Chaotic Mixed Excitation Source for Speech Synthesis

Hemant A. Patil and Tanvina B. Patel

Dhirubhai Ambani Institute of Communication and Technology, (DA-IICT),
Gandhinagar-382007, India

{hemant_patil, tanvina_bhupendrabhai_patel}@daiict.ac.in

Abstract

Linear Prediction (LP) analysis has proven to be very powerful and widely used method in speech analysis and synthesis. Synthesis by LP-based approach is carried by exciting an all-pole model (whose parameters are derived by LP analysis). Synthesis is carried by using mixed excitation source consisting of a sequence of *impulses* for *voiced* regions and *white-noise* source for *unvoiced* regions. In this paper, we present novel chaotic excitation source using *chaotic titration* method. The voiced and unvoiced regions in speech are modeled by *chaos* which is quantified by adding noise of known standard deviation (determined using *chaotic titration* method). It is observed that on an average for synthesized voices (both male and female**)**, MOS increases from *2* to *2.4*, DMOS from *2.1* to *2.4* and preference is increased from *39 %* to *61 %* via A/B test. PESQ score increases from *1* to *1.5* and MCD score decreases from *4.06* to *4.03,* relatively for voices synthesized by proposed chaotic mixed excitation source. The relatively better performance of proposed approach is may be due to the novel chaotic mixed source of excitation.

Index Terms: Linear Prediction, Nonlinear Prediction, Volterra-Wiener-Korenberg Series, Chaos, Speech Synthesis

1. Introduction

Linear Prediction (LP) is an extensively used technique for analyzing the acoustic properties of speech [1]. The acoustic theory assumes that the speech production mechanism to be a *linear* system. LP analysis is very efficient in speech analysis and synthesis as it captures *implicitly* the frequency response of the time-varying vocal tract area function and estimates the frequency, amplitude, and bandwidth of formants.

Apart from the linear system characteristics, the nonlinear characteristics of human speech production system make new insights in speech analysis. Research in this area is initiated due to the original investigations by the Teagers [2], [3], which gave several evidences that speech is outcome of *nonlinear* system. In addition, nonlinearities in speech have potential applications in detecting laryngeal pathologies [4]. Nonlinear dynamic methods such as fractal dimension (FD), Lyapunov exponent (LE), correlation dimension (CD), etc. are used to describe the nonlinear characteristics of speech signal [5]. Studies in [6] reported that fricatives have a high FD of around *1.7* than vowels that have a FD of *1.2*. Parameters like LE may fail to detect chaos from short and noisy data [7]. Traditional nonlinear dynamical methods require that the analyzed time series must be stationary and long which may not be always true for natural speech signal.

One of the former attempts to model nonlinearities in speech was reported by Thyssen *et. al.* for speech coding using Volterra-Wiener (VW) series [8]. There are efforts to model the global behavior of VW series using artificial neural networks (ANN) [9], [10]. In [11], an application was shown to model nonlinear characteristics of electronic devices (such as diode capacitance curve). Thereafter, the instability of VW series nonlinear predictive filter used for speech coding was studied [12]. Thereafter, a novel use of Volterra series was presented to analyze the multilayered perceptron (MLP) to estimate the posterior probability of phoneme for automatic speech recognition (ASR) [13]. The study used Volterra kernels to capture *spectro-temporal* patterns that are learned from the Mel filterbank log-energies (during training of system for each phoneme). Very recently, the authors have shown that nonlinear prediction (NLP) by VW series method is better than LP and it gives relatively *flat* residual spectrum than its LP counterpart [14]. In the same study, the prediction power of LP and NLP is used to estimate chaos in the speech signal. It has been shown that chaos exists in speech as well as its residuals (LP and NLP). The residual error is replaced by a *sequence* of impulses in many speech synthesis schemes. Hence, the synthetic speech loses nonlinear characteristics of the natural speech. This might be a reason that synthetic speech sounds machine-like rather than natural speech [15].

