Modeling of Phoneme Transitions for Natural Synthesis of Speech Signals

H. M. L. N. K. Herath
Postgraduate Institute of Science,
University of Peradeniya,
Peradeniya, Sri Lanka

J. V. Wijayakulasooriya
Department of Electronic and Electrical Engineering
Faculty of Engineering,
University of Peradeniya,

ABSTRACT
Natural synthesis of speech needs to identify the minute variations in phoneme during reproduction, which is affected by many factors. One well-known problem with speech synthesis is the occurrence of audible discontinuities at phoneme boundaries, which lead to the unnaturalness of synthetic speech. This study basically focuses on introducing a novel method with low bit rate to improve the naturalness of synthetic speech.

The research presents a sinusoidal noise based mathematical method to reform the transition regions from one phoneme to another phoneme with lesser number of parameters. The speech information which are amplitude, phase and frequency were extracted using three different algorithms. They are Fast Fourier Transform (FFT) algorithm, Auto Regressive model (AR) with Linear Predictive Coding (LPC) algorithm and Auto Regressive Moving Average model (ARMA) with Steiglitz-McBride method. Polynomial coefficients were estimated to represent the speech information in lesser number of parameters. The results show that the synthesized output is highly correlated to the source signal in FFT method than AR model and ARMA model.

Keywords
Fast Fourier Transform algorithm (FFT), Auto Regressive model (AR), Auto Regressive Moving Average model. (ARMA), Speech Synthesis, Correlation Coefficient, Phoneme Transition

1. INTRODUCTION
The current ambition in speech synthesis research is to improve the quality of the synthesized speech by naturalizing speech on a global level, allowing changes of speaker characteristics by a great amount of flexibility, speaking style and number of voices. Instead of monotonous, incoherent and mechanical sounding speech utterances, these systems produce output that sound relatively close to human speech. Naturalness is described as how much the synthetic voice is similar to the human voice. Thus, the synthesis of natural speech signals which are similar to the human voice is one of the main problems lies today in the speech synthesis itself.

However, some systems produce better synthetic sound today, so that few listeners believed that they hear actual human speakers. But even the simulation is very good; still it is not perfect as human speech. So it is really necessary that the quality needs to be high and the voice need to sound like a human. Most research have shown that people are in fact very sensitive, not just to the words that are spoken, but also to the way they are spoken. Most of the time people who listened to the highly mechanical voices in short time felt irritated and discomforting to listen to it. Furthermore some experiments have shown that user satisfaction increases dramatically when the voice is more "natural". Some particular commercial experiences show that users clearly want natural sounding systems. Hence the main goal of speech synthesis research at present is to improve the intangibility and naturalness of the synthetic speech. In addition to this, most synthesizers currently manipulate a small number of parameters in a highly constrained manner to produce speech and thus it lacks flexibility. Massive research and financial investments were made to improve the naturalness of synthetic speech. But still the holy grail of "true naturalness" in synthetic speech seems so near and not yet so elusive

In this regard, low and high-level synthesis methods have been proposed [1]. Articulator synthesis, Concatenation synthesis (wavetable synthesis in digital music) and mathematical model (Parametric speech synthesis) based speech synthesis are the most wildly used computer-based speech synthesis techniques. Articulatory synthesizers, [2][3] determine the characteristics of the vocal tract filter by means of a description of the vocal tract geometry and place the potential sound sources within this geometry. The advantage of these higher-dimensional models is that the form and position of the articulators can be specified in a very direct fashion. It potentially has the best chance at near perfect synthesis because it takes in the possibility of a complete model of human speech production.

In the concatenation synthesis [4][5] procedure, the optimum set of raw waveform segments (units) corresponding to each phoneme is stored in a database. The synthetic speech is generated by concatenating the selected waveform segments. Waveform segments either by phonemes, diaphones or syllables. The naturalness of the concatenation synthesis systems depends greatly on the database recording. For grate naturalness and intelligibility transitions between waveform segment is important. However waveform segments cannot avoid the transition between units, which often produce auditory discontinuity. It is one issue that is faced in concatenation synthesis and which is the main concern in this research. In addition to that, designed voices for particular application may often sound inappropriate for another application. For an example, voice built from news reader's speech used in automated mobile service provider's system may make the user think they have been interviewed on CNN rather than getting assistance for recharging their mobile accounts credit. Selecting different voices for different applications are time and space consuming with this method. Although it produces more natural speech than the mathematical coding based models, the high capacity needed for storing the speech, [Table 1] and high bit rates involved in transmission of the speech is a main concern [1].

Table 1: Comparison between speech synthesis models [1]

Synthesis type	Bit rate	Database size	Naturalness (segments)	
			Isolated	Coarticulation
Articulatory	Mid	Mid	Good	Mid
Concatenative	High	Large	Good	Good
Mathematical (Parametric)	Low	Small	Mid	Poor

In contrast, the mathematical coding based techniques such as Formant Synthesis, Hidden Markov Model (HMM) speech Synthesis, Auto Regressive (AR) [6][7] model based Linear Predictive Coding (LPC), Auto Regressive Moving Average (ARMA)[6] models are used for speech synthesis in various ways. All of the above mentioned models do not use human speech samples at runtime. Instead, the speech output is created using mathematical parameters like fundamental frequencies (vocal source), duration (prosody), noise level, etc… which are extracted by the human speech samples. Many systems based on mathematical models generate artificial, robotic sounds due to poor co-articulation [Table 1]. However, in all these models, the speech is modeled as a response of a Linear Time Invariant (LTI) system to an input excitation signal for each phoneme. The speech syntheses by this model regrettably remain highly unnatural. This is due each phoneme generated separately and concatenated. It assumes that each phoneme is independent from the neighbor phonemes. As a solution, diaphones which dissects each phoneme at the midpoint is used. However, the speech signals produced by this method also does not take the phoneme transition patterns into account.

