INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Phase Modeling using Integrated Linear Prediction Residual for Statistical Parametric Speech Synthesis

Nagaraj Adiga and S R M Prasanna

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, India
{nagaraj, prasanna}@iitg.ernet.in

Abstract

The conventional statistical parametric speech synthesis (SPSS) focus on characteristics of the magnitude spectrum of speech for speech synthesis by ignoring phase characteristics of speech. In this work, the role of phase information to improve the naturalness of synthetic speech is explored. The phase characteristics of excitation signal are estimated from the integrated linear prediction residual (ILPR) using an all-pass (AP) filter. The coefficients of the AP filter are estimated by minimizing an entropy based objective function from the cosine phase of the analytical signal obtained from ILPR signal. The AP filter coefficients (APCs) derived from the AP filter are used as features for modeling phase in SPSS. During synthesis time, to generate the excitation signal, frame wise generated APCs are used to add the group delay to the impulse excitation. The proposed method is compared with the group delay based phase excitation used in the STRAIGHT method. The experimental results show that proposed phased modeling having a better perceptual synthesis quality when compared with the STRAIGHT method.

Index Terms: All-pass filter, statistical parametric speech synthesis, integrated linear prediction residual, cosine phase.

1. Introduction

Statistical parametric speech synthesis (SPSS) gains significant importance in recent years due to the smooth transition of synthesized speech and its flexibility in adapting to different voices [1]. However, the naturalness of SPSS is not par with concatenative speech synthesis. The main factors for the lack of naturalness are due to the statistical model and the features used for modeling is not sufficient to capture the variations present in the natural speech. Specifically, the phase information is completely ignored while synthesizing speech. The reason for this may be twofold, it is difficult to model the phase due to the phase unwrapping issue. The second reason is the magnitude spectrum, which is perceptually more relevant message information than the phase spectrum. Hence, most of the works are done using only magnitude spectrum.

1.1. Prior art

In recent years, some works are done in modeling the phase component for speech processing applications [2]. These works investigated the usefulness of phase component for speech recognition and speaker verification task [3, 4]. Particularly, in speech coding area, to get the sophisticated excitation signal, phase component is used along with minimum-phase synthesis filter to get the mixed phase characteristics of the speech signal. However, in speech synthesis, more specifically in SPSS, not much exploration is done in modeling the phase component. This is mainly due to the random nature of phase signal and it is not suitable for training directly in hidden Markov model (HMM). Yet, there are some works showed that phase spectrum also has a significant role in the naturalness of synthetic speech. Kawahara *et al.* in [5] showed that adding fixed group delay around the high-frequency region leads to improvement in the naturalness of synthesized speech. In [6–9], even for SPSS, it is shown that phase can be modeled and improvements in the quality of synthetic speech are reported. Further, in recent advances, phase component is used in deep-learning based speech synthesis [10]. In summary, to generate the excitation signal, which attempts to mimic the glottal flow or residual information have to move beyond the minimum-phase assumption and model phase information.

1.2. Our contributions

This paper presents an approach to model the phase information obtained from the cosine phase of an analytical signal, where the analytical signal is computed from the integrated linear prediction residual (ILPR) [11]. This analytical signal is used to model the phase of a source signal. The advantage of the cosine phase of ILPR signal is that it is free from the vocal-tract response and contains only source signal. The phase of a vocal-tract system can be modeled by minimum-phase assumption [12]. Hence, proposed work tries to model the phase of the excitation signal using ILPR signal. In our earlier work, it is shown that modeling the magnitude spectrum of the ILPR signal using residual Mel-cepstral coefficients (RMCEP) is helpful to improve the naturalness of synthetic speech [13]. In this work, the cosine phase signal obtained from the analytical function of ILPR signal is used to improve the perceptual quality of SPSS. Here, modeling of cosine phase is done by assuming it as the output of an all-pass (AP) filter. The filter coefficients (APCs) are estimated by minimizing an entropy based objective function [14]. The obtained APCs are trained in a statistical framework using HMMs. During the synthesis stage, the group delay of impulse excitation is adjusted using APCs obtained from the HMMs. The synthesis quality is compared with zero phase impulse excitation and fixed group delay based excitation used in STRAIGHT. The results show improvement in perceptual quality for the proposed method.

