Available online at www.elixirpublishers.com (Elixir International Journal)

Digital Processing

Elixir Digital Processing 97 (2016) 42080-42084

ISSN: 2229-712X

Auto Regressive Model Based Phoneme Transition Model for Natural Speech Synthesis

H.M.L.N.K Herath[a] and J.V. Wijayakulasooriya[b]

[a]Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka.

[b]Department of Electronic and Electrical Engineering Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka.

ARTICLE INFO

Article history:
Received: 27 June 2016;
Received in revised form:
1 August 2016;
Accepted: 6 August 2016;

Keywords
Speech synthesis,
Auto Regressive model (AR model),
Linear predictive Coding (LPC),
Correlation Coefficient,
Phoneme transition,

ABSTRACT

In parametric speech synthesis algorithms, a sequence of signals corresponding to phonemes is generated. However, synthesized speech tends to be unnatural as the vocal tract transition from one phoneme to another is not considered in most of the existing algorithms. This paper attempts to model the phoneme transition by extracting the speech parameters by means of Linear Time Varying system based Auto Regressive model. To reduce the capacity Speech parameters were represented in polynomial equations. Sinusoidal Noise model was used to reconstruct the phoneme transition region. The results show moderate correlation of reconstructed transition regions with synthesized signal for different orders of polynomial.

Introduction

The goal of speech synthesis is to convert a string of text, or a sequence of words, into natural-sounding speech. However, numerous techniques has been proposed in past decades and still no speech synthesis system which is available today is able to produce speech that could be characterized as natural or completely pleasant. The discontinuities of phoneme boundaries are identified as one of the significant factors affecting the quality of the synthetic speech. This discontinuity arises while connecting speech phonemes or segments to form words. In most of the parametric speech synthesis models, phonemes are represented using Linear Predictive Coding (LPC), synthesized separately and concatenated to form words, phrases and sentences. In this process the segments or phonemes do not consider the phoneme transitions.

Diphone synthesis is used for addressing this problem. Diphones, defined as central point of the steady state part of the phone to the central point of the following one, it contain the transitions between adjacent phonemes. In diphone synthesis concatenation point will be in the most steady state region of signal, which reduces the distortion from concatenation points. It preserves the finest acoustic details of natural speech. But the output is not natural, because diphone synthesis, the co-articulation achieved by considering only the immediately preceding and following phoneme. But in some cases some phonemes strongly affect the articulation of several preceding phonemes. The second reason is that the transition between diphones may not be sufficiently smooth and perceptually disruptive discontinuities arise in the middle of a phoneme when they are concatenated.

The use of large number of units (phones, diphones, polyphone…etc) has produced substantial increase in intangibility and naturalness of synthesized speech. Some of the available speech synthesis applications on the market use parametric synthesis instead of wavetable synthesis (concatenation) due to low storage capacity requirements. However, the former doesn't provide high quality speech compared to concatenation synthesis in every context. One of the most resent models is statistical parametric model based speech synthesis Hidden Markov model (HMM). It consists of spectral, pitch, and durational parameters for a context dependent phoneme. In there the concatenation speech units are done by considering the probability of phone transition[1]. It is generally found that HMM-based speech synthesis is more intelligible but less natural-sounding than unit selection.

All of the above speech synthesis techniques suffer from unnaturalness due to some discontinuities taken place when joining speech segments, phonemes, diaphone etc together. To overcome this issue, speech processing technique called pitch synchronous overlap add (PSOLA) method was originally developed [2][3]. It smoothly concatenates prerecorded samples and provide a good control for pitch and timing directly in the waveform domain, without needing any explicit parametric analysis of the speech. There are several versions of PSOLA algorithms were used in speech segment concatenation purpose and all of them work in essences the same way. The popular concatenation methods such as synchronous overlap add (SOLA)[4], Frequency Domain PSOLA[5], Time Domain PSOLA[6], Linear-Predictive PSOLA[7] etc are based on overlap-add method. Among the above methods, although TD-PSOLA provides good quality speech synthesis, it has limitations which are related to its non-parametric structure; spectral mismatch at segmental boundaries and tonal quality when prosodic modifications are

Tele: +94771306026
E-mail address: lakminiherath0@gmail.com

applied on the concatenated acoustic units [8]. But still the naturalness of synthetic speech is low.

