SOURCE MODELING FOR HMM BASED SPEECH SYNTHESIS USING INTEGRATED LP RESIDUAL

Nagaraj Adiga and S. R. Mahadeva Prasanna

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

ABSTRACT

In this work, new method of source modeling for HMM based speech synthesis is proposed using integrated LP residual (ILPR). The nature of ILPR waveform resembles the glottal flow derivative signal and may keep the speaker characteristics in a better way. The ILPR signal is modeled in the frequency domain by dividing the spectrum into two bands to characterize harmonic and noise components of the voice speech segment. The harmonic components of ILPR signals below the maximum voiced frequency (f_m) is modeled using mel-cepstral coefficients called as RMCEPs, whereas noise component above f_m is modeled by pitch adaptive triangular noise envelope weighted by the strength of excitation (SoE). The RMCEPs and SoE are modeled on the HMM framework along with MCEPs and F_0 representing vocal tract information and fundamental frequency, respectively. The synthesized speech by the proposed source modeling reduces the buzziness and improves the speaker similarity compared to the conventional impulse / noise and mixed excitation source modeling and comparable with STRAIGHT based excitation. This is further reflected in both objective and subjective valuations.

Index Terms: Integrated LP residual, residual MCEPs, source modeling, SoE, HTS

1. INTRODUCTION

The HMM based speech synthesis system (HTS) is quite popular in the modern day speech synthesizer due its small footprint and flexibility. However, the main limitation of HMM based speech synthesis is the vocoder framework due to which synthesis quality is still lagging behind the unit selection speech (USS) synthesis [1]. The vocoder compactly represents the acoustic phoneme units and reconstructs phoneme units from such a compact representation. Therefore, vocoder framework is very significant and it can influence the overall voice quality. In the initial version of the HTS, two state source-filter model is used with simple periodic pulse-train or white Gaussian noise as an excitation for source modeling, which generally gives a buzzy quality to the generated speech [2]. Subsequently, a range of high quality vocoders [3, 4, 5] has been suggested to alleviate this problem. Most of these methods focused on improved excitation schemes such as mixed excitation or residual excitation, using some compact representation of excitation, which can be trainable parameters for modeling. Specifically, the STRAIGHT vocoder [6] cannot be integrated with HMMs directly, because it has a large number of parameters. Therefore, Zen *et al.* [7] proposed to convert the features into mel-cepstral coefficients and band aperiodicity in order to use STRAIGHT with HTS. In case of residual excitation, the residue is modeled using voiced and unvoiced filter parameters trained by closed loop analysis [8].

The primary motivation of this paper is modeling the actual glottal flow derivative signal to enhance the naturalness and improve the speaker similarity of the synthesized speech. However, in practice it is challenging to model the source signal. This may be because the source signal in case of voiced speech consists of a harmonic structure in a low-frequency band representing periodic component and noise component in the high-frequency band due to the turbulence of the glottal airflow. There are some attempts to model the residual signal, which is an approximated source signal, for HMM based speech synthesis. In [3], the residual signal is modeled as harmonic component and noise component, representing the deterministic and stochastic (DSM) part of the source signal, respectively. The spectrum of the residual signal is divided into two bands separated by the maximum voiced frequency fm. The lower band below fm is modeled by processing pitch-synchronous residual frames and keeping it as codebooks. The stochastic component above fm is modeled by random noise with its shape is weighted by pitch adaptive triangular window. In GlottHMM, glottal pulses are extracted from real speech via iterative adaptive inverse filtering and stored as a library of pulses, resulting in improved the synthesis quality [9]. However, storing codebook or glottal pulses need separate memory and a complex optimized algorithm is required to select the codebook or glottal pulses for creating excitation [10].