The rest of the paper is organized as follows, the significance of cosine phase of ILPR signal for speech synthesis and modeling of cosine phase using APCs is described in Section 2 and Section 3, respectively. The experimental evaluation of proposed phase modeling and its comparison with other methods are described in Section 4. The paper is finally concluded in Section 5.

http://dx.doi.org/10.21437/Interspeech.2017-587

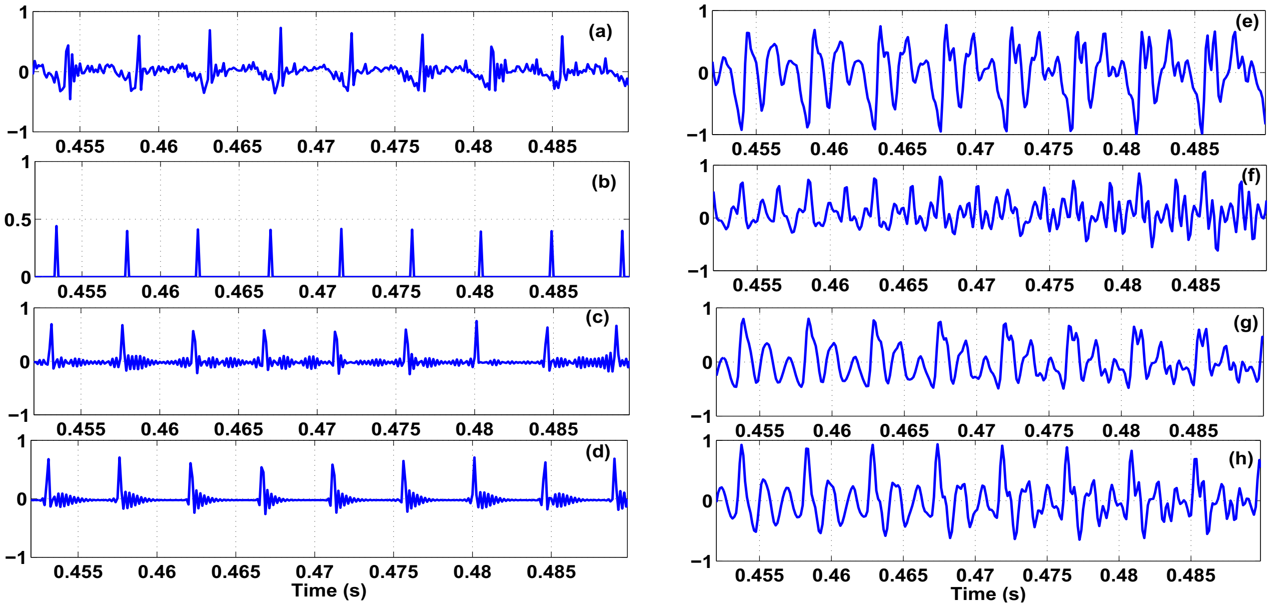

Figure 1: *Comparison of different type of excitation with its synthesized speech: ((a)-(d)) represents the residual, impulse, cosine phase, and group delay phase excitation , respectively; ((e)-(h)) represents synthesized speech for the excitation signal shown in ((a)-(d)), respectively*

2. Phase Modeling using Integrated Linear Prediction residual

In linear prediction analysis, an error signal or residual signal obtained approximately separates the vocal-tract response and contains mostly periodicity, amplitude, and phase information [15]. In addition, the residual source signal obtained still contains high-frequency components due to the pre-emphasis operation. Alternatively, when non-pre-emphasized speech ($s[n]$) is used during the inverse filtering operation, the residual signal obtained is called ILPR signal [11]. It is given by,

$$r_i[n] = s[n] + \sum_{k=1}^{p} a_k s[n-k] \tag{1}$$

where a_k are linear prediction (LP) coefficients obtained from the LP analysis of pre-emphasized speech and p is the order of the LP filter. The source signal obtained is similar to the glottal flow derivative, having quasi-periodic nature and phase information embedded in it. To derive phase information, cosine phase from ILPR signal is computed.