The source-filter model of *speech synthesis* uses *LP residual* as an *excitation source* to the vocal tract (all-pole model) [16]. Speech is synthesized as the output of a linear recursive filter excited by either sequence of quasi-periodic pulses or white-noise source [17]. Extension of this is *mixed excitation* LP-based vocoder model (MELP codec) to achieve low bit-rate [18]. A harmonics plus noise model (HNM) also exists that implicitly assumes a two-band (lower harmonic band and an upper noise-like band) mixed excitation [19]. Earlier a potential work using chaos for speech synthesis is carried out in [20] that describe a novel design of speech synthesizer. Chaos-based speech synthesis techniques are reported in [21], [22]. Oscillator models for speech modeling and synthesis have been exploited for several years [23]. An extension of this model is to re-generate noise-like, i.e., high-dimensional unvoiced component of stationary speech signals [24].

It has been known that vocal fold oscillations process is *nonlinear*. Vocal fold models are also characterized in terms of their *bifurcation* diagrams [25]. Hence, the source of excitation for speech should be *chaotic* in nature. It is known that, the excitation cannot be perfectly presented as an impulse train or white noise and also harmonic-excitation alone is known to be poor. In this paper, we use a numerical chaotic titration procedure such as a *litmus test* for sensitive and robust detection of chaos [26]. The novelty in the present work lies in using our recent work [14], for creating a novel chaotic mixed excitation source for speech synthesis. The method uses a computational procedure, based on comparison of the prediction power of LP and NLP model of the VW form [27]. The voiced regions are replaced by sequence of impulses and unvoiced regions by noise. However, for the proposed mixed excitation strategy, the amount of noise added is controlled by the chaotic behaviour of speech in both voiced and unvoiced regions. The amount of chaos in the speech signal is estimated through chaotic titration method in terms of *Noise Limit* (NL) values. It is shown that instead of using impulses at voiced regions and noise at unvoiced regions, intelligibility is more

by using chaotic voiced source for voiced regions and chaotic noisy source (of known standard deviation) at unvoiced regions. This is verified by Perceptual Evaluation of Speech Quality (PESQ) score, Mel Cepstral Distance (MCD) measure, Mean Opinion Score (MOS) (along with Student *t*-test), Degraded Mean Opinion Score (DMOS) and preference of synthesised voices via A/B test.

2. Prediction of Speech

Linear Prediction (LP) deals with representing a speech sample *s(n)* in terms of the *linear* combination of its previous *p* samples [16]. Hence, an estimate of the speech signal can be represented using *p* predictor memory of $s(n)$ as follows,

$$\hat{s}_{LP}(n) = -\sum_{k=1}^{p} a_k s(n-k), \tag{1}$$

where a_k's are *optimal linear* predictor coefficients obtained by minimizing the l^2 energy of LP residual signal, i.e.,

$$e_{LP}(n) = s(n) - \hat{s}_{LP}(n). \tag{2}$$

Nonlinear prediction (NLP) can be based on Volterra series expansion with the expansion limited to 1^{st} and 2^{nd} order terms for simplicity. A nonlinear system with *k* memory terms can be represented as an extension of power series expansion such as the Taylor series. For time series *s(n)*, a discrete VW series of degree *d* and memory *k* is used as a NLP model to estimate the predicted time series $\hat{s}(n)$ as [28];

$$\hat{s}_{NLP}(n) = a_0 + a_1 s(n-1) + \ldots + a_k s(n-k) + \ldots + a_{k+1} s(n-1)^2 + a_{k+2} s(n-1) \times s(n-2) + \ldots + a_{M-1} s(n-k)^d = \sum_{m=0}^{M-1} a_m q_m(n), \tag{3}$$

where the functional basis $\{q_m(n)\}$ includes all the distinct combinations of the *embedding* space coordinates $\{s(n-1), s(n-2), \ldots, s(n-k)\}$ up to degree *d*, with a dimension of $M=(k+d)!/k!d!$ [27]. Each model is parameterized by *k* and *d* corresponding to the *predictor memory* and *degree* of nonlinearity in the model, respectively. The coefficients a_m's are estimated by Korenberg's fast algorithm using Gram-Schmidt procedure from linear and nonlinear autocorrelation of the speech time series [29]. Hence, the VW series is referred as Volterra-Weiner-Korenberg (VWK) series. Thus, for the *NLP* model, $s_n^{calc} = \hat{s}_{NLP}(n)$ and hence, NLP residual, *viz.*, e_{NLP} is given by