The AR model (LPC algorithm) based method in this paper attempts to model the transition region from one phoneme to another using less number of parameters.

Methodology

To model the transition regions, a parametric mathematical model was developed. For demonstration purposes, short ‘a’ sound words were considered and the transition regions were segmented manually. The set words shown in Table 1 containing phoneme short ‘a’ uttered by a male speaker is used for the analysis

Table 1. Selected phoneme transition sounds and words.

Word stating phoneme	Phoneme category	Transition sound	Words
B	Plosives Voiced constant	Ba	Bat, Bag, Ban, Bad, Back, Band
P	Plosives Unvoiced constant	Pa	Pat, Pad, Pan, Pam, Pal, Pack
F	Fricatives unvoiced constant	Fa	Fat, Fad, Fan, Fact
V	Fricatives voiced constant	Va	Vat, Van
M	Nasals constant	Ma	Mat, Mad, Man, Mam, Map, Mack

Frequency peak points were extracted manually from each segment of the sound wave. The speech parameters were estimated by applying the Linear Predictive Coding (LPC) to the quasi-stationary part of the speech wave form. The basic analysis system for data extraction is shown in Fig.1. The amplitude frequency and phase values were calculated by considering the 5 dominant poles of the Linear predictive Coding. The experiment was carried out changing the number dominate component poles from 5 to 20. In AR model the coefficient of linear predictor (FIR filter) was estimated by applying the general equation

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

where, Fs is sampling frequency

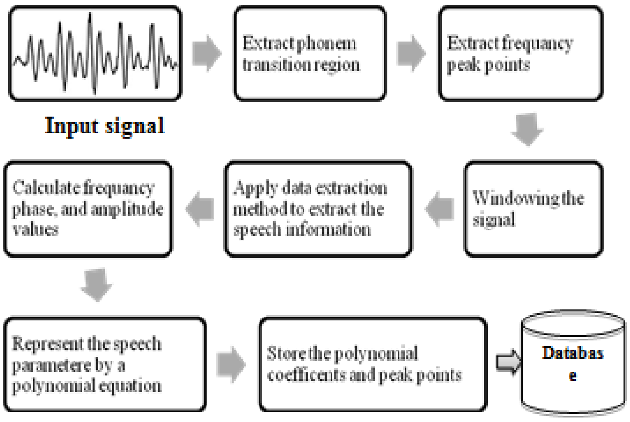

Input signal
Extract phonem transition region
Extract frequancy peak points
Windowing the signal
Apply data extraction method to extract the speech information
Calculate frequancy phase, and amplitude values
Represent the speech parameter by a polynomial equation
Store the polynomial coefficents and peak points
Databas e

Fig 1. Basic Analysis Model

Estimating Speech Parameters

Linear prediction models the human vocal tract as infinite impulse response(IIR) system that produces the speech signal. The linear predictive is so called because it assumes that the output samples can be predicted by a linear combination of the filter parameters and the previous samples. It used an all pole filter to simulate the vocal tract. The basic idea behind the LPC model is that a given speech sample at time n, S(n) can be approximated as a linear combination of the past *p* speech samples, such that,

$$S(n) \sim a_1 s\,(n-1)+ a_2 s\,(n-2)+ a_3 s\,(n-3)+\ldots+ a_p s\,(n-p) \qquad (2)$$

Where, the coefficients $a_1,a_2,\ldots,a_p$ are assumed constant over the speech analysis frame [1].

