering the number of dominant values selected to reconstruct the original signal. (e.g. f1, the number indicates the number of dominant values selected). When the capacity ratio was increased, the Pearson's Correlation Coefficient values were also increased gradually. Highest Pearson's Correlation Coefficient value was found in the highest capacity ratio, for all the phoneme transitions, it was 17.6% from the actual capacity. All the correlation values observed were greater than 0.75. According to the graph a clear cut off point at the capacity ratio 11% can be observed. That is occurred after third point (*p3,b3,v3,f3,m3*) the increment of the Pearson's Correlation Coefficient is very small even more points were added. For an example the Pearson's Correlation Coefficient of the *f4* has no clear significant improvement compared to the *f3*. This is true for all phoneme transitions. By considering the correlation coefficient value and the sound of the reconstructed wave, third point can be selected as the cutoff point.

Then the same procedure was carried out by changing the window size. The window sizes which was selected were 300 and 400. It also shows the same pattern of the correlation coefficient with the capacity ratio (Fig 5). All the observed Pearson's Correlation values were higher than 0.65, but less than the values observed in window size 300.

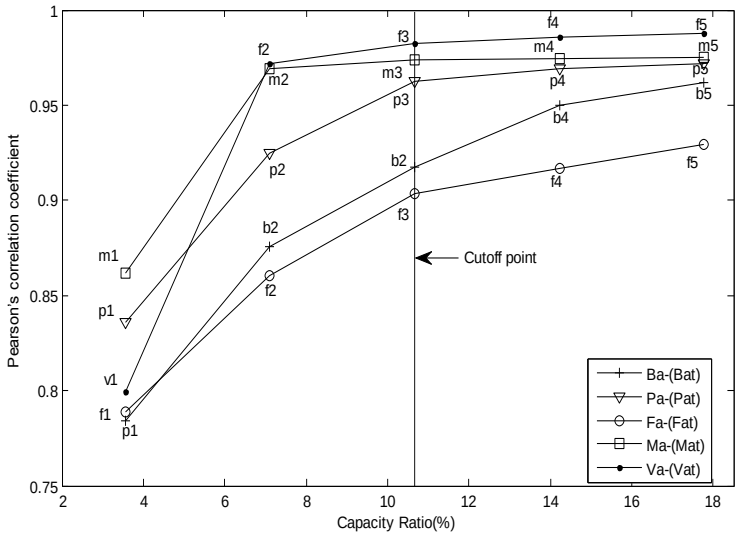

Fig. 4. Pearson's correlation coefficient changes with Capacity ratio in different number of dominant values with frame size 300 (b1-Number indicates the number of dominant values selected)

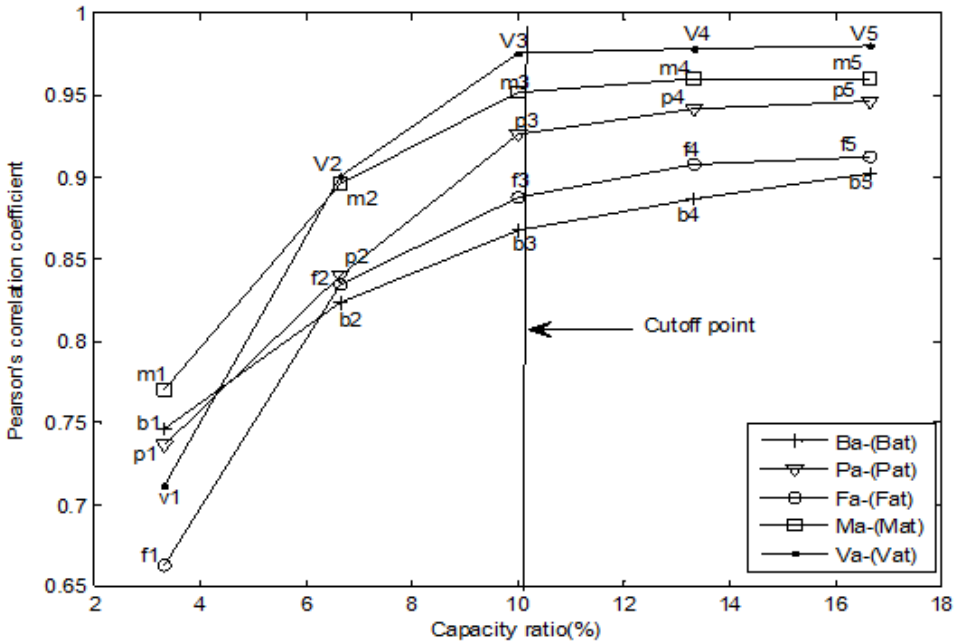

Fig.5. Pearson's correlation coefficient changes with Capacity ratio in different number of dominant values with frame size 400 (b1-Number indicates the number of dominant values selected)

