ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Pitch-Scaled Analysis based Residual Reconstruction for Speech Analysis and Synthesis

Zhengqi Wen[1], *Hideki Kawahara*[2], *Jianhua Tao*[3]

[1,3]National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]Faculty of Systems Engineering, Wakayama University, Wakayama, Japan
{zqwen[1], jhtao[3]}@nlpr.ia.ac.cn, kawahara[2]@sys.wakayama-u.ac.jp

10.21437/Interspeech.2012-136

Abstract

The typical problem in LPC-like vocoder is buzzing sound which is mainly due to the simple pulse train or noise excitation model. One way to improve it is to reconstruct the residual obtained from inverse filtering. So a new parametric representation of speech based on pitch-scaled analysis is proposed in this paper. Pitch-scaled analysis is used to extract the periodic spectrum of residual with half pitch period length. Then these periodic spectrums are de-correlated by principal component analysis (PCA) to reduce their dimension. Aperiodic measure is defined as the harmonic-to-noise ratio in the frequency domain where voicing cut-off frequency (VCO) is used to control the smoothness of aperiodicity. Periodic spectrum and aperiodic measure together with F0 are indicated as excitation parameters in the proposed LPC vocoder. Experimental results show that this proposed vocoder can get a mean opinion score (MOS) of 4.1 for a female voice before dimensionality reduction and keep the high-quality property after parameter compression.

Index Terms: speech parametric representation, pitch-scaled analysis, voicing cut-off frequency, principal component analysis

1. Introduction

Parametric representation of speech has received a great attention in recent years, especially in statistical parametric speech synthesis, e.g., the Hidden Markov model (HMM)-based speech synthesis (HTS, [1]). One of the main degradations in synthesized speech quality of HTS is caused by oversimplified vocoding techniques, e.g., the voices sound buzzing in typical HTS system which is also a typical problem in LPC-like vocoder. It is mainly due to the excitation that is a pulse train or white Gaussian noise for voiced and unvoiced segments, respectively.

There are several high-quality vocoding techniques existing in literature. One is the harmonic plus noise model (HNM) proposed by Yannis Stylianou [2] which has been successfully integrated into concatenating speech system [3] and HTS system [4]. In this technique, spectrum is spilt by voicing cut-off frequency (VCO) into a low-frequency harmonic region and a high-frequency noise region. Another one is STRAIGHT proposed by Hideki Kawahara [5] which also has been integrated into HTS and is state-of-the art HMM-based speech synthesizer [6]. This technique extracts the spectrum without periodic structure both in time domain and frequency domain. Both of these two techniques have kept the detailed harmonic structure of spectrum. In [7], Skoglund et al. mentioned that female voices sound better than male voices in sinusoidal coders where the reconstruction of the harmonic structure of the speech is generally very good, but the pitch-cycle phase is usually modeled with low accuracy. So a way to improve synthesized speech quality for female voice in LPC vocoder is to reconstruct the harmonic structure of residual obtained from inverse filtering.

Pitch-scaled analysis introduced by Jackson et al. in [8], is referred to as an analysis frame that contains a small multiple of pitch period which is different from pitch-synchronous analysis which depends on glottal closure instant (GCI) detection [9]. We will adopt this technique to extract periodic spectrum. Then Principle Component Analysis (PCA) is adopted to de-correlate these periodic spectrums and eigenvalues are kept as periodic parameters. Aperiodic measure is extracted as the harmonic-to-noise ratio (HNR) in the frequency domain and VCO introduced in [2] will be used to control the raucousness of synthesized speech by a sigmoid function [10]. Finally, periodic spectrum and HNR measure together with F0 are indicated as excitation parameters in the proposed LPC vocoder. Experimental results show that the proposed vocoder can generate high-quality speech and take a good property in parameter compression which is very important for the further integration into HTS.

The rest of this paper is organized as follows. Section 2 will give a detailed description of the proposed excitation model for LPC vocoder. In Section 3, experiments will be carried out to testify the effectiveness of the proposed vocoder. Finally, conclusions and future work will be summarized in Section 4.

2. Proposed Vocoder

The excitation model in the proposed vocoder is represented as three parts: F0, periodic spectrum and aperiodic measure. To evaluate the effectiveness of proposed periodic spectrum and aperiodic measure, F0 is generated by manual annotation. The extraction of periodic spectrum and aperiodic measure will be described in the following.

2.1. Periodic Spectrum

To keep the detailed harmonic structure of spectrum, periodic spectrum is defined by concatenating the peak points in the harmonic frequencies of spectrum which can be identified by a peak-searching algorithm. An easy way to extract this envelope is by pitch-scaled analysis introduced in [8].