The main objective of this research is to develop a mathematical method to improve the naturalness of synthetic speech by modeling the co-articulation between the phonemes in lesser number of parameters.

2. METHODOLOGY

The main problem in modeling waveform transitions from one phoneme to another is finding parameters from a speech waveform that represents a quasi-stationary portion of that waveform. And then to use those to reconstruct an approximation that is closer to the original speech.

2.1 Speech Signal Analysis

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 for demonstration purposes, constant to vowel transitions of short /a/ sound words, such as ba (bat, bad, ban, etc) ta (tab, tan, tad, etc), sa(sam, sat, sag etc.), ma(man, mat, mag etc) and ha(hat, ham ,has etc) were considered.

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

Peak points of frequency curve were extracted manually from each segment of the sound wave. The speech signals were analyzed by applying three different data extraction methods.

1. Fast Fourier Transform algorithm (FFT)
2. Auto Regressive model (AR)
3. Auto Regressive Moving Average model. (ARMA)

In the analyzing process, speech parameters such as amplitude, phase and frequency of the speech signal were estimated. The speech parameters were estimated by considering the dominant frequency components of the Fast Fourier Transform (FFT) and dominant poles of the AR and ARMA models.

In AR model the coefficient of linear predictor (FIR filter) was estimated by applying the general equation

$$ncoeff = 2 + Fs / 1000 \quad (1)$$

where, Fs is sampling frequency.

This was applied to the AR model (LPC - method 1) to estimate the amplitude, phase and frequency components. For AR model (LPC – method 2), FIR filter coefficients and the coefficients of IIR filter in ARMA model were found by comparing Pearson's Correlation values between the source and the synthesized signal by changing the number of filter coefficients in the algorithm. The basic analysis process explained in Figure 1 was carried out by changing the data extraction method.

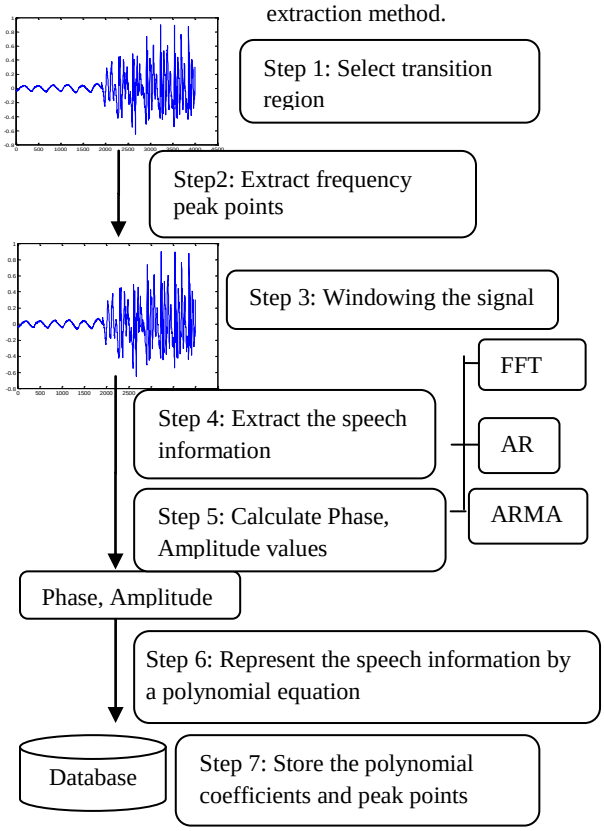

Figure 1: The Basic Analysis Process

Required capacity to store the source waveform and the proposed method, values for speech parameters were compared by calculating the capacity ratio. For that the polynomial curve fitting algorithm was applied to find the best fit curve to represents the values of speech parameters. The polynomial coefficients of the curves were stored in a database for synthesis the speech signal.

Figure 2. has summarized the data extraction methods that were used in the experiment. But for some experiments, apart

from the basic procedure, some additional changes have been done amended analysis process in Figure 3.

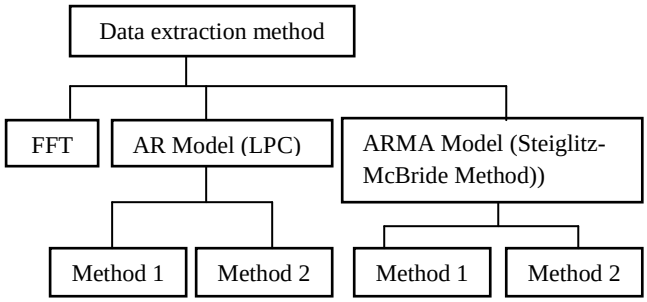

Figure 2: Data Extraction Methods

2.1.1 Fast Fourier Transform (FFT)

Speech information (amplitude, phase) was extracted by applying the FFT algorithm to the source signal by following steps 1 to 7 in basic analysis model (Figure 1). Signal was reconstructed using signal reconstruction system. To reconstruct the signal polynomial coefficients of phase and amplitude was used. Fundamental frequency value and its harmonics were used as the frequency values. The experiment was carried out changing the number dominate frequency components from 5 to 20.

