Pitch-Scaled Spectrum based Excitation Model for HMM-based Speech Synthesis

Zhengqi Wen¹, Jianhua Tao², Horst-Udo Hain³

^{1,2}National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
³linguwerk GmbH Andreas-Schuber-Str.23 01069 Dresden, Germany
{zqwen¹, jhtao²}@nlpr.ia.ac.cn, udo.hain³@linguwerk.de

Abstract—The quality of speech generated from Hidden Markov Model (HMM)-based Speech Synthesis System (HTS) is suffered from ‘buzzing’ problem which is due to oversimplified vocoding technique. This paper proposed an excitation model to improve the parametric representation of speech in HTS. Residual got from inverse filtering keeps some detailed harmonic structure of speech which has not be included in linear prediction (LP) spectrum. Pitch-scaled spectrum can be used as a supplement of LP spectrum in speech reconstruction. This spectrum is compressed by principal component analysis (PCA) and eigenvalues are indicated as periodic parameter. Then, an aperiodic measure is also extracted from pitch-scaled spectrum and a sigmoid function is fitted to this measure as aperiodic parameter. These two parameters are integrated into HTS training as excitation parameter. Listening tests showed that this proposed technique could generate better sound than pulse train excitation model and take a comparable result with STRAIGHT.

Keywords—HMM-based Speech Synthesis; excitation model; pitch-scaled spectrum; linear prediction; principal component analysis

I. INTRODUCTION

Parametric representation of speech has received great attentions in recent years, especially in statistical speech synthesis [1]. One example is Hidden Markov Model (HMM)-based Speech Synthesis (HTS) [2]. The main degradation of speech quality in typical HTS is oversimplified vocoding technique used in synthesis stage, such as pulse train excitation model.

Some vocoding techniques have been proposed in literature to improve the parametric representation of speech in recent years. In the time domain, Drugman et al. used principal component analysis (PCA) to de-correlate pitch-synchronous residual frame and took the PCA coefficients as excitation information into HTS training [3]. In the frequency domain, Zen et al. incorporated STRAIGHT-based vocoding technique [4] into HTS [5]. Other methods extracted these two domains’ parameters simultaneously. In [6], Cabral et al. incorporated the LF-model [7] into STRAIGHT and used the LF parameters to control the glottal source signal in HTS. In [8], Raitio et al. extracted the vocal-tract parameters and glottal parameters separately and used original glottal source pulses as the source signal when synthesizing. All these methods have shown their ability in improving the generated speech quality of HTS. This

paper will try to keep the detailed harmonic structure of residual in frequency domain.

Residual got from inverse filtering ideally should have a flat spectrum because linear prediction (LP) spectrum has taken the formant structure. But residual spectrum showed in Fig. 1 still keeps some detailed harmonic structure which is not consistent with the pulse train excitation model used in typical HTS system that has a flat spectrum. Pitch-scaled spectrum used by Jackson [10] extracts the harmonic structure in the frequency domain and could be a good supplement of LP spectrum for speech reconstruction. In order to integrate into HTS, principal component analysis (PCA) is adopted to reconstruct the pitch-scaled spectrum and eigenvalues are represented as excitation parameter in HTS training. Besides periodic spectrum, aperiodicity is defined as harmonic-to-noise ratio and sigmoid function [11] is adopted to fit the aperiodic measure based on voicing cut-off frequency [12]. To evaluate the effectiveness of the proposed excitation model, integrated HTS is firstly compared with HTS based on pulse train excitation model and then compared with STRAIGHT-HTS.

The rest of the paper is organized as follows: Section II gives a detailed description about the proposed excitation model. In Section III, this model is integrated into HTS. Experiments are carried out in Section IV. Finally, conclusions and future work are summarized in Section V.

Figure 1. Top: a residual frame of two-pitch periods length; Down: DFT of the residual frame with two-pitch periods length.

II. EXCITATION MODEL

A. Periodic Spectrum

Let $s(k), k=1 \dots N$ be a residual frame of two-pitch period length and the corresponding discrete Fourier transform (DFT) of two-pitch period length is $S(n), n=1 \dots N$.

$$\begin{aligned} N &= 2 \times f_s / f_0 \\ f_k &= f_s \times k / N = f_s \times k / (2 \times f_s / f_0) = f_0 \times k / 2 \end{aligned} \quad (1)$$

where f_0, f_s and f_k is the fundamental frequency, the sampling frequency and the frequency of k th point in $S(n)$.