2.1. Cosine phase

Cosine phase is obtained from the ILPR signal by taking Hilbert transform of it to get the analytic signal ($r_a[n]$), which is given by,

$$r_a[n] = r_i[n] + jr_h[n] \tag{2}$$

where $r_h[n]$ is the Hilbert transform of $r_i[n]$. This transform does not change the magnitude of the signal. It just alters its phase, i.e shifts the phase of a positive frequency by -90 and that of negative frequency by +90. Further, to minimize the effect of periodicity and strength of excitation (SoE) from the ILPR excitation signal, cosine phase ($c[n]$) of ILPR signal is computed as follows:

$$c[n] = r_i[n]/h[n] \tag{3}$$

where $h[n]$ be the Hilbert envelope (HE) and it is computed from $r_a[n]$ as follows:

$$h[n] = |r_a[n]| \tag{4}$$

$$h[n] = \sqrt{r_i^2[n] + r_h^2[n]} \tag{5}$$

In a conventional way, STFT phase can be used for modeling phase. However, STFT phase has an unwrapping issue. Further, STFT phase consists of both excitation and vocal-tract phase part. Cosine phase computed from analytic signal represents the phase of ILPR source signal. The obtained cosine phase obtained from the analytic signal contains only the phase component and the periodicity effect present in the residual signal is deemphasized.

2.2. Phase using group delay

To know the significance of the cosine phase, it is compared with the group delay based phase used in the STRAIGHT [16]. Where random phase with a fixed group delay is added to impulse excitation using AP filter. In this process, the random phase is generated and only in higher frequency above some cutoff frequency (default value = 4 kHz), a fixed group delay is added. In [16], it is also shown that the random phase addition to the zero-phase impulse excitation is significant in terms of perceptual quality. Figure 1(d) shows group delay phase excitation obtained by passing impulse excitation with an AP filter. Here, a fixed group delay is added with a standard deviation of 0.5 ms. However, nature of the signal is not similar to LP residual signal shown in Figure 1(a)). Hence, there is a need for phase modeling separately instead of random phase.

2.3. Significance of cosine phase

To illustrate the significance of cosine phase, it is used as an excitation signal by adding it with epoch weighted impulse excitation [17–19]. Here, epoch refers to the glottal closure instant and it is computed from the zero-frequency filter. The epoch weighted impulse excitation is generated as mentioned in our earlier work [20]. Figure 1(c) shows the cosine phase excitation signal obtained after adding cosine phase with the epoch weighted impulse excitation. The figure is shown for the voiced portion of one utterance of SLT speaker taken from the CMU-ARCTIC database and its synthesized speech is shown in Figure 1(g). Here, for synthesis, an excitation signal is passed through LP filter. The excitation signal obtained from the cosine phase is better than group delay phase excitation and even the synthesized speech looks more similar to the synthesized speech obtained from LP residual excitation. Further, to know the significance of cosine phase for synthesis, its quality is evaluated using the perceptual evaluation of speech quality (PESQ). The score for cosine phase excitation signal is shown in Table 1. For comparison, synthesis quality of proposed excitation signal from the cosine phase is compared with the impulse excitation and group delay based phase excitation (Figure 1(h)). From the PESQ score, it can be seen that the perceptual quality of the cosine phase excitation signal is better than both the zero phase impulse excitation and the group delay based phase excitation.

Table 1: *PESQ for different types of excitation*

Excitation signal	PESQ
LP residual	4.5
Impulse	2.83
Cosine phase	3.74
Group delay phase	3.38

3. Analysis and Synthesis Framework

In this section, modeling of cosine phase using the AP filter is described. The AP filter is having unit magnitude response and captures the phase component. Further, obtaining the AP filter coefficients (APC) from the cosine phase signal is shown. Finally, synthesis framework to derive the excitation signal from the filter coefficients is showed to get the mixed phase response of speech signal.

3.1. Analysis

The cosine phase is modeled as an output of an AP filter excited by white Gaussian input sequence. An AP filter response ($H_{ap}[z]$) has a unit magnitude with poles of the system and its zeros are at the complex conjugate reciprocal locations. The filter response is as follows:

$$H_{ap}[z] = \frac{a_N + a_{N-1}z^{-1} + a_{N-2}z^{-2}..... + a_1 z^{-N+1} + z^{-N}}{1 + a_1 z^{-1} + ... + a_{N-1}z^{-N+1} + a_N z^{-N}} \quad (6)$$

where $[a_1, a_2, ..., a_{N-1}, a_N]$ are the desired APC. Further, the cosine phase is modeled as an output of an AP filter $H_{ap}(z)$ excited by Gaussian input sequence $x[n]$.