2.1.1. Pitch-Scaled Analysis

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

$$N = 2 \times f_s / f_0$$
$$f_k = f_s \times k / N = f_s \times k / \left(2 \times f_s / f_0\right) = f_0 \times k / 2 \quad (1)$$

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

The even line of $S(n), n = 1 \ldots N$ 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 and its corresponding DFT spectrum.

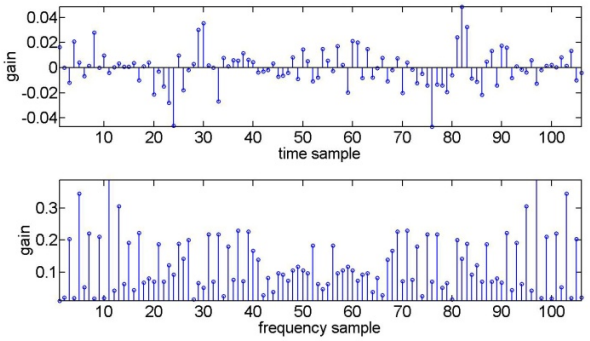

Figure 1: *Top: a residual frame of two-pitch periods length; Down: DFT of the residual frame with two-pitch periods length.*

The residual got from inverse filtering ideally should have a flat spectrum because LPC spectrum has already contained the formant structure. However, the amplitude spectrum showed in Fig. 1 does not reserve this property and has a complicated structure instead. So retaining the half pitch period length of even line of $S(n)$ with multiple fundamental frequencies as periodic spectrum can be a way to keep the detailed harmonic structure of residual.

These periodic spectra would take different length and should be interpolated into a constant length in the frequency domain firstly for the further interpolation in the time domain. Fig. 2 shows the pitch period distribution of 1000 sentences from the female database and the corresponding accumulated ratio. The constant half pitch period length used in this paper is 128 which could cover 99.66% of pitch period values in this database.

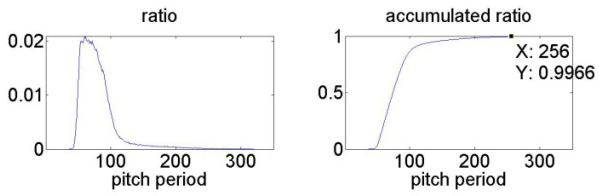

Figure 2: *Left: pitch period distribution of 1000 sentences; Right: the accumulated ratio of the pitch period distribution.*

2.1.2. Principal Component Analysis

Periodic spectrum has been normalized into a constant length of 128 which 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. 3 and 20 eigenvectors can decrease the error into 21.08%. Listening test results also show that this reduction does not produce severe degradation in human perception in Fig. 8.

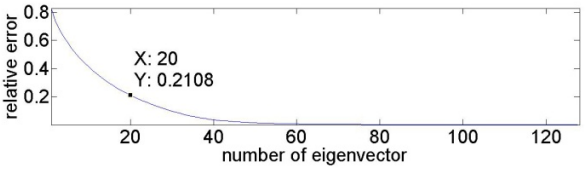

Figure 3: *The error between original periodic spectrum and reconstructed periodic spectrum with the number of eigenvectors increasing.*

2.2. Aperiodic Measure

Aperiodic measure is very important for the naturalness of synthesized speech and is defined as harmonic-to-noise ratio (HNR) in the frequency domain in this paper.

2.2.1. Aperiodic Measure Definition

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. 5 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

$$P_i = \left\{k \mid k_i - 2N/M \le k \le k_i + 2N/M\right\}$$
$$D_i = \left\{k \mid k_{i-1} + 2N/M \le k \le k_i - 2N/M\right\} \quad (2)$$

where k_i is ith multiple fundamental frequency which can be easily indentified 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. An area value and a symmetric score are defined for this triangle in Eq. 3. The aperiodic measure is defined as the ratio between symmetric score with area value and has an increasing property showed as dash-dot line in Fig. 4.

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

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.

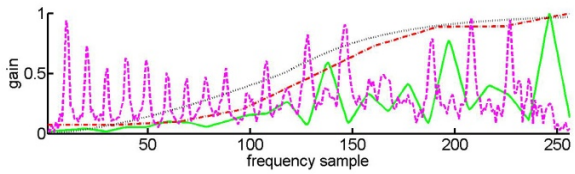

Figure 4: *Dotted line: residual spectrum; Solid line: aperiod-to-period ratio in [13]; Dash-dot line: proposed aperiodic measure; Dashed line: sigmoid fitting result.*

2.2.2. VCO Calculation

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 as the size of dash area showed in the left of Fig. 5 with the minimum value. The dash area is defined in Eq. 4 and one example is showed in Fig. 5.