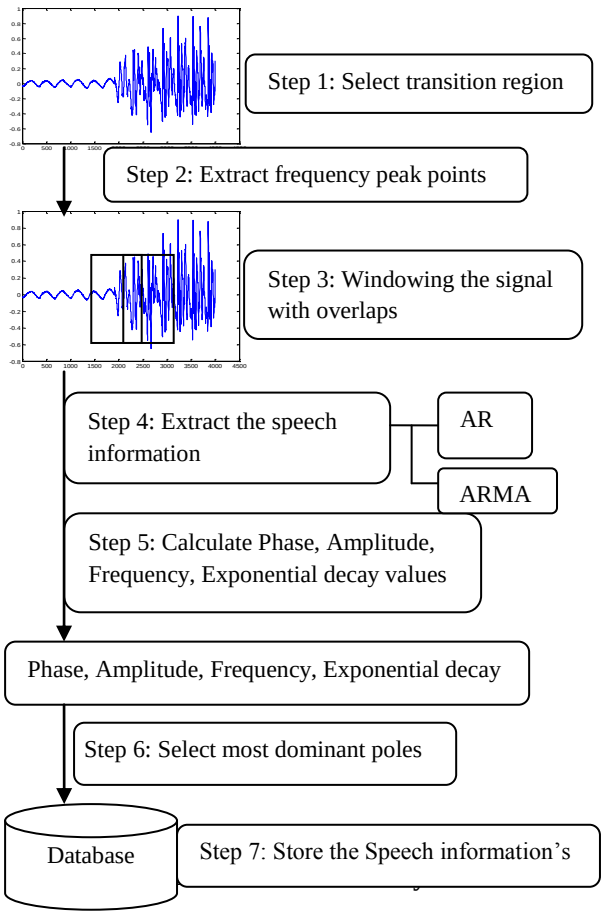

Figure 3: Amended Basic Analysis Model

2.1.2 Auto Regressive Model –(LPC Method)

Instead of FFT algorithm, LPC algorithm was applied to the source signal to extract the speech information as the LPC (method 1).

As the second method (method 2) of LPC algorithm, the basic analysis model was improved as shown in Figure 3. After that the signal was reconstructed by storing the source frequency, phase and amplitude values, instead of calculating polynomial coefficients. The experiment was carried out changing the number dominate component poles from 1 to 5. Same experiment was repeated by changing the frame size (200,300,400,500) and the size of the overlap (100,200,300).

2.1.3 Auto Regressive Moving Average Model – (Steiglitz-McBride Method)

Data extraction method was replaced by Steiglitz-McBride Method and Same procedure was repeated.(Method 1)

In Method 2 of ARMA, the signal was analyzed using the amended basic analysis and instead of storing the original frequency, phase and amplitude values (Figure 3- step 6), polynomial coefficients were used to represent the original frequency, phase and amplitude values(Figure 3 – step 6 and step 7). The signal was reconstructed by recalculating the frequency, phase and amplitude values

2.2 Estimating Speech Parameters

2.2.1 Fast Fourier Transform (FFT)

Amplitude and phase values were estimated by considering the equation (2). FFT algorithm return the Discrete Fourier Transform (DFT) of the vector x which is a complex transform. In this case real and imaginary parts of the complex transform are used to estimate the amplitude A_n and the phase $\emptyset_n$

$$x(j) = \frac{1}{N}\sum_{K=1}^{N} x(k) e^{\frac{(-2\pi i)}{N}(j-1)(k-1)} \quad (2)$$

$$A_n = |x(k)| \quad (3)$$

$$\emptyset_n = \tan^{-1}\left(\frac{x(k)Im_n}{x(k)Re_n}\right) \quad (4)$$

Where, x(k) is a vector of n values at frequency index k corresponding to the magnitude of the sine waves resulting from the decomposition of the time indexed signal. *Re* and *Im* are the real and the imaginary parts of the transform.

2.2.2 Auto Regressive Model –(LPC Method) and Auto Regressive Moving Average Model – (Steiglitz-McBride Method)

Speech parameters frequency, phase amplitude derived according to the equation (5) AR model and equation (6) ARMA model

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1-\sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \quad (5)$$

$$H(z) = \frac{y(z)}{x(z)} = \frac{b_0+b_1 z^1+\cdots+b_q z^{-q}}{1-a_1 z^1-\cdots-a_p z^p} \quad (6)$$

the partial fraction representation H(z) express as,

$$H(z) = \frac{B(z)}{A(z)} = \frac{r_m}{s-p_m} + \frac{r_{m-1}}{s-p_{m-1}} + \cdots + \frac{r_0}{s-p_0} + k(z) \quad (7)$$

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[8]. The real and imaginary parts of the complex transform of residues r_m are used to estimate the amplitude A_n and the phase $\emptyset_n$

$$A_n = |r_m| \quad (8)$$

$$\emptyset_n = \tan^{-1}\left(\frac{r_{im_n}}{r_{Re_n}}\right) \quad (9)$$

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

$$f_n = \tan^{-1}\left(\frac{p_{im_n}}{p_{Re_n}}\right) \times ((Fs/2)/\pi) \quad (10)$$

$$r_n = |p_m| \quad (11)$$

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.