$$e_{NLP}(n) = s(n) - \hat{s}_{NLP}(n). \tag{4}$$

Frechet proved that the set of Volterra functionals is *complete* (i.e., every Cauchy sequence converges to an *accumulation point* belonging to the same *function space*) [30] and can be approximated with arbitrary precision as sum of finite Volterra functionals in *s(t)* (pp. 15, [28]).

3. Chaotic Titration Method

The idea of the chaotic titration method is quite different from the other nonlinear dynamical methods of chaos estimation. This chaos detection method is named titration of chaos because it is analogous to a *chemical* titration process.

3.1. Methodology

After obtaining coefficients of the model [29], its short–term prediction power is measured by the standard deviation of one-step-ahead prediction error, calculated as follows [27],

$$e^2(k,d) \equiv \frac{\sum_{n=1}^{N} (s_n^{calc}(k,d) - s(n))^2}{\sum_{n=1}^{N} (s(n) - \bar{s})^2}, \tag{5}$$

where $s_n^{calc} = \hat{s}(n)$, $\bar{s} = \frac{1}{N}\sum_{n=1}^{N} s_n$ and $e^2(k,d)$ is the *normalized variance* of the residuals. The best model $[k_{opt}, d_{opt}]$ minimizes the following information criterion in accordance with the *parsimony principle* [31]:

$$C(l) = \log[e(l)] + \frac{l}{N}, \tag{6}$$

where $l \in [1, M]$ is the number of polynomial terms of the truncated VWK series expansion for a certain pair $\{k,d\}$. The procedure includes obtaining k^{lin} for the best *linear* model which minimizes *C(l)* with *d*=1. Repeat again, with increasing *k* and *d>1* to obtain the best *nonlinear model*.

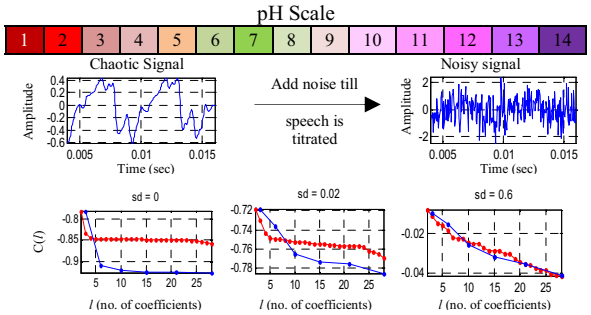

Figure 1: *Analogy of chemical titration and chaotic titration.*

3.2. Estimating the Chaos by Noise Limit

To detect the chaos in the signal it is necessary to estimate the noise limit (NL). The NL indicates as to when the information obtained from the linear and nonlinear predicted coefficients is same. The VWK series will be now used as an indicator (just as *litmus paper* in the chemical titration process).

For a frame of speech signal (duration *10 ms*) both linear and nonlinear dynamics are predicted. If the *l-C(l)* curves of the speech frame are sufficiently apart, we add noise of certain standard deviation (i.e., $sd=\sigma$) to the speech frame and obtain the *l-C(l)* plot again. Noise is added to speech till the nonlinearity estimated in the *l-C(l)* plot goes undetected by the nonlinear indicator, i.e., at a certain σ, the nonlinearity will not be detected by the indicator and the *l-C(l)* plots for LP and NLP will overlap (i.e., the given speech frame is *titrated*). The indication that signal is neutralized by the noise added is when the curves of nonlinear and linear error values are nearly close to each other or in other words, the prediction done by the linear and nonlinear method is nearly the same. The value of the σ of the noise added when the speech is neutralized is known as the '*noise limit*' (i.e., *NL*) or '*noise ceiling*' of the speech signal corresponding to the amount of chaos [26]. Titrating the speech signal frame-by-frame will give an estimate of the *NL vs.* time. More chaotic the signal is, more should be its *NL*.