Including an excitation term G u(n) giving,

$$S(n) = \sum_{i=1}^{p} a_i\, s(n-i) + G\, u(n) \qquad (3)$$

Where, u(n) is a normalized excitation and G is the gain of the excitation. By expressing above equation in the z domain we get the relation,

$$S(z) = \sum_{i=1}^{p} a_i\, z^{-i} s(z) + G\, U(z) \qquad (4)$$

Leading the transform function

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

By minimizing the sum of the squared differences (over a finite interval) between the actual speech samples and the linearly predicated ones, a unique set of predicator coefficients (a_i) is determined. Speech parameters frequency, phase amplitude derived in terms of the predication coefficients a_i. The partial fraction representation H(z) express as,

$$H(z) = \frac{B(z)}{A(z)} = \frac{r_m}{s-p_m} + \frac{r_{m-1}}{s-p_{m-1}} + \cdots + \frac{r_0}{s-p_0} + k(z) \qquad (6)$$

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. 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

$$A_n = |r_m| \qquad (7)$$

$$\phi_n = \tan^{-1}\left(\frac{r_{im_n}}{r_{Re_n}}\right) \qquad (8)$$

Pole locations p_m used to calculate the frequency f_n

$$f_n = \tan^{-1}\left(\frac{p_{im_n}}{pl_{Re_n}}\right) \times ((Fs/2)/\pi) \qquad (9)$$

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 a polynomial equations.

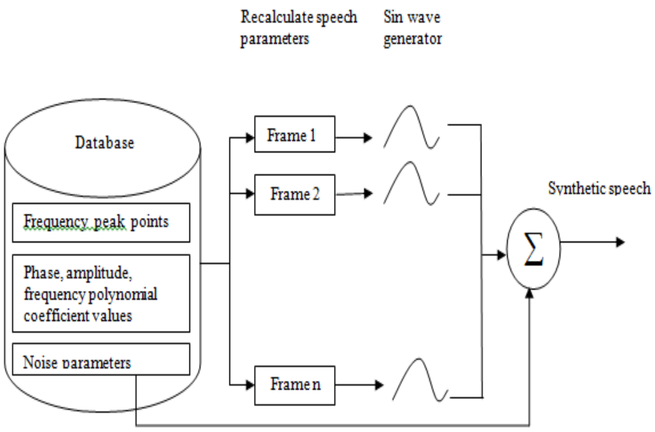


Fig 2. Proposed System.

$$\begin{bmatrix} A_i \\ \phi_i \\ f_i \end{bmatrix} = \begin{bmatrix} b_m & b_{m-1} \cdots & b_0 \\ c_m & c_{m-1} \cdots & c_0 \\ d_m & d_{m-1} \cdots & d_0 \end{bmatrix} \begin{bmatrix} x^m \\ x^{m-1} \\ \vdots \\ 1 \end{bmatrix} \qquad (10)$$

Where $b_1 \ldots\ldots b_m$ and $c_1 \ldots c_m$ are polynomial coefficients.

Signal Reconstruction

Sinusoidal Noise Modeling

The sinusoidal noise model is a parametric speech synthesis model originally proposed by McAulay & Quatieri for speech coding purposes and by Smith & Serra 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 [10]. 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 [11].

$$x(t) = \sum_{i=0}^{N} A_i(t)\cos\left(\theta_i(t)\right) + r(t) \qquad (11)$$

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,

$$x(t) = \sum_{i=0}^{N} A_i(t)\cos\left(\varpi_i t + \phi_i\right) + r(t) \qquad (12)$$

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

$$x(t) = \sum_{i=0}^{N} A_i(t)\cos\left(2\pi f_i t + \phi_i\right) + r(t) \qquad (13)$$

where, f_i is the oscillation frequency in ith sinusoidal component.

$$x(t) = \sum_{i=0}^{N} A_i(t)\, e^{-\alpha t}\cos\left(2\pi f_i t + \phi_i\right) + r(t) \qquad (14)$$

Equation 14 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 2). White Gaussian noise was applied to generate the noise residuals using mean and standard deviation of the noise.