To estimate both APC and $x[n]$, some prior information about either $x[n]$ or APC are required. In this work, the estimation of APC is done similar to the approach given in [21] with the output of the AP filter being assumed as cosine phase. The energy in $x[n]$ should be concentrated around a few samples. Therefore, the input-output relation between $x[n]$ and $c[n]$ can be written in-terms of APC as follows:

$$c[n] = -\sum_{k=1}^{N} a_k c[n-k] + x[n-N] + \sum_{k=1}^{N} a_k x[n-N+k]. \quad (7)$$

The input signal to the AP filter is obtained by the stable and non-causal inverse filtering of phase signal as reported in [21]. The input signal $x[n]$ can be written as

$$x[n] = -\sum_{k=1}^{N} a_k x[n+k] + c[n+N] + \sum_{k=1}^{N} a_k c[n+N-k]. \quad (8)$$

Here, the energy of both $c[n]$ and $x[n]$ is same since $x[n]$ is passed through the AP filter to give $c[n]$. The energy of the input sample $x[n]$ is given by:

$$e[n] = x^2[n]. \quad (9)$$

In order to concentrate the energy around a few samples, the entropy of $e[n]$ is minimized. The entropy, in turn being a function of APC, can be expressed in terms of an objective function ($J[a_k]$) given by,

$$J[a_k] = -\sum_{n=1}^{N} e[n] log e[n]. \quad (10)$$

To minimize the entropy, the gradient based minimization is carried out with respect to a_k's. The function $J[a_k]$ will be minimized for a particular set of $[a_k]$'s. In this chapter, the gradient descent algorithm is used as mentioned in [14] to minimize the entropy. The APC are iteratively updated until the minimum error is achieved below some epsilon value. In our work, epsilon value is chosen as 10^{-6} determined empirically. All the APC are in the form of a Gaussian function, thus implying that they converge and become suitable for training in HMM.

3.2. Synthesis using cosine phase

In synthesis stage, APC are used to synthesize speech by adding it as a phase to the excitation signal. The fundamental frequency (F0), SoE, and voicing decision are obtained from zero-frequency filter [22, 23]. In the voiced signal, an impulse excitation is generated according to F0 and passed through the AP filter with filter coefficients derived from the modeled APC to get the excitation signal. The resulting signal consists of zero phase impulse excitation signal added with phase. To derive the unvoiced excitation, white Gaussian noise is used. The excitation input signal is convolved with a time-varying filter, which in this chapter is the MLSA filter [24]. The coefficients of this MLSA filter are MCEP obtained from short-term Fourier transform. The block diagram for the synthesis module is shown in Figure 2. In order to get the harmonic and noise representation, in our earlier work [13], magnitude spectrum of ILPR signal is modeled with RMCEP parameters and to model the noise component, white Gaussian noise is weighted by the pitch adaptive triangular envelope. To this framework, cosine phase is integrated.

4. Experimental Evaluation

The evaluation of the proposed phase modeling in SPSS is implemented using open source toolkit named hidden Markov model based speech synthesis system (HTS) [25]. Two speakers from ARCTIC database (1 male BDL speaker + 1 female SLT speaker) were used for evaluation [26]. The database consists of 1132 sentences (around 1.5 hr) for both speakers. During training the HTS system, 1000 sentences were used while remaining 132 sentences were used for evaluation. The parameters F0, MCEP, RMCEP, SoE, and APC are extracted for a frame

