Prosody Modification for Vocoder Based on Amplitude Spectrum of Residual Signal

Zhengqi Wen [1], *Jianhua Tao* [2]

[1, 2] National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
zqwen,jhtao@nlpr.ia.ac.cn

Abstract

This paper describes the prosody modification (pitch and duration) for vocoder based on amplitude spectrum of residual signal. In this vocoder, period component is represented as amplitude spectrum of half pitch period length and aperiod component is estimated from the difference of amplitude spectrum between the constructed period signal and the residual signal. Then, pitch modification is conducted by re-sampling the period spectrum according to desired pitch period length in frequency domain and duration modification is conducted by adjusting the frame shift length in time domain. Listening tests show that the speech quality of proposed vocoder after modification is not decreased so much and can get comparable performance with STRAIGHT.

Index Terms: prosody modification, pitch, duration, amplitude spectrum

1. Introduction

It is well-know that prosodic factors such as word stress, phrase accent, phrasal position, and speaking style, have systematic effects on the acoustic effect features of prosody. And these features involve variation in duration, energy, pitch, formant frequencies and so on. The objective of prosody modification is to alter the utterance to the desired prosody features without affecting the shapes of the short-time spectral envelopes [1]. Such technique has been widely used in text-to-speech (TTS) synthesis, voice conversion, expressive speech synthesis, speech rate modification and so on [2, 3].

There are several approaches proposed in the literature for prosody modification [3]. In [4, 5], authors concluded that approaches like overlap and add (OLA), synchronous overlap and add (SOLA) and pitch synchronous overlap and add (PSOLA) operating directly on the waveform to incorporate the desired prosody information are time domain methods [6]. However, these methods operating in the time domain need the prior knowledge of the instants of significant excitation (glottal closure instant, GCI [7-8]) and the quality of modified speech depends on the accuracy of GCI detection. Another group of prosody modification approaches base on parametric representation of speech signal, such as harmonic plus noise model (HNM) [9], sinusoidal model [10], speech transformation and representation using adaptive interpolation of weighted spectrum (STRAIGHT) [11] and so on. All these parametric form methods have shown their ability in reproducing high-quality speech.

In this paper, we will devote our efforts into speech prosody modification in parametric form, especially pitch and duration modification. A vocoder based on amplitude spectrum of residual signal has been proposed in [12]. The speech signal is represented as spectrum parameter (LPC), pitch, period spectrum and aperiod spectrum. Period spectrum is extracted from discrete Fourier transform (DFT) of residual frame of two pitch period length and amplitude spectrum of half pitch period length is enough to conserve the period information. Amplitude spectrum is calculated from the difference of spectrum between the constructed period signal and the residual signal. In this vocoder, pitch modification and duration modification are conducted in frequency and time domains, respectively. In frequency domain, period spectrum is re-sampled to the desired pitch length and excitation frame of two pitch period length is synthesized from the amplitude spectrum based on zero-phase harmonic representation [10]. Then in time domain, frame shift length is expanded or contracted according to the desired duration length when reconstructed the residual signal. Experiments about the quality of speech after modification are firstly carried out. Then the prosody modification of proposed vocoder is compared with that of STRAIGHT-based vocoding technique. Listening tests show that the speech quality of proposed vocoder after modification is not decreased so much and can get comparable performance with STRAIGHT.

The rest of the paper is organized as follows: Section 2 will give a brief description of proposed vocoder based on amplitude spectrum. Prosody modification for the vocoder includes pitch modification and duration modification will be given in Section 3. Experiments are carried out and results are presented in Section 4. Finally, conclusions and future work will be summarized in Section 5.

2. Vocoder Based on Amplitude Spectrum

In source-filter model, speech production can be made of a sound source and a linear acoustic filter [13]. Equ.1 is the model expressed in frequency domain.

$$S(\omega) = D(\omega)G(\omega)V(\omega)R(\omega) \tag{1}$$

where $D(\omega)$ is the Fourier Transform (FT) of an impulse train, $G(\omega)$ is the FT of a glottal pulse, $V(\omega)$ is the vocal tract transfer function and $R(\omega)$ is the radiation characteristic.