Experiment was repeated using several other words selected from each phoneme category. Patten of change of the average correlation coefficient with capacity ratio shown in Fig 6 and Fig 7 were similar to the pattern in Fig 4 and Fig 5. When the number of selected points exceeds 3, the correlation values increase in small amount. In addition to that the variability of the standard deviation also minimum compared to the other methods. So Fig 6 and Fig 7 also prove that S3 can be selected as the cut-off point.

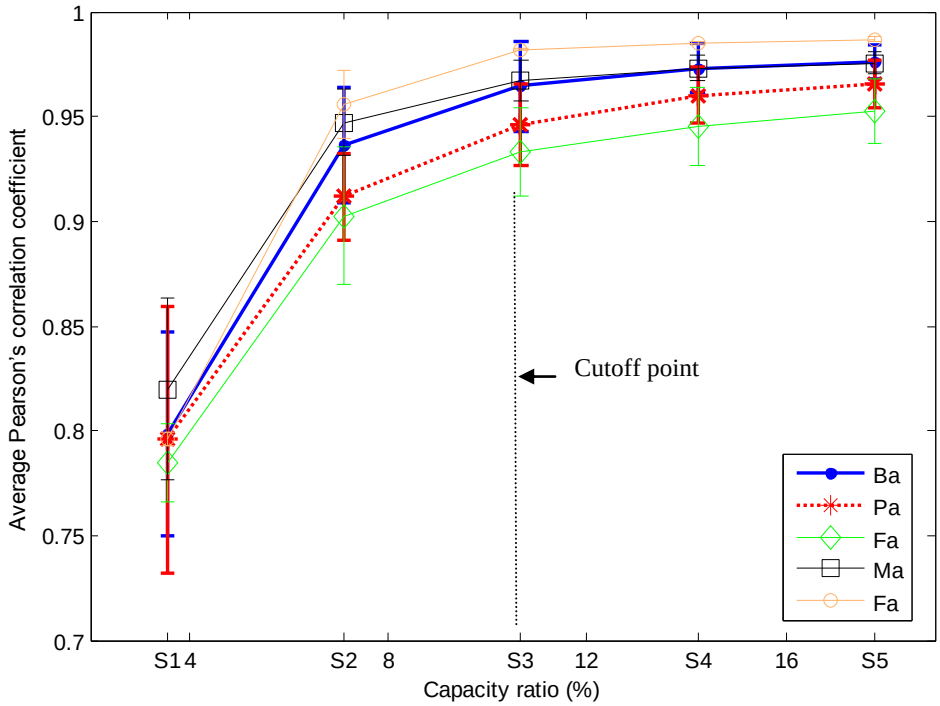

Fig.6. Average Pearson's correlation coefficient changes with Capacity ratio in different number of dominant values with frame size 300(S1 -Number indicates the number of dominant values selected)

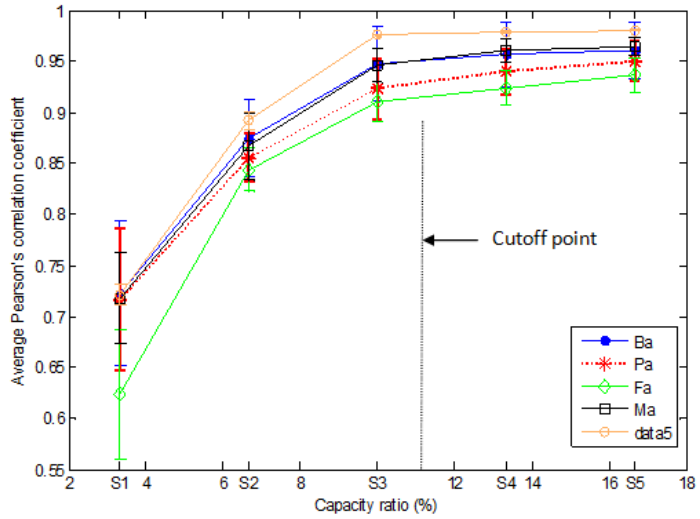

Fig. 7. Average Pearson's correlation coefficient changes with Capacity ratio in different number of dominant values with frame size 400(S1 -Number indicates the number of dominant values selected)

In Fig 8, it is clearly shown that how the Pearson's Correlation Coefficients were changed with the capacity ratio in both AR model (LPC algorithm) and the ARMA model (Steiglitz-McBride algorithm). The graph clearly indicates that the ARMA model (Steiglitz-McBride algorithm) provides better results compared to the AR model (LPC algorithm). All the correlation values obtained in the ARMA model (Steiglitz-McBride algorithm) were grater then 0.8 but in AR model (LPC algorithm) all the values were between 0.3 and 0.85.