Variation of estimated speech parameters (phase, amplitude and frequency of i^{th} sinusoidal component) in each time window were represented using polynomial equations.

$$A_i = c_m X^m + c_{m-1} X^{m-1} + \cdots + c_1 X + c_0 \quad (12)$$

$$\emptyset_i = d_m X^m + d_{m-1} X^{m-1} + \cdots + d_1 X + d_0 \quad (13)$$

$$f_i = e_m X^m + e_{m-1} X^{m-1} + \cdots + e_1 X + e_0 \quad (14)$$

Where, c_m, d_m, e_m are polynomial coefficients.

To capture the most important features from the speech signals, it needs to extract the most dominant poles gained from AR and ARMA model. First the residuals were converted to frequency, phase amplitude and exponential decay values. Then the non-negative frequency values and exponential decay (attenuation coefficient) greater than 0.95 and the corresponding phase, amplitude values were found. In order to find the most dominant values, speech information were sorted considering amplitude and frequency values.

2.3 Signal Reconstruction

In the speech signal synthesis, speech signals were re-synthesized using a sinusoidal noise model [9]. White Gaussian noise was applied to generate the noise residuals using mean and standard deviation of the noise.

All the speech parameters were recalculated using the stored polynomial coefficients. Thereafter Pearson's Correlation Coefficient values between source signal and the synthesized signal were calculated. Also the capacity ratio between the source signal and the parameter of the proposed method was evaluated.

In addition to that for AR and ARMA models, instead of calculating the polynomial coefficients, the original speech information were stored to reconstruct the speech signal. Then for the same information, Pearson's Correlation Coefficient and capacity ratio between source signal and the synthesized signal were calculated.

3. RESULTS AND DISCUSSION

The corresponding polynomial coefficients of amplitude and phase that are calculated from FFT were retrieved. Then amplitude, phase values were recalculated using proposed methodology to model the synthesize speech signal. The average correlation coefficients with the number of dominant frequency components of FFT changes from 5 to 20 of 'Ba' transition were illustrated in Figure 4. As a summary, all phoneme transition sounds have average correlation coefficients greater than 0.85 [13].

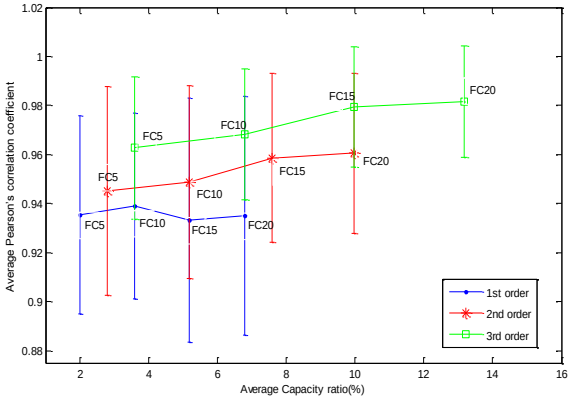

Figure 4: Average Pearson's correlation coefficient change with number of dominant frequency component poles of FFT in different polynomial orders for 'Ba' phoneme transition

The average correlation coefficient values of the AR model (method 1) of 'Ba' sound were greater than 0.65 as illustrated in Figure 5. Considering the average correlation coefficient values of all the phoneme transition regions, highest average correlation values were observed in signals that were constructed by calculating the speech parameters using 1^{st} order polynomial coefficients. This was true for all the phoneme transitions. When number of selected dominant poles was changed from 5 to 20, the average correlation value was changed within 0.05 [15].

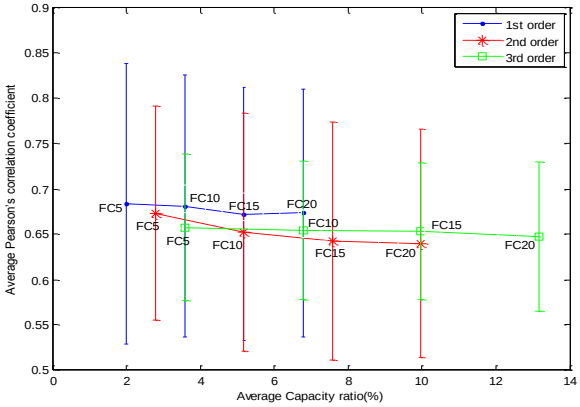

Figure 5: Average Pearson's correlation coefficient change with number of dominant frequency component poles of LPC in different polynomial orders for 'Ba' phoneme transition

While using AR Model (method 2) and ARMA (method 1) a best result was found when the window size is 300 for all the transition regions as in Figure 6 [16].