It is infeasible to carry out the experiment for all those words, thus some words were selected by considering the phoneme classification. Then Pearson's correlation coefficient between original wave and the reconstructed wave were calculated. The required capacity to store the source wave form and the proposed method speech parameters were compared by calculating the capacity ratio.

Results and Discussion

Figure 3 and Figure 4 shows how the capacity ratio changes with the Pearson's correlation coefficient in different polynomial orders. All observed correlation values were less than grater than 0.5 and less than 0.8. This interprets there are moderate positive correlation between the original signal and the constructed signal. When the capacity ratio was increased, the correlation values were increased only in 'Ba' sound of word 'Bat'. For other sounds when the capacity ratio was increased with the polynomial order the correlation was reduced. The signal quality wasn't improved when the capacity ratio was increased.

Figure 4 illustrates how the average correlation coefficient changes with polynomial order for all the phoneme transitions. Generally average correlation values lies within the range of 0.5 and 0.8. According to the error plot the variability from the mean value was within ±0.15.The moderate correlation was observed between the source signal and the reconstructed signal for all of the phoneme transition sounds. 1st order polynomial coefficients recalculate the amplitude and phase values which are more similar to the source values. Because the signals synthesized using 1st order coefficients have higher correlation than other polynomial orders

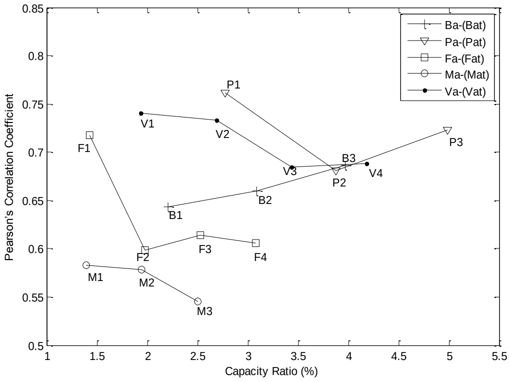


Fig 3. Pearson's correlation coefficient changes with Capacity ratio in different polynomials, considering five dominant poles of LPC. (M1- Number indicates the order of the polynomial).

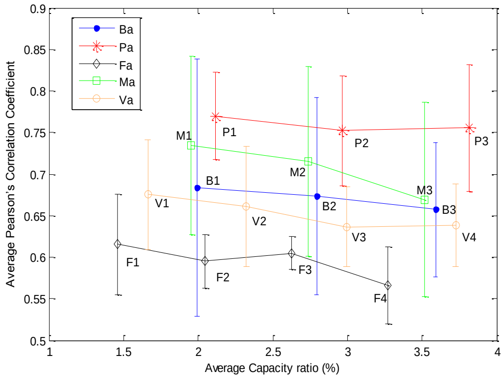
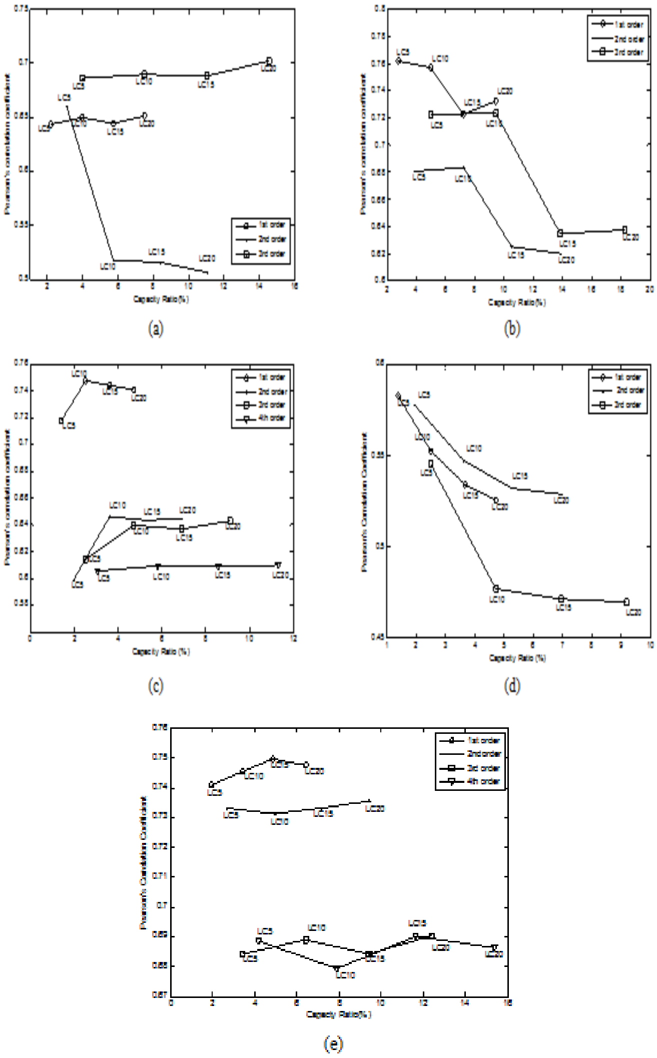