In this work, the actual source signal itself modeled in HMM by parametrization of the source signal. The integrated LP residual (ILPR) is used as a source signal and its time domain waveform is similar to the glottal flow derivative signal [11]. The ILPR signal is parametrized using Mel-cepstral coefficients (MCEPs). However, MCEPs capture only the harmonic content of the ILPR signal and to capture the aperiodic or stochastic component, noise modeling has to be done. Hence, in this work the harmonic noise model (HNM) approach is used to model the spectrum of the ILPR signal by dividing its spectrum into two bands. The lower band representing the harmonic components is modeled by MCEPs instead of keeping codebooks whereas the upper band representing noise component is modeled by a triangular shaped random noise weighted by the epoch strength, which represents the actual strength of random noise around the epoch locations. The proposed source model is evaluated by both subjective and objective evaluation and compared with the impulse/noise, mixed, and STRAIGHT based excitation source model.

The rest of the paper is organized as follows, proposed source modeling for HTS using ILPR is described in Section 2. The integration of the proposed source modeling to HTS framework is explained

This work is part of the ongoing project on the Development of TTS for Assamese and Manipuri languages funded by TDIL, DEiTy, MCIT, GOI. The authors wish to thank Prof. Keiichi Tokuda and Prof. Hideki Kawahara for providing us to use the HTS and STRAIGHT code, respectively.

in the Section 3. The experimental evaluation of proposed source modeling and its comparison with other methods are described in Section 4. The paper is finally concluded in Section 5.

2. SOURCE MODELING USING ILPR SIGNAL

This section describes modeling of the source signal using the ILPR signal. Motivated by the fact that the voiced source signal consists of both harmonic and noise component, ILPR signal is divided into two bands, harmonic and noise component in frequency domain based on voicing frequency (f_m). In the ILPR signal of voiced speech, the frequency component below the f_m contains harmonic components, whereas frequency component above f_m contains the random noise spectrum. Hence, the noise component also has to be modeled to represent the ILPR signal fully. In the rest of the section, the nature of the ILPR source signal and the need for the two band excitation scheme for the ILPR signal into two bands is described.

2.1. ILPR signal

The source signal can be approximately separated from vocal-tract response using LP-based inverse filtering on pre-emphasized speech [12]. However, 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,

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

where a_k are 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 glottal flow derivative, having both quasi-periodic nature and harmonic structure. Fig. 1(b) shows the ILPR signal obtained after passing a non pre-emphasized speech through the inverse LP filter. The signal looks similar to the DEGG signal, which is shown in Fig. 1(a).

The ILPR source signal contains both periodic and aperiodic components in the voiced speech. Fig. 2(a) shows the ILPR signal for a voiced speech segment having quasi-periodic waveform with noise component embedded in it. Further, to know these two components, the ILPR signal is passed through a low-pass and a high-pass filter with a cut-off frequency of voicing frequency (f_m= 4 kHz). The low-pass and high-pass filtered ILPR signal is shown in the Fig. 2(b) and (c), respectively. The low-pass filtered signal retains the periodic nature of voiced signal and turbulence noise is de-emphasized to some extent. The turbulence noise is preserved in the high-pass filtered signal shown in Fig. 2(c). It can be observed that the high-pass filtered ILPR signal consist of noise component synchronized with pitch period of speech and having a variable amplitude due to excitation around the epoch region. The epoch here corresponds to the glottal closure and a small number of excitation instants around them in voiced speech [13]. Due to these events being present in the production of voiced speech, turbulence and frication noise are partially produced at the time instants of opening and closing of the vocal folds [13, 14].

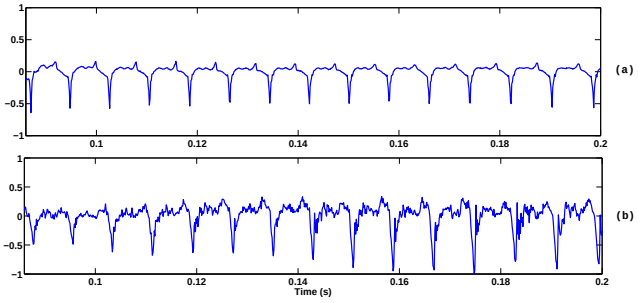

Fig. 1. Source signal representation of a speech segment of voiced regions: (a) and (b) reference DEGG source signal and ILPR signal for a same speech segment, respectively.