Figure 7 illustrates an overall summary of the results obtained in the experiment for 'Ba' phoneme transition. The results show that the source wave was highly correlated with the

reconstructed signal by FFT data extraction method (5 dominant frequency components) and Steiglitz-McBride Method (Method 1). All the observed values in both methods were greater than 0.75. Correlation values observed in the Method 1 and the Method 2 of the AR model (LPC method) were less than 0.75. The capacity ratios of the both methods were not overlapped with each other for same wave portion. This was because in the FFT method values were represented as polynomials and in the ARMA model (Steiglitz-McBride Method-method 1), the original data points were considered and the overlapped windows were selected.

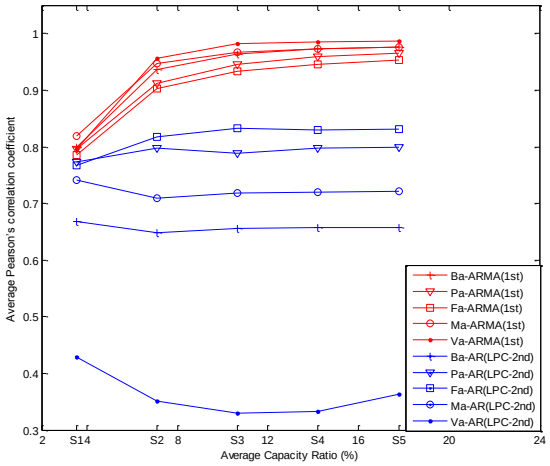

Figure 6: Average Pearson's correlation coefficient changes with Average Capacity ratio in AR model (LPC - 2nd method) and ARMA model (Steiglitz-McBride Method - 1st method). (S1 -Number indicates the number of points selected)

Figure 8, Figure 9, Figure 10 and Figure 11 show the overall summary for other phoneme transitions. For all of the phoneme transitions the observed correlations values were greater than 0.75 for the FFT method and the ARMA model (Steiglitz-McBride Method –method 1). All the phoneme transitions have similar result patterns for all methods. In all transition sounds, for ARMA model (Steiglitz-McBride Method - Method 2) and AR model (LPC –method 1) lower correlation values were observed.

Comparatively better results were derived by FFT data extraction method, when sound quality, capacity ratio and correlation values were taken into account. Among the other methods, ARMA Model (Steiglitz-McBride method – method 1) was proved to be better than the AR model (LPC method) and ARMA model (Steiglitz-McBride method – method 2). The signal quality was higher in the 3rd point of ARMA model (Steiglitz-McBride method – method 1). The best sound quality output for FFT was also obtained in the 3rd order. But the capacity of the 3rd order was same as the original number of points extracted by the source signal, thus it is unworthy of using a polynomial equation. In ARMA (Steiglitz-McBride method – method 1) needs more capacity than the FFT data extraction method and the capacity ratios were 5% in FFT method and 12% for ARMA model (Steiglitz-McBride method – method 1). In FFT method the original source information were stored as polynomial coefficients and then speech information were recalculated. But in ARMA (Steiglitz-McBride method- method 1) the original source data were used. So to reconstruct the signal, FFT needs more computational resources than the ARMA Model (Steiglitz-McBride method – method 1). But the FFT method reconstructs the speech transitions in high quality with less number of points. Same pattern was observed for all other transition sounds.

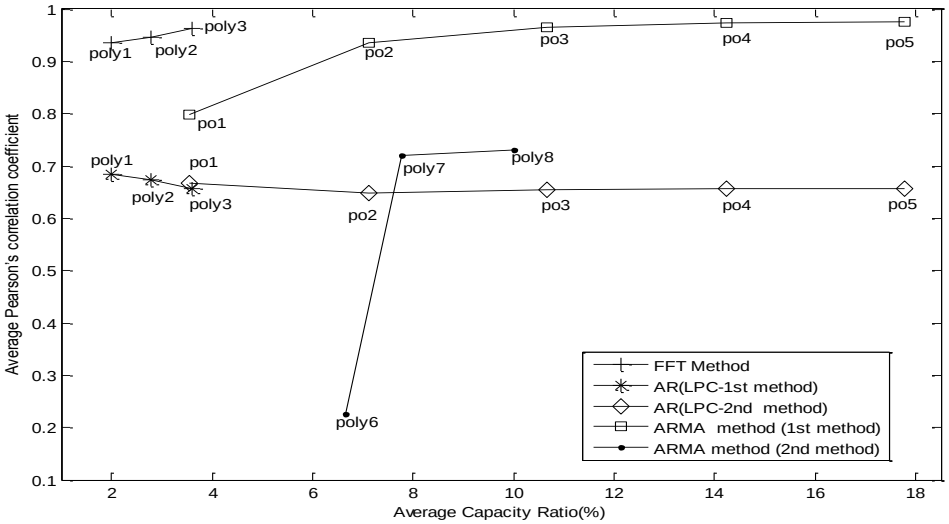

Figure 7: Pearson's correlation coefficient changes with Capacity ratio in all methods for 'Ba' phoneme transition (po1- Number indicates the number of points selected , poly1-Number indicates the order of the polynomial)