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

where Ap is the aperiodic measure line.

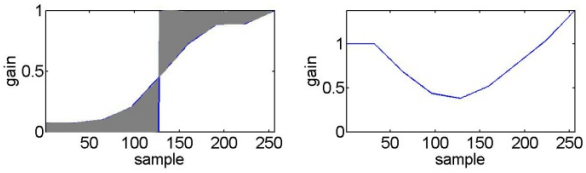

Figure 5: *Left: Aperiodic measure and corresponding dash area. Right: Dash area measure moving from left to right.*

After rough calculation of VCO, time smoothing which adopts the Viterbi algorithm is used to smooth the VCO contour. The target score T_Cost and concatenate score C_Cost used in Viterbi are defined in Eq. 5.

$$\begin{aligned} T\_Cost(i,j) &= Dash(i,j) \\ C\_Cost(j,k) &= \exp\left(abs(j-k)\right)\times\alpha \\ Score(i,j) &= \arg\min\left(T\_Cost(i,j)+C\_Cost(j,k)+Score(i-1,k)\right) \end{aligned} \quad (5)$$

where i is the frame index, j,k are the candidate index and α is used to control the smoothness of the VCO contour. The smoothed VCO contour is calculated with minimum *Score* .

2.2.3. Sigmoid Function Fitting

Sigmoid function introduced in [10] is used to fit the aperiodic measure in Eq. 6.

$$r(f) = \frac{(f/f_c)^{\alpha}}{1+(f/f_c)^{\alpha}} \quad (6)$$

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

We make two modifications in Eq. 6. One is to replace f_c by VCO calculated in Section 2.2.2 to make sure the aperiodic measure has a smooth property. The other is to replace the low part of sigmoid function by a parabolic function defined in Eq. 7 in order to reduce the noise in the low-frequency region. An example is showed as dashed line in Fig. 4.

$$r(f) = 0.5\times(f/f_c)^{\beta} \qquad 0 \le f \le f_c \quad (7)$$

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

2.3. Vocoder

The workflow of proposed vocoder is showed in Fig. 6. Input speech is represented as LPC, F0, periodic spectrum and aperiodic measure. In synthesis stage, every pitch circle of periodic excitation is synthesized by the IDFT of one pitch periodic spectrum and then added together by overlap add (OLA) method. 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.

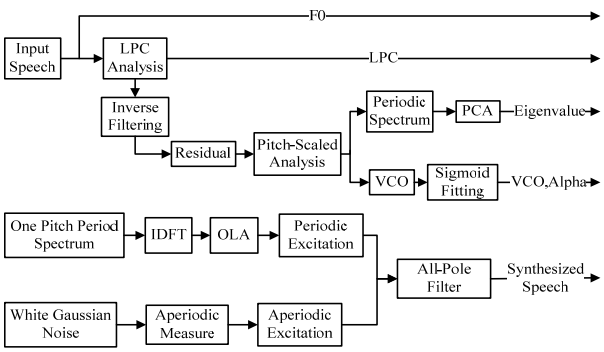

Figure 6: *The workflow of vocoder based on the proposed excitation model.*

3. Experiments and Results

3.1. Proposed Vocoder

To evaluate the effectiveness of proposed vocoder, fifteen sentences lasted about two minutes are randomly selected from a female database and a male database individually. Then these sentences are analyzed and synthesized based on the proposed vocoder without dimensionality reduction and STRAIGHT-based vocoding technique. Ten participants are asked to listen to these two versions of synthesized speech and give out a mean opinion score (MOS, Table 1) to every sentence.

The result of the listening test is showed in Fig. 7. It is indicated from this figure that female voice sounds better than male voice in the proposed vocoder and can get closing performance with STRAIGHT in female voice but not male voice. This result is consistent with the comment mentioned in [7] that keeping the detailed harmonic structure can get better synthesized results for female voice than male voice.