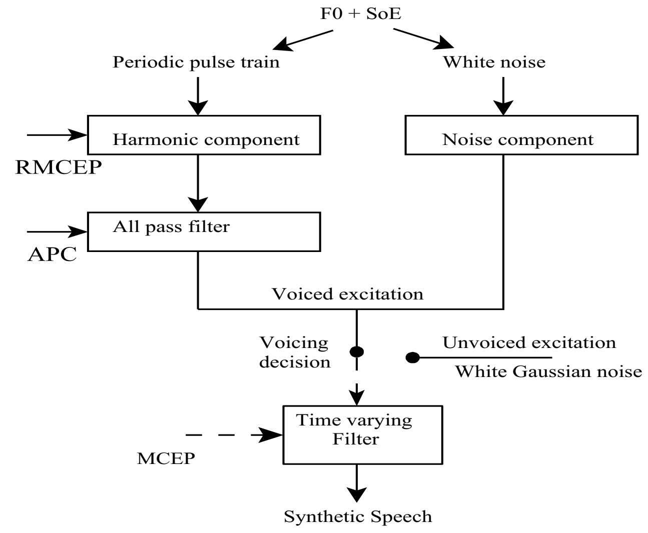

Figure 2: *Integration of AP model to the proposed vocoder*

rate of 5 ms. These parameters are trained in the HMM framework. The extracted parameters for each phoneme is modeled using 5 states, with each state consists of 7 streams. Basic features MCEP and F0 are modeled in first 4 streams. MCEP are modeled as continuous distribution, whereas, F0 is modeled as MSD [27]. In the fifth and sixth stream, RMCEP and SoE along with its derivatives are modeled, respectively. Along with these 6 streams, 20 APC parameters and its derivatives are modeled in the seventh stream to represent the cosine phase of the ILPR signal.

For comparison purpose, impulse excited (with MLSA filter) and STRAIGHT (fixed group-delay based phase with aperiodicity excitation) systems are also developed in the HMM framework. In this work, version 40 of STRAIGHT is used to generate fixed group delay based excitation. To do a fair comparison of proposed phase modeling, in all the methods F0 is extracted from the zero-frequency filter. The synthesized speech from different methods can listen from the following link[1].

4.1. Subjective evaluation

In this evaluation, two tests were conducted, namely, mean opinion score (MOS) and preference test (PT) for all the HTS system. In MOS test, 25 sentences which are not used in training are given to subjects along with the original waveform and asked to give the mean opinion score in the scale of 1 to 5. A total of 10 subjects were used in the subjective evaluation. For evaluations, listeners were asked to examine the naturalness of each file and give their scores accordingly. The average scores obtained from listeners are given in Table 2 along with standard deviation. From the table, it can be seen that the proposed phase modeling outperform the impulse and STRAIGHT method. In the preference test, for each sentence, subjects were requested to listen two versions of the system shuffled randomly from three systems at a time and asked to choose any one system or prefer same as their preference. The percentage of preference scores with p value from listeners can be viewed in Table 2. A clear improvement of the proposed method over zero-phase and random phase methods can be observed from the preference score and the p values given by hypothesis tests.

Table 2: *Subjective evaluation results of MOS and PT for different HTS systems with 95% confidence interval*

Experimental Evaluation	**Different types of phase information**				**p value**
	Impulse	**STRAIGHT**	**Cosine**	**Same**	
MOS	2.71±1.01	3.16±0.92	3.31±0.95	-	
PT	11%	-	80%	9%	2.16×10^{-9}
	26%	64%	-	10%	9.22×10^{-7}
	-	32%	39%	29%	3.67×10^{-2}

4.2. Objective evaluation

In this work, two objective measure is used, namely, PESQ [28] and log spectral distance (LSD) [29]. However, the duration of the synthesized speech and the original speech may not be of the same length. Hence, alignment of the original and the synthesized speech is done using the dynamic time warping algorithm. The PESQ measure should be interpreted as an MOS regarding the similarity to the original waveform. The PESQ scores obtained for all types of phase modeling are tabulated in Table 3. It can be observed from the table that proposed cosine phase model is having a relatively high PESQ score of 1.74 with the standard deviation of 0.03 compared to zero phase excitation and group delay phase excitation, which signifies the improvement in the synthesis quality for the proposed method.

Second objective evaluation is the LSD measure, which gives the distortion error in the spectral domain. This measure is evaluated between reference original speech and the synthesized speech for the same text. The average LSD for all the three methods are given in Table 3 along with standard deviation. The LSD of the proposed phase model has lesser distortion of 1.83, indicating the better phase modeling of the proposed method compared to the impulse and a fixed group delay phase addition.