The chaotic titration method is illustrated in figure 1. The underlying concept is that, *NL>0* indicates chaos, while *NL=0* indicates that the data series either is not chaotic or the chaotic component is neutralized by background noise [26]. Thus, *NL>0* gives sufficient test for *chaoticity*. In [26], the titration method is verified by a bifurcation diagram of the logistic map to obtain a relation between the largest Lyapunov exponent (LLE) and the *NL* values. In addition, as noise itself is used as a titrant for chaos, the *NL* test is robust to measurement noise. Thus, the chaotic titration method used in this paper for the estimation of *NL* is convenient to determine the *chaos embedded* in the speech signal.

4. Speech Synthesis

4.1. Synthesis by LP analysis

LP-based approach for speech synthesis uses vocal tract as an all-pole filter with system function as [16],

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} = \frac{G}{A(z)}, \quad (7)$$

where p is the number of poles, G is the filter gain, and a_k*'s* are pole determining parameters. Two models exists to model *voiced* and *unvoiced* speech sounds. Voiced speech being periodic in nature is generated by exciting the all-pole filter model by quasi-periodic impulse-train with period equal to the desired pitch period and unvoiced speech sounds are generated by exciting the all-pole filter model by the *noise-like* source.

4.2. Iterative Adaptive Inverse Filtering (IAIF)

For synthesis by LP analysis, the locations of the quasi-periodic impulse train are obtained from the glottal closure instants (GCI) of the speech signal. There are several methods to estimate the GCI locations, *viz.,* glottal closure instants of the glottal flow waveform (GFW), peaks of LP residual [1], Hilbert envelope of LP residual [32], Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) [33], etc. In this paper, we have used the GFW of the speech signal obtained from IAIF method which decomposes speech into its glottal source signal and vocal tract system [34]. The key motivations for using the IAIF method are (1) no need of ground truth (such as an electroglottograph (EGG)), (2) it is computationally efficient and completely automatic and (3) it naturally fits into the philosophy of LP-based speech synthesis. In the IAIF method, the effect of the vocal tract system and lip radiation is cancelled from the speech signal to estimate the GFW. Using Glottal Inverse Filtering (GIF), the derivative of the GFW is obtained and the negative peak of the differenced EGG gives the location of the impulses in the excitation source. The block diagram of IAIF method is shown in figure 2 [35].

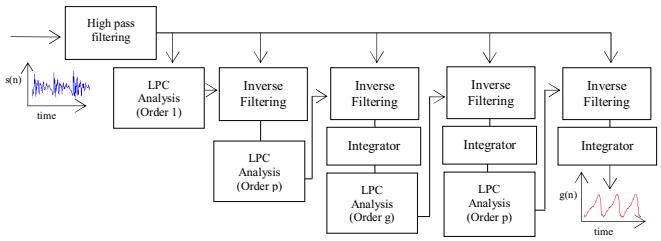

Figure 2: *Block diagram of IAIF method. The GFW g(n) estimated through repetitive procedure of canceling vocal tract and lip radiation effects from speech s(n) (p=20, g=4). (After* [35]*).*

Figure 3 shows a speech segment of a vowel /aa/ from the TIMIT database [36]. The impulse-train obtained from the closed phase of the GFW act as an excitation source (due to *sudden* closure of the vocal folds) to the all-pole model of speech to get the synthesized vowel as shown in figure 3 (d).

4.3. Synthesis by Chaotic Source

For synthesis by chaotic excitation, we add chaos to the sequence of impulses (i.e., proposed chaotic excitation source). For an utterance /aa/ (in figure 4(a)) the *NL* values are estimated for every *10 ms* of speech frame and noise with *NL* standard deviation is added to corresponding frame of impulses. *NL* being an estimate of chaos, adding *NL* standard deviation noise, makes the excitation source *chaotic*. The chaotic source and synthetic speech are shown in figure 4 (c) and figure 4(d), respectively.