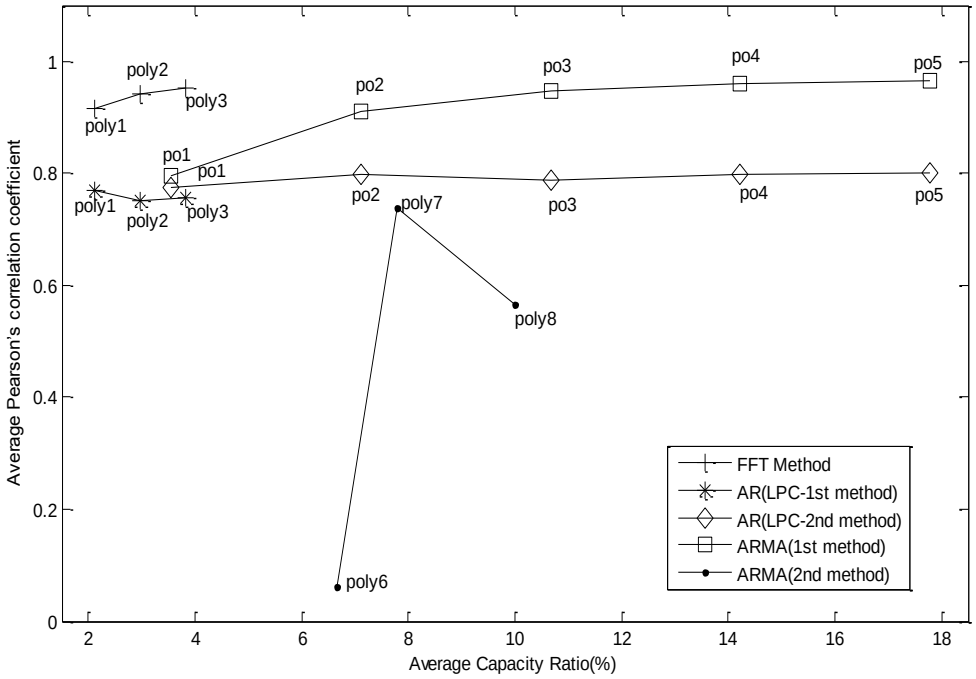

Figure 8: Pearson's correlation coefficient changes with Capacity ratio in all methods for 'Pa' phoneme transition (po1-Number indicates the number of points selected , poly1-Number indicates the order of the polynomial)

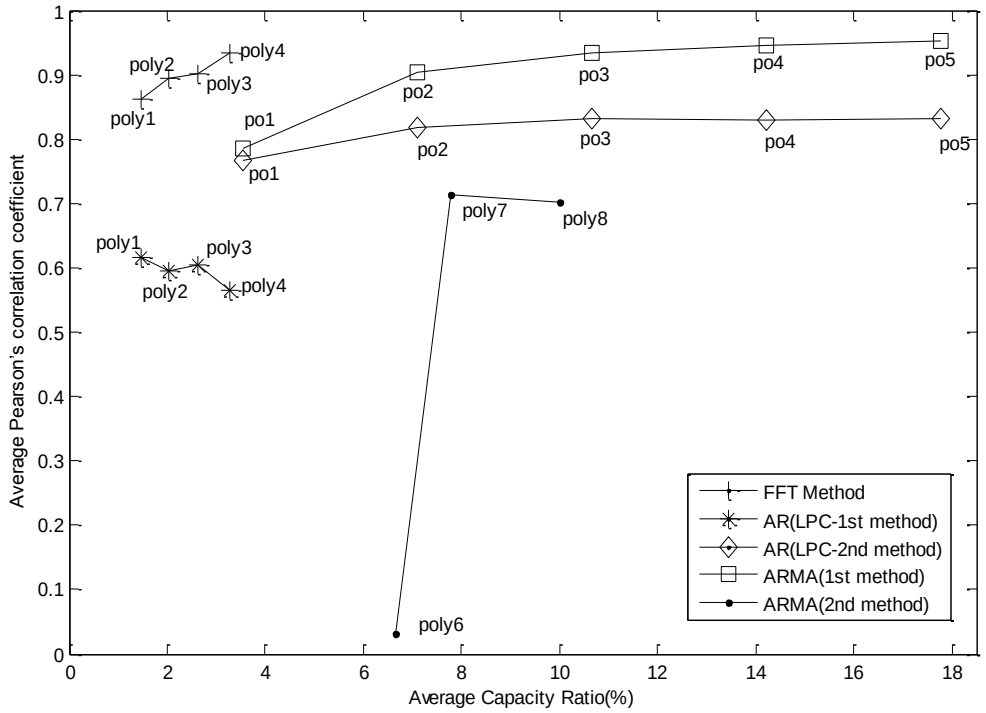

Figure 9: Pearson's correlation coefficient changes with Capacity ratio in all methods for 'Fa' phoneme transition (po1-Number indicates the number of points selected (po1), poly1-Number indicates the order of the polynomial)