Hence, in this work to make ILPR signal suitable for parametrization and then model in HTS, it is modeled with two components, harmonic and noise. The harmonic component $r_h[n]$ represents the periodicity in voiced speech and noise component $r_{no}[n]$ tries to capture the aperiodic nature present in the voiced speech and it is given by

$$r[n] = r_h[n] + r_{no}[n] \tag{2}$$

2.1.1. *Residual MCEPs representing harmonic component*

To represent the harmonic structure of ILPR signal, the residual signal is divided into two bands in frequency domain based on voiced frequency f_m. The lower band of the residual signal below f_m is parametrized using MCEPs in frequency domain. The MCEPs approximates spectrum of signal in the frequency domain with very small error and it is called as RMCEPs in this paper. It captures the harmonic structure of the source signal. The value of voicing frequency (f_m) is fixed in this work to 4 kHz as mentioned in [15]. Moreover, here performance of harmonic representation of ILPR signal using RMCEPs for the source modeling is shown rather than the effectiveness of variable voiced frequency.

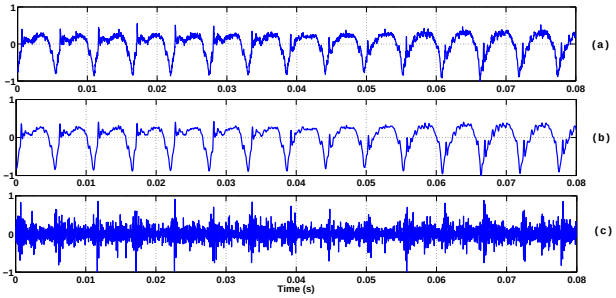

Fig. 2. Periodic and noise component of the ILPR source signal for voiced regions: (a) the ILPR source signal for a segment voiced speech; (b) and (c) periodic and noise component of the ILPR signal obtained by low-pass filtering and high-pass filtering the ILPR signal, respectively.

2.1.2. *Noise component*

The noise modeling of the ILPR signal $r_{no}[n]$ is followed similar to noise modeling done in the HNM model [15]. In the HNM model, it is assumed that white Gaussian noise $b[n]$ is convolved with an auto-regressive model $q[n]$ and its time domain envelope is modulated by weighting function $w[n]$:

$$r_{no}[n] = w[n](q[n] * b[n]) \tag{3}$$

where $w[n]$ is the noise envelope, which is a pitch dependent triangular function, trying to fit the noise component present in the ILPR signal shown in the Fig. 2. Since, fm is fixed to 4 kHz in the proposed method and also the spectrum of the ILPR signal is flat over the entire frequency band, the auto-regressive model has assumed to be having the same effect for all the frames. Hence, $q[n]$ is considered as a high-pass filter (beyond fm) slightly attenuated in the very high frequencies. In this work, instead of using the constant envelope amplitude for triangular function, variable amplitude obtained from the strength of excitation of the ILPR signal around the epoch region is used as envelope amplitude.

2.1.3. Strength of Excitation (SoE)

In the voiced region, due to the rapid movement of vocal fold, significant excitation occurs during the closing of the vocal fold. This results in high strength in the source signal near the epoch location. This can be observed Fig. 2(c), showing the high amplitude around the epoch region for the noise component in the high-pass filtered ILPR signal. Moreover, the amplitude of the noise component around the epoch location is variable, so estimating this amplitude may help in representing noise component in a better way. The strength near an epoch can be obtained from the ILPR signal by passing it through the zero-frequency filter (ZFF) and taking the slope of the filtered signal near the epoch location [16]. The strength of excitation ($s_e[k]$) is defined as the slope of the ZFF signal ($y[n]$) given by:

$$s_e[k] = |(y[k+1] - y[k])|, \tag{4}$$

where k is the epoch location. $s_e[k]$ gives the strength of the impulse-like excitation at the epoch location. The SoE parameter gives a variable amplitude to pitch adaptive triangular noise envelope.