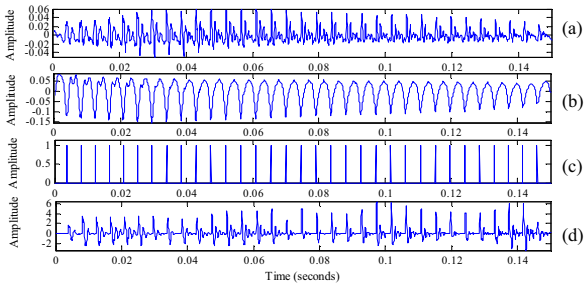

Figure 3: *a) Speech utterance for vowel /aa/ from TIMIT database* [36] *(b) GFW for signal in (a), (c) the impulses at the location of closing of GFW and (d) synthetic speech generated with (c) as a source to all-pole filter with predictor order p=20.*

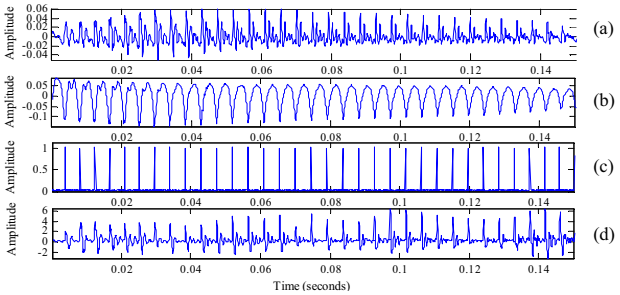

Figure 4: *a) Speech utterance for vowel /aa/ from TIMIT database* [36] *(b) GFW for signal in (a), (c) the impulses at the location of closing of GFW and chaos added to sequence of impulses of NL standard deviation (d) synthetic speech generated with (c) as a source to all-pole filter with predictor order p=20.*

5. Results and Discussions

5.1. Experimental Setup

In this work, we synthesize *20* utterances from *CMU-ARCTIC* database (*1* male and *1* female) [37] and we evaluated these sentences to test the naturalness and speech quality. For short segments, especially vowels alone, the improvement in naturalness may not be perceived well, i.e., any significant improvement does not show up in a single high amplitude vowel. Similarly, for unvoiced sounds, (*viz.,* fricatives /f/, /s/), the synthesized signal is perceived as noise. Therefore, we synthesize long sentences to estimate the importance of chaos in excitation source. Here, *20* utterances from *CMU-ARCTIC* database [37] were synthesized with impulse and noisy source (Method 1: M1) and impulsive source with noise of known σ derived from chaotic titration method (i.e., impulse plus chaos) (proposed Method 2: M2). Figure 5 shows the waveforms of one of twenty synthesized utterances. It is seen that in figure 5(b), that the unvoiced regions (as shown by dotted circles) are more noisy than in case of figure 5c. Thus, the speech synthesized by M2 has less of background noise due to addition of controlled noise in the source excitation.

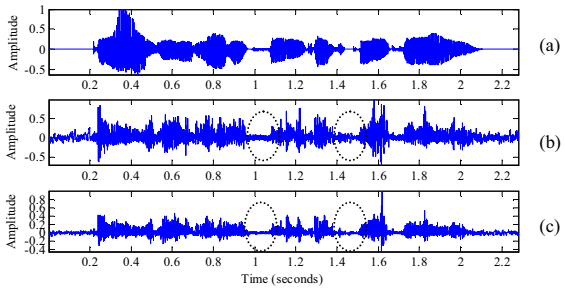

Figure 5: *a) Speech utterance from CMU-ARCTIC database (arctic_0009 (Fs=16 kHz))* [37] *(b) speech synthesized by M1, (c) speech synthesized by M2.*

5.2. Subjective Evaluation

Twenty synthesized utterances from each M1 and M2 system were taken and MOS analysis was done to get an estimate of the naturalness of the synthesized speech. Files from different systems are played randomly and subjects were asked to give score in *1* to *5*. As quality of synthetic speech depends on that of the original speech, the MOS is normalized to that of natural speech. This is referred to as degraded MOS (DMOS) which gives the performance of synthesized signal relative to the natural voice [38]. The MOS and DMOS have been evaluated by *15* listeners. Table 1 show that the MOS and DMOS are more for M2 system than that of the M1 system. Hence, the naturalness in synthesized speech is more in case M2 system (impulse+chaotic source) than in that of M1 system (impulse+noise source). It is observed that there is an improvement around *14 %* for speech synthesized by M2 than by M1. This is quantified by measures shown in Table 1.