In sinusoidal model, speech signals can be decomposed into harmonically-related period component and noise-related aperiod component [9] [10] [14].

$$s(n) = \sum_{k=1}^{K} A_k \cos(\omega_k(n)n + \Phi_k(n)) + e(n) \tag{2}$$

where $s(n)$ is the speech signal, A_k , ω_k and Φ_k are the amplitude, frequency and phase at k th harmonic component, K is the harmonic number and $e(n)$ is the residual signal.

The model that we propose combines the ideas listed above. LPC is firstly extracted from speech signal and an all-pole filter which is indicated as filter in source-filter model. Then residual signal which is indicated as sound source in source-filter model is obtained by inversing filtering the speech signal. After these, sinusoidal model is adopted to reconstruct the residual signal. Residual signal decomposed in

frequency domain will be detailed described in the following paragraphs.

2.1. Period Spectrum

Residual frame of two-pitch period length is extracted and Discrete Fourier Transform (DFT) of two-pitch period length is calculated on this residual frame. The generated amplitude spectrum could be found in Fig. 1.

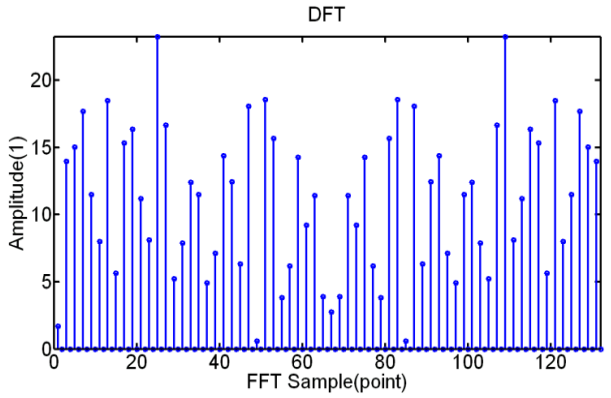

Fig. 1: The amplitude spectrum of a residual frame with two-pitch period length.

The amplitude spectrum showed in Fig. 1 could be divided into two parts: the odd line contains period component and the even line contains aperiod component. The value of even line approximates to zero which indicates this aperiod measure is useless. In addition, the amplitude spectrum shows a symmetrical character. So the amplitude spectrum of half pitch period length is enough to conserve the period information.

2.2. Aperiod Spectrum

Aperiod spectrum is estimated by minimizing the difference of spectrum between the constructed period signal and the residual signal expressed in the following equations.

$$Ap[n] = min\sum_{n=1}^{N}(R[n] - \alpha P[n])^2 \tag{3}$$

$$\alpha = \frac{\sum_{n=1}^{N} R[n]P[n]}{\sum_{n=1}^{N} P[n]P[n]} \tag{4}$$

where $Ap[n]$ $R[n]$ and $P[n]$ are the amplitude spectrum of aperiod component, residual signal and constructed period signal respectively. N is the length of Discrete Fourier Transform.

2.3. Proposed Vocoder

A vocoder based on amplitude spectrum of residual signal can be found in Fig. 2. Input speech is analyzed and LPC, pitch, period spectrum and aperiod spectrum are extracted. In synthesis stage, period frame is synthesized from F0 and period spectrum with zero-phase harmonic representation [10] and OLA is adopted to construct the period signal. Aperiod signal is generated from white Gaussian noise with the aperiod measure. Then these two components are added together as excitation signal to get through an all-pole filter constructed from LPC to generate speech.