Fig 4. Average Pearson's Correlation Coefficient changes with Polynomial Order considering first five dominant frequency component poles of LPC in different phoneme categories.

Following figures (Fig 5) show how the average correlation changes with the capacity ratio when the number of dominant poles changes from 5 to 20. For 'ba' (word Bat) transition the highest correlation was found when the number of poles changes 5 to 20 in 3rd order. In 2nd order the highest correlation value was observed in first five dominant poles.

When the number of dominant poles was increased, the correlation values were fluctuated with the capacity ratio. 'ma' (word Mat) , 'pa'(word Pat) phoneme transitions all the correlation values were decreased with the capacity ratio. This is because when the number of poles was increased, the algorithm extracted some unwanted information that disturbs the other important information. For all transition sounds the Pearson's correlation values were less than 0.8.



(a) (b) (c) (d) (e)

Fig 5. Pearson's correlation coefficient changes with Capacity ratio in different polynomials in different number of dominant poles of LPC for (a) 'ba' of word 'Bat' (b) 'pa' of word 'Pat' (c) 'fa' of word 'Fat'(d) 'ma' of word 'Mat' (e) 'va' of word 'Vat'.

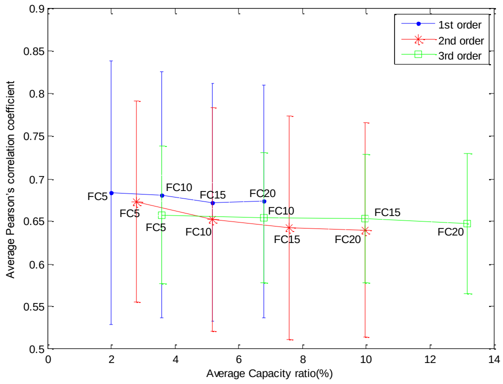


Fig 6. Average Pearson's correlation coefficient change with number of dominant frequency component poles of LPC in different polynomial orders for 'ba' phoneme transition.

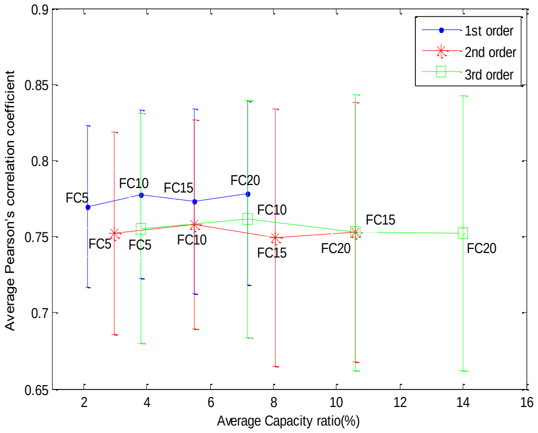


Fig 7. Average Pearson's correlation coefficient change with number of dominant frequency component poles of LPC in different polynomial orders for 'pa' phoneme transition.