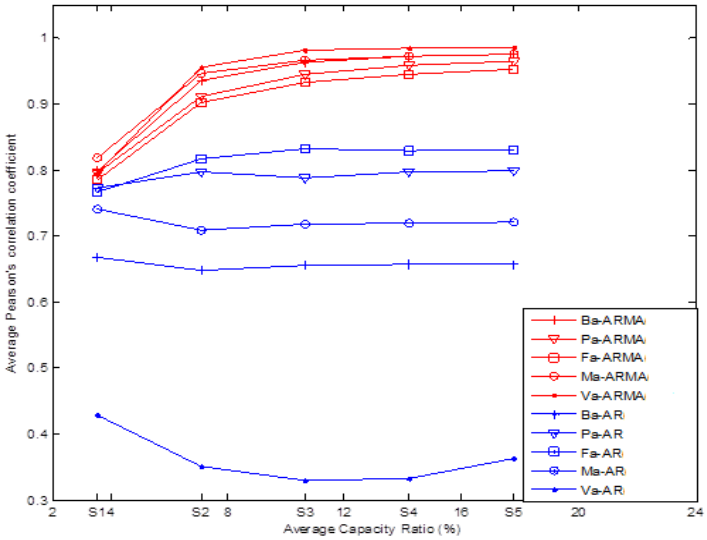

Fig.8. Average Pearson's correlation coefficient changes with Average Capacity ratio in AR model (LPC Algorithm) and ARMA model (Steiglitz-McBride Algorithm). (S1 -Number indicates the number of dominant values selected)

IV. Conclusion

This paper has discussed two data extraction methods that can be used to extract the model dominant speech information between consecutive phonemes. The proposed method is capable of synthesizing transition region based on the sinusoidal noise model with lesser number of parameters. Speech parameters were extracted using AR model (LPC algorithm), the observed correlation coefficient values conclude that the constructed signal was moderately correlated with the source signal. Significant improvements cannot be observed by increasing the number of dominate LPC poles. In contrast the signals constructed by the ARMA model was highly correlated with the source signal. When the sound of the output signal was compared, the ARMA gives a better quality output than the AR method. This study conclude that the ARMA model extract the most dominant features of the transition regions in less number of parameters than AR model, while the synthesized output is almost identical to the source signal.

References

[1]. P.Taylor, *Text to Speech Synthesis*, Cambridge University Press, 2009.
[2]. J. Holmes, W. Holmes, *Speech Synthesis and Recognition*, Second Edition, Taylor & Francis, 2001.
[3]. J.Benesty, M. M. Sondhi, Y. Huang, *Springer Handbook of Speech Processing.* Springer.2008.
[4]. L. Rabiner, B. Juang, *Fundamentals of speech Recognition*, Prentice Hall International, 1993
[5]. J. K. Sharma, Business statistics. Pearson Education India. 2007.
[6]. M. Tatham, K. Morton, **,***Development in Speech Synthesis*, John Wiley & Sons Ltd, 2005.
[7]. A. O'Cinneide, D. Dorran, M. Gainza, "Linear Prediction: The Problem, its Solution and Application to Speech*", DIT Internal Technical Report*,2008.
[8]. T. Phung, M. C. Luong, M. Akagi, "An Investigation on Perceptual Line Spectral Frequency (PLP-LSF) Target Stability against the Vowel Neutralization Phenomenon*", 3rd International Conference on Signal Acquisition and Processing (ICSAP 2011)*: 2011,pp 512-514
[9]. T. Phung, M. C. Luong, M. Akagi, "On the Stability of Spectral Targets under Effects of Coarticulation*", International Journal of Computer and Electrical Engineering*, Vol. 4, No. 4, 2012 pg 537-541.
[10]. M. Shannon , H. Zen, W. Byrne,"Autoregressive Models for Statistical Parametric Speech Synthesis*", IEEE transactions on audio, speech, and language processing*, vol. 21 (3); 2013 pg 587-597
[11]. M. Wang , "Speech Analysis And Synthesis Based On ARMA Lattice Model", *Master's Thesis*, University Of Windor, 2003