Table 1. *The average MOS, DMOS and A/B test (from 15 listeners) for 20 utterances of CMU-ARCTIC database* [37].

Method	MOS		DMOS		A/B Test (%)	
	M	F	M	F	M	F
M1	2.09	2.04	2.10	2.06	37	41
M2	**2.42**	**2.28**	**2.43**	**2.30**	**63**	**59**

In addition to the MOS and DMOS, A/B test had been performed to judge the preference of the listeners when the listeners were subjected to the synthesized files played randomly from the systems M1 and M2. It is observed that on an average about *61 %* of the people preferred the synthesized files of the proposed method which shows the importance of the controlled noise via chaotic titration as excitation source.

To quote statistical significance of the results, Student's *t*-test is used. The test performed along the sentences for M1 and M2 systems (male and female) showed that the probability of accepting the null hypothesis is less than *0.0001* and *0.0004* for male and female systems, respectively. Hence, the M1 and M2 systems have different means (as shown in figure 6(a)).

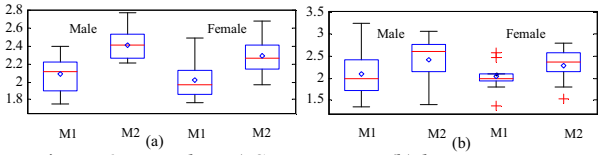

Figure 6: *Box-plots a) Sentencewise (b) listenerwise.*

To know the competence of listeners in evaluating, Student's *t*-test was computed listenerwise for male and female systems. Here, the null hypothesis was accepted with a probability of *0.14* and *0.11* for male and female systems, respectively. Hence, the listeners evaluated all systems almost similarly.

5.3. Objective Evaluation

To evaluate the quality of synthesized voices, the PESQ score [39] and widely used MCD measure is used [40]. MCD is used for measuring the accuracy of the spectral envelope of the synthetic speech with respect to natural speech. Here, the MCD between the speech frame of the natural and synthetic frames are estimated [40]. It is observed from Table 2 that the speech synthesized by controlled addition of noise (M2) is better in terms of PESQ score than M1 due to less noise (i.e., PESQ increases). In addition, the MCD scores for the speech synthesized by M2 method is less than the MCD scores for speech synthesized by M1 (i.e., distance decreases and the M2 system performs better). For MCD there is an average improvement of *0.55* % of M1 over M2. Similarly for PESQ there is an improvement of *30* % of M1 over M2.

Table 2. *The average MCD score and the PESQ score for 20 utterances from CMU-ARCTIC database* [37].

Source	MCD		PESQ	
	M	F	M	F
M1	4.03734	4.0735	1.08	1.19
M2	**4.014278**	**4.0521**	**1.41**	**1.54**

The correlation coefficient 'ρ' (Pearson's correlation) between the subjective quality ratings S_d (i.e., MOS from Table 1) and the objective measure O_d is given by the following [41] ,

$$\rho = \frac{\sum_d (S_d - \bar{S}_d)(O_d - \bar{O}_d)}{\left[\sum_d (S_d - \bar{S}_d)^2\right]^{1/2}\left[\sum_d (O_d - \bar{O}_d)^2\right]^{1/2}}, \quad (8)$$

where $\bar{S}_d$ and $\bar{O}_d$ are mean values of S_d and O_d, respectively. Table 3 shows that ρ of objective measures improves for the proposed method. In case of MCD (male system) ρ decreases, while for PESQ, ρ increases in both female and male system using the proposed excitation method. On the whole, PESQ scores using the proposed source excitation *correlates well*.