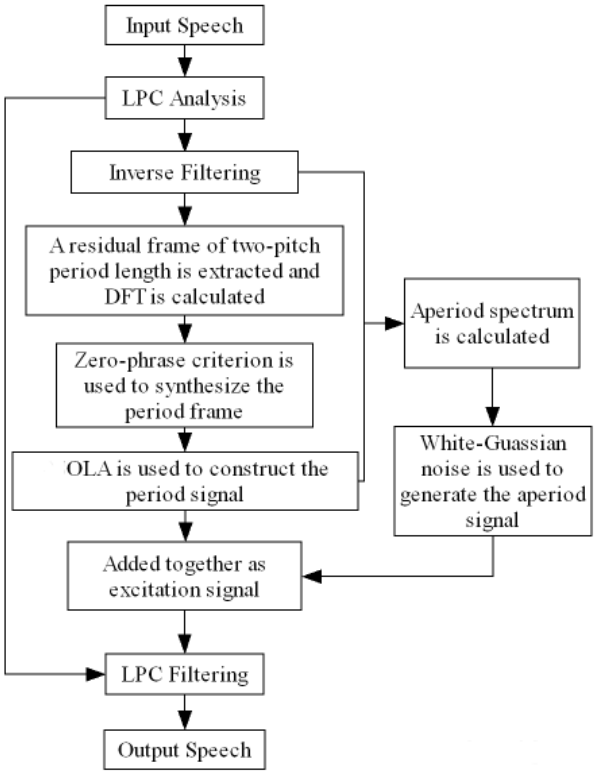

Fig. 2: The workflow of proposed vocoder based on amplitude spectrum of residual signal.

3. Prosody Modification

The acoustic effect features of prosody are consisted of duration, energy, pitch, formant frequencies and so on. And in this paper pitch modification and duration modification are only considered. In frequency domain, pitch period can be easily controlled by the amplitude spectrum length when synthesizing and pitch modification can be conducted just by re-sampling the extracted amplitude spectrum to the desired pitch length. Excitation frame of two-pitch period length has been synthesized and duration modification can be conducted by adjusting the frame shift length when generating excitation signal in time domain. Detailed descriptions can be found in following.

3.1. Pitch Modification

In Fig.1, amplitude spectrum of half pitch period length is reserved as period component and can be re-sampled to the desired pitch period length. Following equation is the re-sampling process.

$$Am\_new(n) = \frac{Pitch\_new}{Pitch\_old} Am\_old(n) \tag{5}$$

where Am_old , Am_new are the extracted amplitude and re-sampled amplitude, respectively.

If the $Pitch\_new$ is larger than $Pitch\_old$, some adjacent points in adjusted amplitude spectrum have the same value. Because there are no interpolation methods used in Euq.3. In this paper linear interpolation is used to solve this problem. The extracted amplitude spectrums are normalized into a const length of 257 in Fig. 3. In this way, the frequency band of this normalized spectrum is extended into $15 \sim \infty$ Hz which is enough to represent the frequency band of human voice.

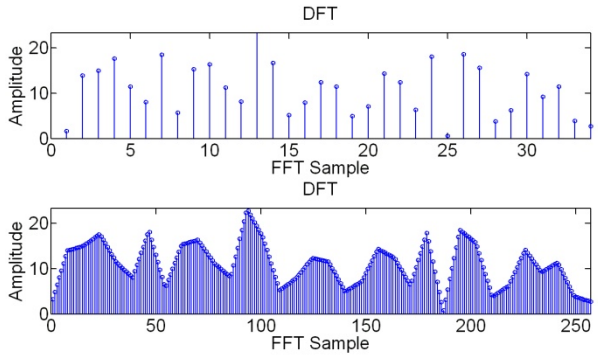

Fig. 3: The half pitch period amplitude spectrum of a residual frame with corresponding normalized amplitude spectrum of 257 points length.

3.2. Duration Modification

In synthesis stage, peak marks are determined according to the pitch contour and frame shift period in Equ.6 and excitation signal is generated based on these peak marks.

$$Peak(n) = \begin{cases} Peak(n-1) + Pitch(i) \\ if\,(Frame(i+1) \geq Peak(n-1) + Pitch(i)) \\ Peak(n-1) + \frac{1}{2}(Pitch(i) + Pitch(i+1)) \\ if\,(Frame(i+1) < Peak(n-1) + Pitch(i)) \end{cases} \quad (6)$$

where $Peak(n), Pitch(i), Frame(i)$ are the n th peak point, i th frame pitch and i th frame start point, respectively.

Duration modification is done by stretching or compressing the length of utterance. Instead of deleting or inserting excitation frame directly, new peak marks can be decided by adjusting the frame shift period in Equ.7 based on Equ.6.

$$Frame\_new = ratio \times Frame\_old \quad (7)$$

4. Experiments

Our experiments are divided into two parts. Firstly, the quality of speech after modification is evaluated. Then the modification performance of proposed vocoder is compared with that of STRAIGHT-based vocoding technique.