3. ILPR SOURCE MODELING FOR HMM BASED SPEECH SYNTHESIS

The HMM based speech synthesis is experimented with publicly available open source tool kit HTS [17]. In the base version of the HTS, each phoneme is modeled with 5 states and in each state 4 streams are used to model the different features of phonemes. The first stream is used for MCEPs and its derivatives, representing vocal-tract transfer function. The next three streams are used for the fundamental frequency (F_0), its delta, and delta-delta, respectively, to represent the source information. Here, F_0 is modeled in Multispace distribution (MSD), which models, both voiced and unvoiced regions in single model [18]. In this work, along with these 4 streams, RMCEPs and its derivatives are modeled in the fifth stream to represent the harmonic component of the source signal and in the last stream SoE and its derivatives are used, which gives the varying amplitude to noise model. The voice / unvoiced decision to generate excitation is modeled by the weight of MSD, whereas duration is modeled by a single Gaussian distribution for each state. The number of speech parameters used in the training of HMM systems per frame is summarized in the Table 1. To represent the harmonic component, 20 RMCEPs parameters are used in the proposed source model. In addition, one SoE parameter is used to represent the varying amplitude of noise component. During the synthesis, parameters are generated by the maximum likelihood algorithm, as described in [2]. The generated parameters are fed into a proposed vocoder to produce the speech for a given text.

Table 1. Speech parameters used per frame for training the HTS

Feature	Number of parameters
Fundamental frequency (F_0)	1
Strength of excitation (SoE)	1
Residual mel-cepstral coefficients (RMCEPs)	20
Mel-cepstral coefficients (MCEPs)	35

3.1. Proposed source modeling using ILPR signal

A block diagram of the proposed source modeling using the ILPR signal is shown in the Fig. 3. The impulse train is generated according to $F0$ is passed through the low-pass filter and weighted with the residual spectrum generated from RMCEPs to represent the harmonic part of excitation. The noise component $r_{no}[n]$ is generated by high-pass filtering the white Gaussian noise and modulated by a spectrum of pitch adaptive triangular envelope weighted by the SoE. Both harmonic and noise components are added to the spectral domain. The added spectrum of excitation is passed through the Mel-Log Spectrum Approximation (MLSA) filter and then overlap-added to obtain the synthesized speech. In case of unvoiced regions, only white Gaussian noise is used as the excitation. The voice / unvoiced decision is made based on the MSD weight of fundamental frequency.

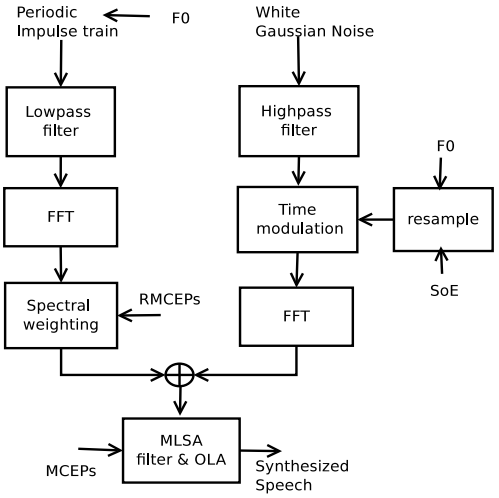

Fig. 3. The work flow of the source modeling using ILPR signal

4. EXPERIMENTAL EVALUATION

To evaluate the proposed vocoder, HTS system is built for two speakers: SLT (US female) and BDL (US male). The speakers SLT and BDL are taken from the CMU ARCTIC database available publicly [19], which consist of 1132 sentences. The first 25 sentences are used for testing and remaining 1107 sentences are used for training. The parameters proposed in the previous sections are analyzed for a frame size of 25 ms with a frame rate of 5 ms and trained in the HMM framework. For the comparison purpose, along with the proposed method, 3 more systems, based on impulse / noise, mixed, and STRAIGHT excitation source model is developed in the HMM framework. In the impulse / noise source model, impulse and white Gaussian noise are used as excitation, for voiced and unvoiced speech, respectively. The mixed excitation is based on a