Table 3. *The correlation coefficient ρ of the MCD and PESQ score with the subjective scores of MOS.*

Method	ρ (MCD)		ρ (PESQ)	
	M	F	M	F
M1	-0.10479	-0.24895	0.070451	-0.20424
M2	**-0.25177**	0.042142	**0.09283**	**-0.04846**

6. Summary and Conclusions

In this work**,** we presented a novel chaotic excitation source for synthesizing speech. It is known that the vocal fold movement is chaotic in nature and hence the source of excitation should be chaotic too. The use of NL values from chaotic titration method justifies this. The excitation method is *mixed excitation* (i.e., adding noise to voiced regions as well). However, an interesting future research direction to further improve speech synthesis can be modeling a chaotic system (i.e., for vocal tract) along with the proposed chaotic source that would represent the speech production mechanism more effectively and hence synthesize speech very close to natural.

7. Acknowledgements

The authors thank Department of Electronics and Information Technology (DeitY), New Delhi to carry out the research work (which was partly supported by two DeitY sponsored projects headed by Prof. Hema Murthy and Prof. B. Yegnanarayana) and authorities of DA-IICT, Gandhinagar for their cooperation.

8. References

[1] J. Markhoul, "Linear prediction: A tutorial review," in *Proc. of the IEEE*, vol. 63, no. 4, pp. 561-580, April 1975.

[2] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. on Acoust. Speech and Signal Processing*, vol. 28, no. 5, pp. 599-601, 1980.

[3] H. M. Teager and S. M. Teager, "Evidence for nonlinear production mechanisms in the vocal tract," in *Speech Production and Speech Modelling in NATO ASI Series*, vol. 55, pp. 241-261, 1990.

[4] P. Henriquez and et. al., "Characterization of healthy and pathological voices," *IEEE Trans. on Audio and Speech Process.*, vol. 17, no. 6, pp. 1186-1195, 2009.

[5] A. Mullick and S. Kumar, "Nonlinear dynamical analysis of Speech," *J. Acoust. Soc. Amer.*,vol.100, no.1, pp. 615-628, 1996.

[6] P. Maragos and A. Potamianos, "Fractal dimensions of speech sounds: Computation and application to automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1925-1932, 1999.

[7] M. T. Rosenstein, J. J. Collins, and C. J. De Luca, "A practical method for calculating largest Lyapunov exponents from small data sets," *Physica D*, vol. 65, pp. 117-134, 1993.

[8] J. Thyssen, H. Nielsen, and S. D. Hansen, "Nonlinear short-term prediction in speech coding," in *Proc. IEEE Int. Conf. on Acous., Speech and Signal Process., ICASSP'94*, vol. 1, Adelaide, South Australia, Australia, pp. 185-188, 1994.

[9] V. Z. Marmarelis and X. Zhao, "Volterra models and three layers perception," *IEEE Trans. on Neural Networks*, vol. 8, no. 6, pp. 1421-1433, 1997.

[10] N. Z. Hakim, et. al., "Volterra characterization of neural networks," in *the 25th Asilomar Conf. on Signals, Systems and Computers*, vol. 2, pp. 1128-1132, 1991.

[11] G. Stegmayer, "Volterra series and neural networks to model an electronic device nonlinear behavior," in *Proc. of IEEE Conf. Neural Networks*, vol. 4, pp. 2907-2910, 2004.

[12] G. H. Alipoor and M. H. Savoji, "Speech coding using non-linear prediction based on Volterra series expansion," in *13th Int. Conf. on Speech and Computer (SPECOM)*, St. Petersburg, pp. 367-370, 25-29th June 2006.

[13] J. Pinto, et. al., "Volterra series for analyzing MLP based phoneme posterior probability estimator," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process., ICASSP'09*, Taipei, Taiwan, pp. 1813-1816, 2009.

[14] H. A. Patil and T. B. Patel, "Nonlinear prediction of speech using Volterra-Wiener series," in *INTERSPEECH*, Lyon, France, pp. 1687-1691, 2013.