The even line of $S(n), n=1 \dots N$ in Eq. 1 which takes multiple fundamental frequencies can be indicated as periodic component and the odd line can be indicated as aperiodic component. Fig. 1 shows an example of pitch-scaled analysis for a residual frame (top) and its corresponding DFT spectrum (down).

Periodic spectrum extracted from the even line of pitch-scaled spectrum would take different length and should be interpolated into a constant length in the frequency domain firstly for further interpolation in the time domain. Here a constant length of 128 is used. But it is still too large to incorporate into speech synthesis system. Therefore, dimensionality reduction is necessary. In [11], Fodor listed a number of dimensionality reduction techniques, for example principle component analysis (PCA) [12] is the best linear dimensionality reduction techniques in the mean-square error sense. In this paper, PCA is adopted to de-correlate the periodic spectrums and reduce their dimension. The relative error between original periodic spectrum and reconstructed periodic spectrum with different number of eigenvector is showed in Fig. 2 and 16 eigenvectors can decrease the error into 30.79%.

Figure 2. The error between original periodic spectrum and reconstructed periodic spectrum with the number of eigenvectors increasing.

B. Aperiodic Measure

In [13], Hermus et al. proposed an aperiodic measure based on the ratio between aperiodic energy and periodic energy of two-pitch scaled spectrum in Fig. 1. However, directly estimating the ratio as the solid line showed in Fig. 3 will introduce some vibrating ratios. So we expand the window length and DFT length and adopt two measures to define the aperiodicity.

Let M, N be the window length and DFT length which are both multiple of pitch period. Here 10 multiple of pitch period is used. So harmonic region P_i and noise region D_i can be identified as

$$\begin{aligned} P_i &= \{k \mid k_i - 2N/M \leq k \leq k_i + 2N/M\} \\ D_i &= \{k \mid k_{i-1} + 2N/M \leq k \leq k_i - 2N/M\} \end{aligned} \quad (2)$$

where k_i is i th multiple fundamental frequency which can be easily identified in pitch-scaled analysis and $2N/M$ is the bandwidth of window.

The peak value in one harmonic region together with peak value in the left and right neighboring noise region can construct a triangle. The aperiodic measure is defined as the ratio between an area measure and a symmetric score in Eq. 3. An example is showed as dash-dot line in Fig. 3.

$$\begin{aligned} \text{Symmetry} &= (P_{\text{left}} - P_{\text{right}}) / P_{\text{harmonic}} \\ \text{Area} &= 2 \times (P_{\text{harmonic}} - P_{\text{right}}) - 0.5 \times 2 \times (P_{\text{left}} - P_{\text{right}}) \\ &\quad - 0.5 \times (P_{\text{harmonic}} - P_{\text{right}}) - 0.5 \times (P_{\text{harmonic}} - P_{\text{left}}) \\ \text{Aperiodicity} &= \text{Symmetry} / \text{Area} \end{aligned} \quad (3)$$

where P_{harmonic} , P_{left} and P_{right} are the peak value in one harmonic region, left neighboring noise region and right neighboring noise region.

Figure 3. Dotted line: residual spectrum; Solid line: directly aperiodic-to-period ratio; Dash-dot line: proposed aperiodic measure; Dashed line: sigmoid fitting result.

This aperiodic measure is normalized to 0-1 based on the assumption that strong harmonic structure exists in low-frequency region and pure noisy structure exists in high-frequency region. VCO is defined as the frequency where the aperiodic measure has the maximum slope and can be estimated in Eq. 4. After rough calculation of VCO, time smoothing which adopts the Viterbi algorithm is used to smooth the VCO contour.

$$\text{Dash}(k) = \text{sum}(\text{abs}([0(1,k); 1(1, \text{length} - k + 1)] - Ap)) \quad (4)$$

where Ap is the aperiodic measure line.

Sigmoid function introduced in [10] is used to fit the aperiodic measure in Eq. 5.

$$r(f) = \frac{(f / f_c)^\alpha}{1 + (f / f_c)^\alpha} \quad (5)$$

where α is a transition slope parameter and f_c is boundary frequency parameter.