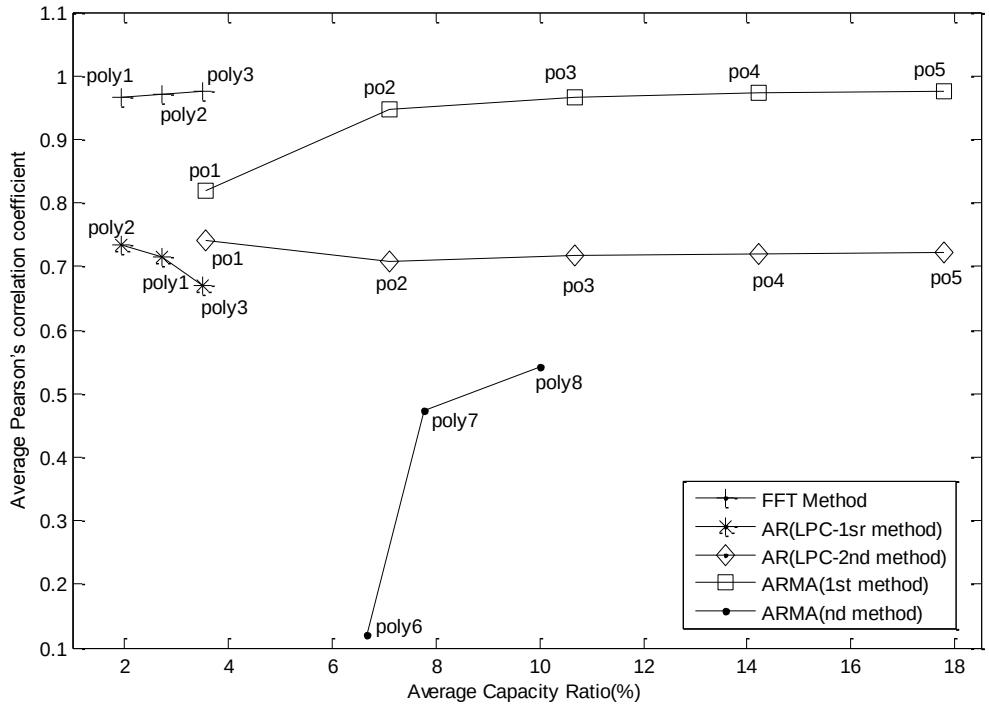

Figure 10: Pearson's correlation coefficient changes with Capacity ratio in all methods for 'Ma' phoneme transition (po1-Number indicates the number of points selected (po1), poly1-Number indicates the order of the polynomial)

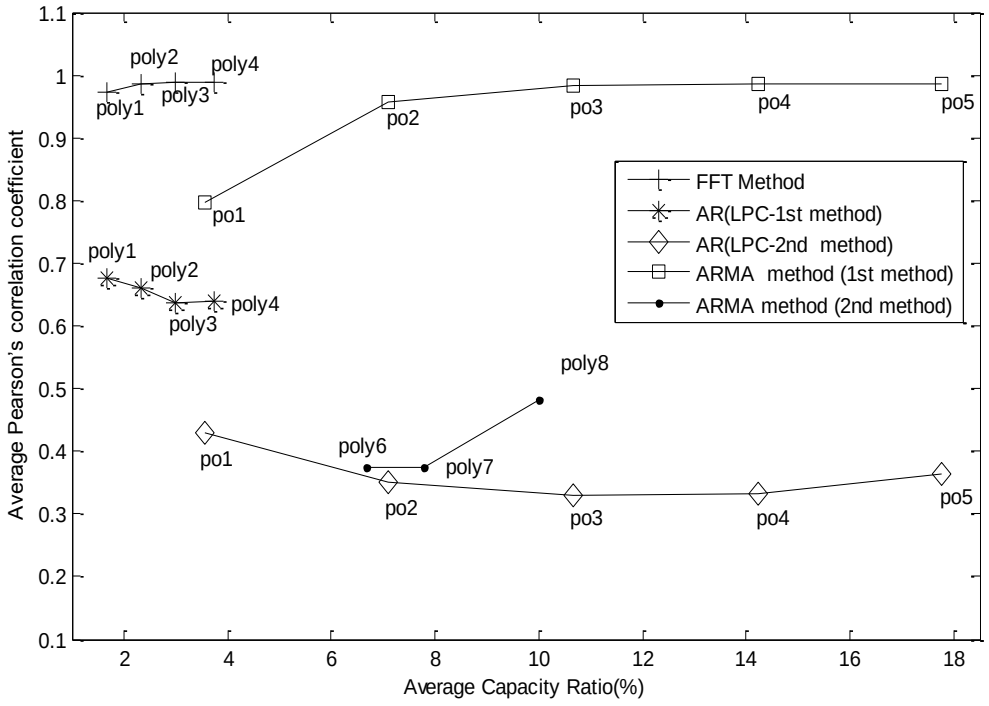

Figure 11: Pearson's correlation coefficient changes with Capacity ratio in all methods for 'Va' phoneme transition (po1-Number indicates the number of points selected ,poly1-Number indicates the order of the polynomial

4. DISCUSSION

In this research, a new parametric method has been proposed based on the sinusoidal noise model to synthesis transition region of consecutive phonemes with lesser number of parameters. To extract the speech information, most popular data extraction methods, Fast Fourier Transform (FFT), Auto Regressive model (LPC method), Auto Regressive Moving average (Steiglitz-McBride Method) were used.