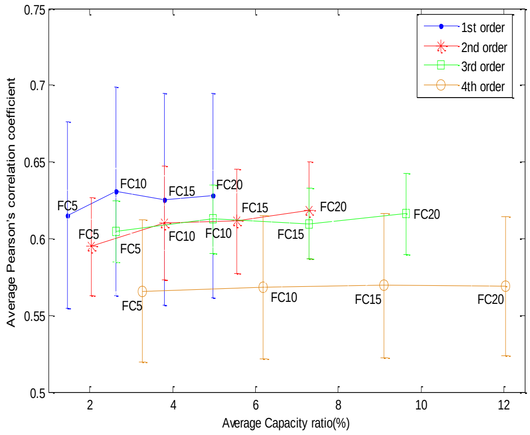


Fig 8. Average Pearson's correlation coefficient change with number of dominant frequency component poles of LPC in different polynomial orders for 'fa' phoneme transition.

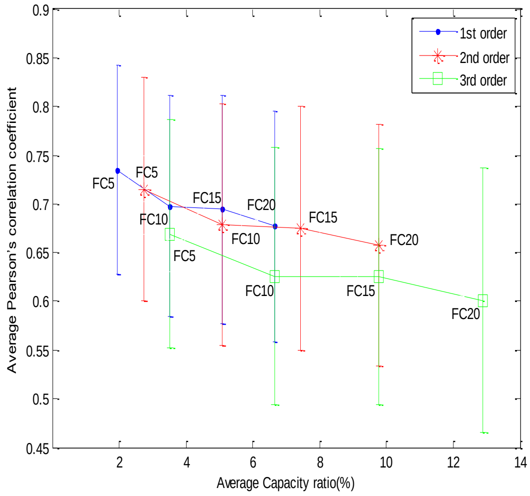


Fig 9. Average Pearson's correlation coefficient change with number of dominant frequency component poles of LPC in different polynomial orders for 'ma' phoneme transition..

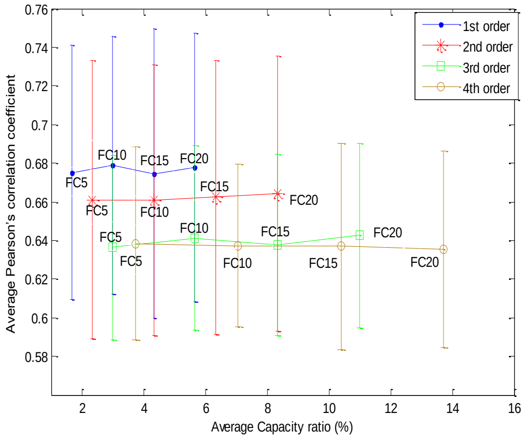


Fig 10. Average Pearson's correlation coefficient change with number of dominant frequency component poles of LPC in different polynomial orders for 'fa' phoneme transition

Conclusion

In this paper, a new parametric method has been proposed based on the sinusoidal noise model to synthesis transition region of consecutive phonemes with less number of parameters. Speech parameters were extracted using LPC algorithm. The constructed transition region was deviated from the source signal in moderate amount. Results didn't show any significant pattern with the capacity ratio and the polynomial order. The observed correlation coefficient values were less than 0.8 which concludes that the constructed signal was moderately correlated with the source signal. Significant improvements cannot be observed by increasing the number of LPC coefficients or the order of the polynomial.