The speech sentences used in these experiments are got from a female Mandarin database used for speech synthesis. Pitch contour is generated by manual annotating and labeling staffs are asked to listen to the synthesized speech from the annotated pitch contour and then adjust the pitch contour if necessary. For each sentence the pitch period was modified by two factors, 0.8 and 1.2. Similarly, the duration was modified by two factors, 0.6 and 1.4. In the listening tests, ten participants are ask to listen to two versions of modification speech (Proposed Vocoder, STRAIGHT) and judge the naturalness, distortion and quality of speech for various modification factors according to a five-point scale given in Table 1 [15].

Table 1. *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

Listening test results are given in Fig. 4 and Fig. 5. Fig. 4 shows the Mean Opinion Score (MOS) for each of duration modification factors and Fig. 5 shows the MOS for each of pitch modification factors.

The MOS for duration modification factors is about 3.8 which are corresponding to the original synthetic speech of 4.0. Results show that the quality of speech after duration modification is not decreased so much. This is because in construction of excitation signal for duration modification, only excitation frame is deleted or inserted comparing to original excitation signal and the spectral envelope is not adjusted.

The MOS for pitch modification factors about 3.0 which means the quality of speech after pitch modification is not as good as that after duration modification. This is mainly to that amplitude spectrum for pitch modification has been adjusted. Detailed description for this will be given in the next paragraph when compared to STRAIGHT.

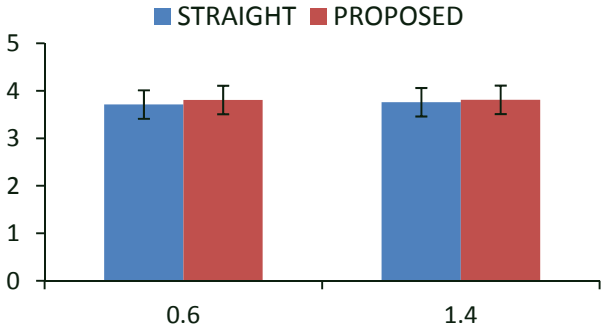

Fig. 4: The Mean Opinion Score (MOS) for the duration modification factors, 0.6 and 1.4.

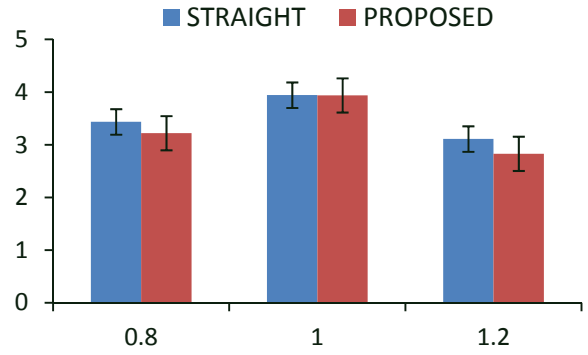

Fig. 5: The Mean Opinion Score (MOS) for the pitch modification factors, 0.8, 1.0 and 1.2.

STRAIGHT is one of the most successfully speech manipulation tool and has showed its power in speech reconstruction. Fig. 6 and Fig. 7 show the preference scores between proposed vocoder and STRAIGHT for each of pitch and duration modification factors, respectively.

In duration modification, our proposed vocoder is much more preferred and gets a higher MOS than STRAIGHT. However in pitch modification, the results are reverse. In our proposed vocoder, amplitude spectrum extracted from residual signal only covers the harmonic-related points for the original pitch period and when the pitch period has been changed, the amplitude spectrum is constructed from the linear interpolation but not from original spectral envelope. In STRAIGHT, the extracted spectrum is a spectral envelope without period information in time domain and frequency domain and when in pitch modification, the extracted amplitudes for harmonics with new pitch period are agreed with the spectral envelope of original speech. So the STRAIGHT can get pitch modification results a little better than our proposed vocoder.

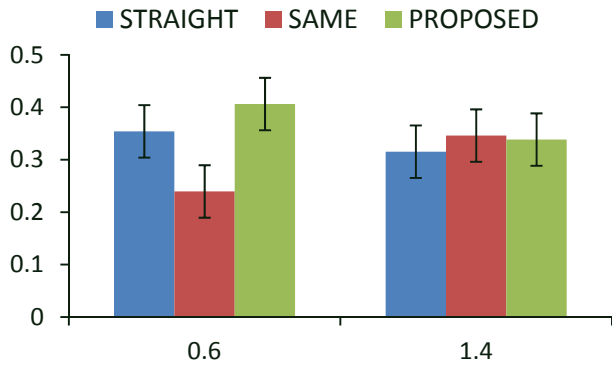

Fig. 6: The preference score between STRAIGHT with proposed vocoder for duration modification factors, 0.6 and 1.4.