simple two band excitation for voiced speech with low-pass filtering the impulse train below the voicing frequency (f_m=4 kHz) for the lower band, whereas in the higher band white Gaussian noise is high-pass filtered above the voicing frequency. In the STRAIGHT based excitation, impulse excitation is convolved with bandpass filter weighted by aperiodic components with random phase is added generated from fixed group delay is used as excitation [20]. In all the systems, vocal-tract system is modeled by MCEPs computed on the STFT spectrum of speech. The synthesized files for all the 4 methods can be accessed from the following link [1].

4.1. Subjective evaluation

In this evaluation, two tests are conducted, namely, comparative mean opinion score (CMOS) and preference test (PT). In CMOS 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. For evaluations, 10 people participated and asked them to observe naturalness, speaker similarity, and perceptual distortions present in each file and give their scores accordingly. The average scores obtained from all the subjects are given in the Table 2 along with standard deviation, which is computed for scores present within a 95% confidence interval of the mean. From the table, it can see that ILPR based source modeling outperform the impulse / noise based source model. Moreover, the proposed source model is slightly better than the mixed excitation with CMOS of 3.11, which signifies the addition of the harmonic and the noise component helped in improving the naturalness and speaker similarity. Further, proposed method performs almost similar to STRAIGHT based excitation. The slight degradation may be due to the fact that random phase is not used in the proposed method for excitation, whereas random phase component is added to the excitation with the help of group delay in STRAIGHT method.

In addition, to know the distribution of score for male and female speaker bar chart is plotted in the Fig. 4. The bar plot shows the CMOS with standard deviation of all the 4 systems. From the bar plot, it can see that proposed method for female speaker is significantly better than the mixed excitation, whereas, for male speaker proposed and mixed excitation performs almost similarly. This is due the fact that the number of RMCEPs parameter used for both male and female speakers has a constant value of 20. For male speaker, pitch period is high and more number of harmonics will be present within voiced frequency, increasing the RMCEPs parameter may improve the synthesis quality, however, in this work only fixed RMCEPs are used for comparison purpose. In the preference test, for each sentence subject were asked to listen two system shuffled randomly from 4 systems at a time and asked to choose any one system or prefer none of them as their preference. The percentage of preference scores can be viewed in Table 2. A clear improvement of the proposed method over the traditional impulse / noise and mixed excitation source model can be observed from the table, whereas it perform equally effective with respect to STRAIGHT method.

Table 2. Subjective evaluation results of CMOS and PT

Experimental Evaluation	Source model using different types of excitation				
	Impulse / noise	Mixed	STRAIGHT	ILPR	none
CMOS	2.31±0.28	2.96±0.23	3.15±0.15	3.11±0.17	-
PT	9%	-	-	85%	6%
	-	32%	-	61%	7%
	-	-	45%	42%	13%

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

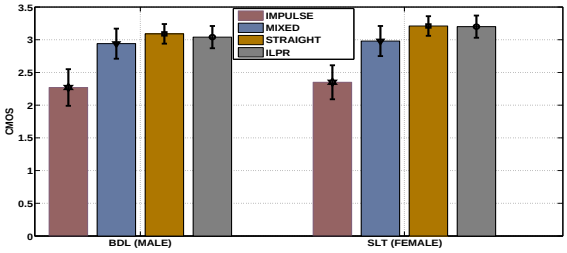

Fig. 4. Average CMOS of 4 HTS systems, impulse / noise, mixed, STRAIGHT and ILPR,respectively,for SLT and BDL speaker