Table 3: *Objective evaluation results of PESQ and LSD with standard deviation for different HTS systems*

Experimental	**Different types of phase information**		
Evaluation	**Impulse**	**STRAIGHT**	**Cosine**
PESQ	1.26±0.04	1.38±0.06	1.74±0.03
LSD	2.57±0.27	2.35±0.28	1.83±0.21

5. Conclusions

In this work, the role of phase modeling for speech synthesis is analyzed to improve the naturalness of synthetic speech. The phase signal is obtained from the cosine phase of ILPR signal and modeled using the AP filter coefficients. Here, the filter coefficients are optimized using the gradient descent algorithm. Initially, in the analysis-synthesis framework, the significance of cosine phase for speech synthesis is compared with zero phase impulse excitation and group delay based phase excitation. Then the APCs are modeled in hidden Markov model along with excitation and vocal-tract spectrum coefficients. The obtained phase signal from the modeled APCs is compared with zero phase impulse excitation and group delay phase excitation used in the STRAIGHT method. The synthesis quality of the proposed method is better than both the methods. The future work may focus on integrating the phase component in the deep neural framework.

[1] http://www.iitg.ernet.in/cseweb/tts/tts/Assamese/apcshts.php

6. References

[1] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101-5, pp. 1234–1252, 2013.

[2] K. K. Paliwal and L. D. Alsteris, "Usefulness of phase spectrum in human speech perception." in *Proc. Interspeech*, 2003.

[3] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Commun.*, pp. –, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639316000364

[4] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, 2001, pp. 133–136.

[5] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, 1997, pp. 1303–1306.

[6] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP Journal on Audio Speech Music Process.*, no. 1, pp. 1–16, 2014. [Online]. Available: http://dx.doi.org/10.1186/s13636-014-0038-1

[7] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proc. Interspeech*, 2009.

[8] T. Raitio, L. Juvela, A. Suni, M. Vainio, and P. Alku, "Phase perception of the glottal excitation and its relevance in statistical parametric speech synthesis," *Speech Commun.*, pp. –, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639316000303

[9] R. Maia, M. Akamine, and M. J. F. Gales, "Complex cepstrum as phase information in statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, March 2012, pp. 4581–4584.

[10] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016.

[11] A. Prathosh, T. Ananthapadmanabha, and A. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 12, pp. 2471–2480, Dec 2013.

[12] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.

[13] N. Adiga and S. R. M. Prasanna, "Source modeling for HMM based speech synthesis using integrated LP residual," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016.

[14] C.-Y. Chi and J.-Y. Kung, "A new identification algorithm for allpass systems by higher-order statistics," *Signal Process.*, vol. 41, no. 2, pp. 239–256, 1995.

[15] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[16] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Commun.*, vol. 27(3-4), pp. 187–207, 1999.

[17] M. Rothenberg, "Glottal noise during speech," *Quarterly Progress Status ReportSpeech Transmission Laboratory of the Royal Institute of Technology, Stockholm*, p. 1, 1974.

[18] P. J. Murphy, "Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2866–2881, 1999.

[19] D. O'shaughnessy, *Speech communication: human and machine*. Universities press, 1987.

[20] N. Adiga and S. R. M. Prasanna, "Significance of instants of significant excitation for source modeling," in *Proc. Interspeech*, 2013.

[21] K. Vijayan, V. Kumar, and K. S. R. Murty, "Allpass modelling of Fourier phase for speaker verification," in *Proc. Odyssey*, 2014, pp. 112–117.

[22] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, pp. 1602–1613, November 2008.

[23] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Process. Lett.*, vol. 17, no. 3, pp. 273–276, March 2010.

[24] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 8, pp. 93–96, 1983.

[25] HTS. [Online]. Available: http://hts.sp.nitech.ac.jp/

[26] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224. [Online]. Available: http://festvox.org/cmu_arctic/index.html

[27] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, Mar 1999, pp. 229–232 vol.1.

[28] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," in *Proc. European Congress on Acoust.*, 2005, pp. 2725–2728.

[29] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Commun.*, vol. 10-5, pp. 819–829, 1992.