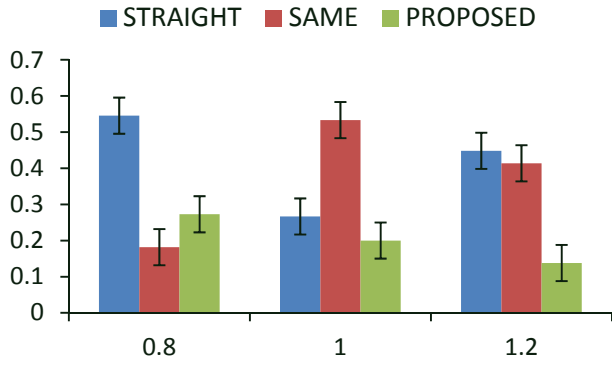

Fig. 7: The preference score between STRAIGHT with proposed vocoder for the pitch modification factors, 0.8, 1.0 and 1.2.

5. Conclusions and Future Work

In this paper, a vocoder based on amplitude spectrum of residual signal is described and prosody modification for it is detailed illustrated. Experiments are conducted to verify that if the quality of speech after pitch and duration modification has been decreased and to compare the modification results with STRAIGHT-based vocoding technique. The MOS valuation show that the quality of speech after prosody modification do not decrease so much and the preference score show that prosody modification of proposed vocoder can get comparable performance with that of STRAIGHT.

In our future work, we will try to modify the prosody of formant frequencies and then modify the emotion of speech based on these techniques and construct an expressive text-to-speech (TTS) system.

6. Acknowledgements

The work was supported by the National Science Foundation of China (No. 60873160 and No.90820303) and China-Singapore Institute of Digital Media (CSIDM).

7. References

[1] Quatieri, T. F. and McAulay, R. J., "Sharp invariant time-scale and pitch modification", IEEE Trans. Signal Proc., 40(3):497-510, 1992.

[2] Childers, D. G., Wu, K., Hicks, D. M. and Yenanarayana, B., "Voice Conversion", Speech Commun., 8:147-158, 1989.

[3] Smits, R., and Laroche, J., "Non-parametric techniques for pitch-scale and time-scale modification of speech", Speech Communication, 16:175-205, 1995.

[4] Rao, K.S., and Yenanarayana, B., "Prosody modification using instants of significant excitation", IEEE Trans. Audio, Speech, Language Proc., 14(3):972-980, 2007.

[5] Prasanna, S. R. M, Govind, D., Rao, K. S., and Yegnanarayana, B., "Fast prosody modification using instants of significant excitation", in Speech Prosody, 2010.

[6] E. Moulines, F. Charpentier, "Pitch-Synchronous WaveForm Processing Techniques for Text-to-Speech Synthesis using Diphones", Speech Communication, Vol.9, pp.453-467, 1990.

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

[8] Murty, K. S. R., and Yegnanarayna, B., "Epoch extraction from speech signals", IEEE Trans. Audio, Speech, Language Proc., 16(8):1602-1613, 2008.

[9] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. diss., Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.

[10] J.M. Robert, and F.Q. Thomas, "Speech Analysis/Synthesis Based on a Sinusoidal Representation", IEEE Transaction on Acoustics, Speech and Signal Processing, vol. 4, no. 34, 1986.

[11] 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.

[12] Z. Q. Wen, and J. H. Tao, "An Excitation Model Based On Inverse Filtering for Speech Analysis and Synthesis," Proc. of MLSP, 2011.

[13] Douglas, B. P., "The Spectral Envelope Estimation Vocoder", IEEE Transaction on Acoustics, Speech and Signal Processing, vol. 29, no. 4, 1981.

[14] P. J. B. Jackson, and C. H. Shadle, "Pitch-Scaled Estimation of Simultaneous Voiced and Trubulence-Noise Components in Speech," IEEE. Trans. Speech and Audio Processing, vol. 9, no.7, pp. 713-726, 2001.

[15] Goodman, D., and Nash, R. D., "Subjective quality of the same speech transmission condition in seven different countries", IEEE. Trans. Communications, vol. 30, No. 4, 1992.