4.2. Objective evaluation

In this work, two objective measure is used, namely, perceptual evaluation of speech quality (PESQ) and log spectral distance (LSD) [21]. The PESQ measure should be interpreted as a MOS regarding the similarity to the original waveform. The PESQ scores obtained for 4 types of source modeling are tabulated in the Table 2. It can be observed from the table that proposed ILPR based source model having a relatively low PESQ score of 1.45 with the standard deviation of ±0.04. However, even the scores obtained by the impulse and mixed excitation source model itself are relatively lower than the proposed excitation, which signifies the improvement in the synthesis quality of the proposed method. The STRAIGHT method performed slightly better than the proposed method, this is due to the fact that phase information is also modeled in STRAIGHT, which is ignored in the proposed method.

The second objective evaluation is the LSD measure, which gives the distortion error in the spectral domain. Note that the distortion is normalized and lower values indicate smaller distortion and better the synthesis quality. This measure is evaluated between the original speech and the synthesized speech for the same text. The average LSD for all the 4 source model are given in Table 2 along with standard deviation. The LSD of the proposed excitation is lesser with distortion of 2.01, indicating the better spectral modeling of the proposed method comparable to that of impulse and mixed excitation. Whereas it performed almost equal with STRAIGHT method.

Table 3. Objective evaluation results of PESQ and LSD

Experimental Evaluation	Source model using different types of excitation			
	Impulse / noise	Mixed	STRAIGHT	ILPR
PESQ	1.21±0.03	1.32±0.04	1.47±0.05	1.45±0.04
LSD	2.20±0.24	2.13±0.25	2.01±0.23	2.03±0.24

5. CONCLUSION

This paper proposes the source modeling for HMM based speech synthesis using ILPR signal. The source modeling is done by dividing the spectrum of an ILPR signal into two bands, harmonic and noise component. The harmonic component is represented by RMCEPs and the noise component by SoE weighted triangular shaped random noise. The proposed ILPR excitation source model is compared with impulse, mixed, and STRAIGHT excitation source modeling. Through the subjective and objective tests, the proposed method was shown to clearly outperform the base version of the HTS system and mixed excitation source model both in terms of naturalness and speaker similarity, gave an almost similar performance with STRAIGHT method. Future work, may focus on modeling more details of the phase information of the excitation signal.

6. REFERENCES

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. 373–376, 1996.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51(11), pp. 1039–1064, 2009.

[3] 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.

[4] R. Maia, T. Toda, H. Zen, Y. Nankaku, and Tokuda.T, "An excitation model for HMM-based speech synthesis based on residual modeling," in *Proc. 6th ISCA Workshop Speech Synth.*, 2007.

[5] J. Nurminen, H. Silen, E. Helander, and M. Gabbouj, "Evaluation of detailed modeling of the LP residual in statistical speech synthesis," in *IEEE International Symposium Circuits and Systems*, 2013.

[6] 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.

[7] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the nitech HMM-based speech synthesis system for the blizzard challenge 2005," *IEICE Trans. Info. Sys.*, vol. E90-D No.1, pp. 325–333, 2007.

[8] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "A trainable excitation model for HMM-based speech synthesis," in *Proc. Interspeech*, 2007.

[9] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19-1, pp. 153–165, 2011.

[10] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Comparing glottal-flow-excited statistical parametric speech synthesis methods," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 7830–7834.

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

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

[14] Y. Pantazis and Y. Stylianou, "Improving the modeling of the noise part in the harmonic plus noise model of speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, March 2008, pp. 4609–4612.

[15] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Process.*, vol. 8, no. 2, pp. 184–194, April 2014.

[16] K. S. R. Murthy, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal processing letters*, vol. 16, no. 6, pp. 469–472, June 2009.

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

[18] 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.

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

[20] H. Kawahara, 2010. [Online]. Available: http://www.wakayama-u.ac.jp/~kawahara/puzzlet/STRAIGHTtipse/'

[21] 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.