[15] C. Tao, J. Mu, and G. Du, "Chaotic characteristics of speech signal and its LPC residual," *Acoustical Letter in Acoustical Science and Technology*, vol. 25, no. 1, pp. 50-53, 2004.

[16] T. F. Quatieri, *Discrete-Time Speech Signal Processing*. Prentice-Hall, 2002.

[17] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, no. 2B, pp. 637-655, 1971.

[18] A.V. McCree and T.P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 4, pp. 242-250, 1995.

[19] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 1-16, 1999.

[20] Y. Stylianou, J. Laroche and E. Mouline, "High-quality speech modification based on a harmonic + noise model," in *Proc. EUROSPEECH*, pp. 451–454, 1995.

[21] K. Kelber, et. al, "Synthesis of unvoiced speech phonemes using programmable chaos generators," in *Proc. 6th Int. Workshop on Nonlinear Dynamics of Electronic Systems (NDES'98)*, vol. 1, Budapest, Hungary, pp. 105-108, 1998.

[22] M. Crisan, "New aspects of phoneme synthesis based on chaotic modeling," in *Instrumentation, Measurement, Circuits and Systems*, vol. 127, pp. 605-614, 2012.

[23] G. Kubin, "Nonlinear processing of speech," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), ch. 16, pp. 557–610, Amsterdam: Elsevier 1995.

[24] E. Rank and G. Kubin, "An oscillator-plus noise model for speech synthesis," *Speech Comm.*, vol. 48, no.7, pp. 775–801, July 2006.

[25] J. Lucero, "Dynamics of the two-mass model of the vocal folds: equilibria, bifurcations and oscillation region," *J. Acoust. Soc. Amer.*, vol. 94, no. 6, pp. 3104-3111, 1993.

[26] C.-S. Poon and M. Barahona, "Titration of chaos with added noise," in *Proc. of National Acad. Science*, vol. 98, pp. 7107-7112, 2001.

[27] M. Barahona and C.-S. Poon, "Detection of nonlinear dynamics in short, noisy time series," *Nature*, vol. 381, no. 6579, pp. 215-217, May 1996.

[28] V. J. Mathews and G. L. Sicuranza, *Polynomial Signal Processing*. John Wiley & Sons, 2000.

[29] M. J. Korenberg, "Indentifying nonlinear difference equation and functional expansion representation: The fast orthogonal algorithm,"*Ann. Biomed Engg.*, vol.16, no.1, pp. 123-142, 1988.

[30] M. Frechet, "Surles les fontionelles continues," *Annales Scientifiques de L'Ecole Normale Sup.*, vol. 27, pp. 193-216, 1910.

[31] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, vol. 19, no. 6, pp. 716-723, 1974.

[32] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 23, no. 6, pp. 562–570, Dec. 1975.

[33] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech Lang. Process*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[34] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse," *Speech Comm.*, vol. 11, no. 2-3, pp. 109–118, 1992.

[35] H. Auvinen et. al., "Automatic glottal inverse filtering with the Markov chain Monte Carlo method," *Comp. Speech and Lang.,*vol 28, no. 5, pp. 1139-1155, 2014 .

[36] J. S. Garafolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," in *National Institute of Standards and Technology (NIST), Gaithersburgh*, MD, USA, 1988.

[37] "CMU-ARCTIC Speech Synthesis Database. Online: http://festvox.org/cmu_arctic/index.html (Last Accessed: February 20, 2014),".

[38] "DMOS evaluation. Online: http://www.itu.int/rec/T-REC-P.800-199608-I/en," (Last Accessed: February 23, 2014)).

[39] "Perceptual Evaluation of Speech Quality (PESQ). Online: http://www.itu.int/recT-REC-P.862/ (Last Accessed: February 23, 2014),".

[40] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Conf. on Comm., Computers and Sig. Process.*, Victoria, BC, pp. 125-128, 1993.

[41] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 16, no. 1, pp. 229-228, Jan. '08.