We make two modifications in Eq. 5. One is to replace f_c by VCO to make the aperiodic measure having a smooth property. The other is to replace the low part of sigmoid function by a parabolic function defined in Eq. 6 in order to reduce the noise in the low-frequency region. An example is showed as dashed line in Fig. 3.

$$r(f) = 0.5 \times (f / f_c)^\beta \quad 0 \leq f \leq f_c \quad (6)$$

where β is a slope parameter which is set as 1.5 in our experiment.

C. Workflow of Proposed Vocoder

The workflow of proposed vocoder is showed in Fig. 4. In analysis stage, input speech is split into single frame and windowed with hanning window. Then every frame is analyzed and represented as LPC, F0, PCA coefficient and aperiodic measure. In synthesis stage, one pitch period spectrum is constructed from the eigenvalue and eigenvector. Every pitch circle of periodic excitation is synthesized by the IDFT of one pitch periodic spectrum with zero-phase criterion and then added together by overlap add (OLA) method to construct periodic excitation. Aperiodic excitation is generated by an IIR filter constructed from aperiodic measure filtering white Gaussian noise. Speech is generated by an all-pole filter filtering the excitation.

Figure 4. The workflow of vocoder based on the proposed excitation model.

III. INTEGRATION INTO HTS

Our speech synthesis system is based on HTS toolkit available in [2]. Some modifications are needed when integrating the proposed excitation model into it.

A. Parameters Preparation

The vocal tract parameter used in Section II is linear prediction coefficient (LPC) extracted from STRAIGHT spectrum. The excitation parameters which are used to reconstruct the residual are PCA coefficient and sigmoid function parameters.

B. HTS Training

Speech features described in Tab. 1 including spectral, F0, periodic and aperiodic parameters, are taken into HTS training in four streams. Spectral stream and aperiodic stream are treated similar to the baseline HTS system as one single Gaussian distribution with diagonal covariance. F0 stream and periodic stream have zero values in unvoiced region and should be treated differently between voiced and unvoiced regions. So multi-space-probability distribution (MSD) is adopted to train these two streams.

Context dependent HMMs have been trained for feature vectors when taking contextual factors into account which

would affect spectral parameters and excitation parameters. These contextual factors included phone identity factors, stressed-related factors, location-related factors and so on [2]. These trained HMMs are clustered by using a tree-based context clustering technique based on minimum description length (MDL) principle.

In synthesis stage, input text is analyzed and transited to a context dependent label sequence which is used to construct a sentence HMM. Parameter generation algorithm is adopted to generate the spectral and excitation parameters. The excitation is generated and speech is synthesized from obtained excitation and spectral parameters as described in Section II.

TABLE I. SPEECH FEATURES AND THE NUMBER OF PARAMETERS IN HTS TRAINING

Feature	PROPOSED	STRAIGHT
F0	1	1
Energy	1	1
Spectrum	24	40
Period	16	0
Aperiod	2	5

IV. EXPERIMENTS

The effectiveness of proposed vocoder in HTS is firstly evaluated by comparing with pulse train excitation model and then comparing with HTS integrated STRAIGHT-based vocoding technique.

A. Database

The database used in these experiments is a female Mandarin corpus lasted about 2 hours. F0 is generated by manual annotating and labeling staffs are asked to listen to the synthesized speech from the annotated F0 contour and then adjust the F0 contour if necessary.

B. Compared with Pulse Train

In this experiment, ten sentences are randomly selected from the database and synthesized based on two versions of excitation models (proposed and pulse train). Ten participants are asked to listen to these synthesized sentences and to give a Mean Opinion Score (MOS) based on Tab. 2.

TABLE II. MEAN OPINION SCORE (MOS)

Ratio	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly Annoying
2	Poor	Annoying
1	Bad	Very Annoying

The results showed in Fig. 7 indicated that the improvement of generated speech's quality from proposed excitation model is obvious when compared with pulse excitation model. Even though increasing the order of spectrum parameter could improve the generated speech quality, but the improvement is very small when compared with the proposed excitation model. This is because the spectrum generated from the spectrum

parameter is a smooth line and proposed excitation model which takes the detailed harmonic structure could enrich more spectral information than directly increasing the order of spectrum parameter.