References

[1]. Taylor, P., (2009). *Text to Speech Synthesis*, Cambridge University Press first edition

[2]. Holmes, J., Holmes, W. (2001). *Speech Synthesis and Recognition*, Second Edition, Taylor & Francis,.

[3]. Benesty, J., Sondhi, M. M., Huang, Y., *Springer Handbook of Speech Processing.* Springer

[4]. Roucos,S.,Wilgus, A., (1985) *High-Quality Time Scale Modification of Speech*, in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'85, 236-239.

[5]. Charpentier,F. J.,Stella, M.G., (1986) *Diphone synthesis using an overlap add technique for speech waveforms*, in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'86, 2015-2018.

[6]. Hamon, C.,Moulines, E.,Charpentier, F., (1989) *A diphone synthesis system based on time-domain prosodic modifications of speech*, in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'89, 238-241.

[7]. Havelock, D., Kuwano, S., Volander. M.,(2008). *Handbook of Signal Processing in Acoustics*, Springer,

[8]. Klatt,D., (1987). *Review of Text-to-Speech Conversion for English*, Journal of the Acoustical Society of America, JASA, 82 (3), 737-793.

[9]. Torres, R.C.,Seixas, J.M. de,Netto, S.L.,Freitas, D.R. da S.,Brasil, E.F., (2008). *Portable implementation of a text-to-speech system for Portuguese*, in Proc. of EUSIPCO 2008,

[10]. Serra, X., *Musical Sound Modeling with Sinusoids plus Noise. Musical signal processing*. Published in C. Roads, S. Pope, A. Picialli, G.DePoli Editors, Musical signal processing Swets & Zeitlinger Publishers,

[11]. Turi Nagy M., Rozinaj.G., (2004). *An Analysis/Synthesis System of Audio Signal with Utilization of an SN Model*, Radio engineering, Pattern Recognition Association of South Africa (PRASA)conference, Vol. 13, No. 4,

[12]. Phung, T., Luong, M. C., Akagi, M.(2011), *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): 512-514

[13]. Phung, T., Luong, M. C., Akagi, M.(2012) *On the Stability of Spectral Targets under Effects of Coarticulation*, International Journal of Computer and Electrical Engineering, Vol. 4, No. 4, (537-541)

[14]. Shannon M, Zen H, Byrne W,(2013) *Autoregressive Models for Statistical Parametric Speech Synthesis*, IEEE transactions on audio, speech, and language processing, vol. 21 (3); (587-597)

[15]. Verhelst,W., Roelands,M., (1993). *An Overlap-Add Technique based on Waveform Similarity (WSOLA) for High-Quality Time-Scale Modification of Speech*, in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'93, 554-557

[16]. Tatham, M., Morton, K., (2005) *Development in speech synthesis.* John Wiley & Sons Ltd, England, Chapter 4, pg 43-44.

[17]. Schnell, N., Peeters G., Lemouton, S., Manoury, P., Rodet, X., (2006) *Synthesizing a Choir in Real-Time using Pitch Synchronous Overlap Add (PSOLA)*, http//www.ircam.fr,

[18]. Mousa, A., (2010), *Voice Conversion Using Pitch Shifting Algorithm By Time Stretching With Psola And Re–Sampling*, Journal of Electrical Engineering, VOL. 61, NO. 1,57–61

[19]. Klatt, D., (1987), *Review of Text-to-Speech Conversion for English. Journal of the Acoustical Society of America*, JASA vol. 82 (3), pp.737-793,

[20]. Torres, R.C., Seixas, J.M. de, Netto, S.L., Freitas, D.R. da S., Brasil, E.F., (2008) *Portable implementation of a text-to-speech system for Portuguese*, in Proc. of EUSIPCO .

[21]. Herath,H.M.L.N.K.,Wijayakulasooriya,J.V, (2014) *A Sinusoidal Noise Model Based Speech Synthesis for Phoneme Transitions*, Postgraduate Institute of Science research congress 2014, pg 49.

[22]. 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.