This study points out that the new model reconstructs the phoneme transition regions that are much closer to the original phoneme transition using FFT method. It is also shown that high quality outputs were observed in higher order polynomials with higher capacity ratio. This is not in lined with the objective, but

a reasonable correlation coefficient value was observed in lower order polynomials with low capacity. In addition to that, the correlation values as well as the capacity ratio were increased as the number of dominant frequency component poles of FFT was Furthermore the observed correlation values depict that the quality of the output signal was not highly dependent on the number of dominant frequency component poles of FFT. It also shows that the signals constructed by lower order polynomials have correlation values exceeding 0.85. Hence it can be concluded that the transition regions can be modeled by lesser number of parameters using FFT data extraction method. The results of Auto regressive model (LPC method) in first attempt were opposite to the results of the FFT method. Thus the results conclude that the first attempt of the LPC method was not a suitable method to reconstruct the speech transition regions with low capacity.

The second attempt of the LPC method shows a fast improvement in the observed correlation values compared to the first attempt, but it is still not acceptable compared with the FFT method. The final method, which is the ARMA model (Steiglitz-McBride Method), shows a large improvement in the output than the LPC second attempt. This trail needs a higher capacity ratio compared to the FFT method but the output signal was better in quality. The result of the first attempt of the ARMA model was better than the second attempt and it concludes that the ARMA model first attempt provides better results than AR model (LPC method). When the sound of the output signal was compared, the ARMA gives a better quality output than the FFT method. As the capacity ratio is taken into account, FFT needs lesser capacity than the ARMA method. Overall it can be concluded that the FFT method provides better results than other methods. The proposed FFT model consumes lesser space to store information of each phoneme transition while the output is almost identical to the source signal.

It has investigated how to improve the naturalness of synthetic speech using lesser number of parameters. But more improvements for the methodology must be done. It has only considered a few transition regions collected from different phoneme categories. It is recommended to carry out the experiment for all the phoneme transition regions to develop a database. Here only the consonant to vowel transition of consonant to short /a/ sound was studied, but for all other transitions between phonemes must to investigated. Moreover in the experiment, all transitions between two phonemes were occurred in the beginning of the word but it needs to study how parameters change when the transition occurs in the middle of a word and in the end of the word. In the ARMA method still the capacity ratio is 10% from the original wave so further study should be carried out to reduce the capacity of the stored parameters. This preliminary experiment shows that there is a way of reconstructing the phoneme transition regions as it is. This is a low bit rate technique which is more useful for most of the speech synthesis systems, because of utilization of lesser number of parameters. Future studies should be carried out to investigate and to improve this novel method for future use

5. REFERENCES

[1] Tatham, M., Morton, K., Development in Speech Synthesis, John Wiley & Sons Ltd, 2005, 43-44.

increased. Increment factor of the correlation values was not very high and hence the generated signals from lower order polynomials can be accepted.

[2] Kröger, B., Minimal Rules for Articulatory Speech Synthesis, Proceedings of EUSIPCO, 1992, 92 (1), 331-334.

[3] Rahim M., Goodyear C., Kleijn B., Schroeter J., Sondhi M., On the Use of Neural Networks in Articulatory Speech Synthesis, Journal of the Acoustical Society of America, JASA vol, 1993, 93 (2), 1109-1121.

[4] Lemmetty, S., Review of Speech Synthesis Technology, M.Sc. Thesis, Helsinki University of Technology, 1999

[5] Holmes, J., Holmes, W., Speech Synthesis and Recognition, Second Edition, Taylor & Francis, 2001

[6] Wang, M., Speech Analysis And Synthesis Based On ARMA Lattice Model, Master's Thesis, University of Windor, 2003.

[7] Rabiner, L., Juang. B., Fundamentals of speech Recognition, Prentice Hall International, 1993.

[8] Sinha P., *Speech Processing in Embedded Systems*, Springer 2010 .

[9] O'Saughnessy D., Speech Communications – Human and Machine, Hyderabad Universities press, 2001

[10] Taylor, P., Text to Speech Synthesis, Cambridge University Press, 2009.

[11] Keller. E., Baily, G., Monaghan, A., Huckvale, M., Improvements in Speech synthesis, COST 258: The Naturalness of Synthetic Speech, John Wiley & Sons, LTD.

[12] Flanagan, J., Speech Analysis, Synthesis and Perception, Springer-Verlag, Second edition, 1972.

[13] Herath,H.M.L.N.K.,Wijayakulasooriya,J.V, (2014) A Sinusoidal Noise Model Based Speech Synthesis for Phoneme Transitions, International Journal Of Scientific & Technology Research Volume 3, Issue 7, July 2014 (Full Paper)

[14] Herath,H.M.L.N.K.,Wijayakulasooriya,J.V , (2015) . Comparison Of The Applicability Of FFT And LPC Methods For Natural Human Voice Synthesis.Proceeding of the Peradeniya University International Research Session (iPURSE).Vol 19. Pg 295(Abstrct –Poster)

[15] Herath, H.M.L.N.K.,Wijayakulasooriya,J.V. (2016), Auto Regressive Model Based PhonemeTransition Model For Natural Speech Synthesis. Elixir International Journal-Digital Processing, August issue – 2016

[16] Herath, H.M.L.N.K.,Wijayakulasooriya,J.V. (2017), Auto Regressive Moving Average Model Based Speech Synthesis for Phoneme Transition. IOSR Journal of Computer Engineering, volume 19 , issue I (Jan-Feb-2017) pp 103-109