Figure 5. Mean opinion scores obtained by the two vocoders. Left: synthesized with 24th LPC and pulse excitation model; Middle: synthesized with 40th LPC and pulse excitation model; Right: synthesized with 24th LPC and proposed excitation model with 16th PCA coefficient.

C. Compared with STRAIGHT

STRAIGHT is one of most successfully speech manipulation tool and has shown showed its ability in improving the generated speech quality of HTS. In this experiment, two HTS systems based on proposed vocoder and STRAIGHT showed in Tab. 1 are trained and fifteen sentences lasted about two minutes are synthesized. Ten participants are asked to listen to these two versions of synthesized speech and choose one which is better in this ABX preference test.

Figure 6. Mean preference scores obtained by these two HTS synthesizers in the preference test.

The results showed in Fig. 8 indicated that proposed excitation model could generate even better sound than STRAIGHT-based vocoding technique. This result is consistent with the experiment when compared with pulse train. Directly increasing the order of spectrum parameter could improve the generated speech's quality very little. But if add the spectral information in other ways, e.g. improving the excitation technique and keeping the detailed harmonic structure of residual, the improvement is very prominent [8].

V. CONCLUSIONS AND FUTURE WORK

This paper proposed an excitation model based on pitch-scaled spectrum for HMM-based speech synthesis. This technique is compared with pulse train excitation model and STRAIGHT-based vocoding technique in two experiments. The result is very prominent and this technique could generate

better sound than pulse train excitation model even the order of spectral parameter has increased and also better than STRAIGHT. So it can be concluded that periodic parameter based on pitch-scaled spectrum which has kept the detailed harmonic structure of residual is a good supplement of LP spectrum and this technique provides a new parametric representation of speech for HTS.

However, there are still some shortcomings. One is the representation of pitch-scaled spectrum. In this paper, principal component analysis is used. But it is a linear dimension reduction method. So in our future work, we will dig some nonlinear dimension reduction methods and present the pitch-scaled spectrum more effectiveness in HTS.

ACKNOWLEDGMENT

The work was supported by NSFC-JSPS joint project (No.61011140075), the National Science Foundation of China (No. 60873160 and No.90820303) and China-Singapore Institute of Digital Media (CSIDM). The authors would like to thank Hideki Kawahara.

REFERENCES

- [1] Z. Heiga, T. Keiichi, and W.B. Alan, "Statistical Parametric Speech Synthesis", Speech Communication, vol. 51, no. 11, pp. 1039-1064, 2009.
- [2] [Online], "HMM-based Speech Synthesis System (HTS)", <http://hts.sp.nitech.ac.jp/>.
- [3] T. Drugman, A. Dutoit, and T. Wilfart, "Using a Pitch-Synchronous Residual Codebook for Hybrid HMM/Frame Selection Speech Synthesis," Proc. of ICASSP, pp.3793-3796, 2009.
- [4] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, vol. 27, pp. 187-207, 1999.
- [5] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis for Blizzard Challenge 2005," IEICE Trans. on Inf. and Systems, vol. E90-D, Jan. 2007.
- [6] J. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "HMM-based Speech Synthesizer Using Glottal Source," Proc. of ICASSP, 2011.
- [7] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," STL-QPSR, KTH, Stockholm, 1985.
- [8] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing Glottal Source Pulse Library for Generating Improved Excitation Signal for HMM-based Speech Synthesis," Proc. of ICASSP, 2011.
- [9] T. Toda and K. Tokuda, "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis", IEICE TRANS.INF&SYST, vol.E90-D, NO.5, 2007.
- [10] Jackson, P.J.B. and Shadle, C.H., "Pitch-Scaled Estimation of Simultaneous Voiced and Turbulence-Noise Components in Speech", IEEE Trans. Speech Audio Processing, 9(7):713-726, 2001.
- [11] Kawahara, H., Morise, M., Takahashi, T., Banno, H., Nisimura, R. and Irino, T., "Simplification and extension of non-periodic excitation source representations for high-quality speech manipulation systems", Proc. of Interspeech2010, 2010.
- [12] P. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," IEEE. Trans. Speech and Audio Processing, vol. 15, no. 1, pp. 34-43, 2007.
- [13] K. Hermus, L. Girin, H.V. Hamme, S. Irhimeh, "Estimation of the voicing cut-off frequency contour of natural speech based on harmonic and aperiodic energies", Proc. of ICASSP, 2008.