Table 1. *Mean Opinion Score.*

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

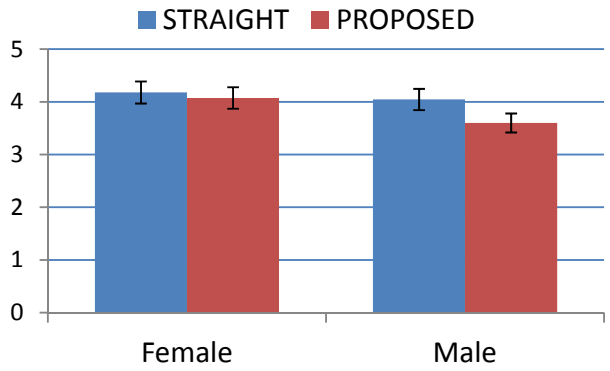

Figure 7: *Mean opinion score obtained from two versions of vocoding techniques for a female database and a male database.*

3.2. Dimensionality Reduction

To evaluate how dimensionality reduction of periodic spectrum influence the speech quality, ten sentences randomly selected from the female database and are analyzed and synthesized in different number of eigenvector showed in Fig. 3. The LPC order is kept as 20 and aperiodic measure is kept same in these experiments. Only the number of eigenvector is changed into 5 groups. Participants are asked to give a mean opinion score again to every sentence.

The result showed in Fig. 8 indicates that the synthesized speech's quality decreases with the number of eigenvector decreasing. But even the number of eigenvector has reduced to 20, the MOS is still closed to 4 and the synthesized speech still sounds in high-quality. It can be concluded that the proposed vocoder can generate high-quality speech in low dimension and take a good property in parameter compression.

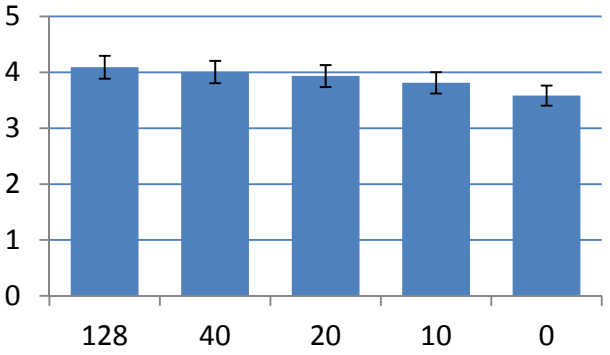

Figure 8: *Mean opinion score obtained from proposed vocoding technique for a female database with different number of eigenvector.*

4. Conclusion and Future Work

In this paper, an excitation model based on pitch-scaled analysis is proposed to reconstruct the residual obtained from inverse filtering for LPC vocoder. Listening test showed that proposed vocoder which has kept detailed harmonic structure of residual could generate high-quality speech for female voice which is comparable with STRAIGHT-based vocoding technique. This vocoder is also efficient for parameter compression. After de-correlation and dimensionality reduction of periodic spectrum by PCA, the synthesized speech quality does not degrade a lot and still maintains in high-quality standard in human perception. This property is very important for further integration into HTS. So in the future work, the effectiveness of proposed vocoder in HTS will be evaluated.

5. Acknowledgements

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).

6. References

[1] [Online], "HMM-based Speech Synthesis System (HTS)", http://hts.sp.nitech.ac.jp/.

[2] Stylianou, Y., "Harmonic plus Noise Model for Speech, combined with Statistical Methods, for Speech and Speaker Modification", P.h.d thesis, 1996.

[3] Stylianou, Y., "Applying the Harmonic plus Noise Model in Concatenative Speech Synthesis", IEEE Trans. Speech Audio Processing, 9(1), 2001.

[4] Hemptinne, C., "Integration of the Harmonic plus Noise Model (HNM) into the Hidden Markov Model-based Speech Synthesis, System (HTS)", Master thesis, 2006.

[5] Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A., "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.

[6] Zen, H., Toda, T., Nakamura, M. and Tokuda, K., "Details of the Nitech HMM-based speech synthesis for Blizzard Challenge 2005," IEICE Trans. on Inf. and Systems, vol. E90-D, Jan. 2007.

[7] Skoglund, J. and Bastiaan, W. K., "On Time-Frequency Masking in Voiced Speech", IEEE Trans. Speech Audio Processing, 8(4):361-369, 2000.

[8] Jackson, P.J.B. and Shadle, C.H., "Pitch-Scaled Estimation of Simultaneous Voiced and Trubulence-Noise Components in Speech", IEEE Trans. Speech Audio Processing, 9(7):713-726, 2001.

[9] Naylor, P., Kounoudes, A., Gudnason, J. and Brookes, M., "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.

[10] 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", Interspeech2010, 2010.

[11] Fodor, I.K., "A survey of dimension reduction techniques", LLNL technical report, 2002.

[12] Jackson, J.E., "A User's Guide to Principal components", New York: John Wiley and Sons, 1991.

[13] Hermus, K., Girin, L., Hamme, H.V. and Irhimeh, S., "ESTIMATION OF THE VOICING CUT-OFF FREQUENCY CONTOUR OF NATURAL SPEECH BASED ON HARMONIC AND APERIODIC ENGRGIES", Proc. of ICASSP, 2008.