A SPECTRALLY MIXED EXCITATION (SMX) VOCODER WITH ROBUST PARAMETER DETERMINATION

Yong Duk Cho, Moo Young Kim and Sang Ryong Kim

Human & Computer Interaction Lab., Samsung Advanced Institute of Technology
San 14, Nongseo, Kiheung, Yongin, Kyungki, 449-712, Korea
{ydcho, moo, srkim}@saitgw.sait.samsung.co.kr

ABSTRACT

Sinusoidal speech coders have been widely studied for low-bit rate coding around 4 kbit/s. However, the estimation error of the sinusoidal model parameters would seriously degrade the speech quality. In general, the estimation errors are caused by the effects of various types of speech signal or backgound noise. In this paper we propose a sinusoidal speech coder with robust parameter determination methods. They consist of spectro-temporal autocorrelation method for robust pitch determination, frequency shifting method for robust voicing level measurement, and residual-spectrum magnitude coding method for spectral magnitude compensation. From the experimental results, we can find the robustnesses of the proposed techniques. In addition, informal listening test of the synthesized speech confirms the effectiveness of the incorporated schemes.

1. INTRODUCTION

Sinusoidal speech coders, such as the multiband excitation (MBE) vocoder [1], the sinusoidal transform coder [2], and the prototype waveform interpolation coder [3], have been widely used for low-bit rate coding. The widespread use of these coders are due to the fact that the speech quality of these coders is comparable to that of mid-rate code-excited linear predictive (CELP) coders. It can, however, be only successful under the condition where the model parameters are estimated accurately. In other words, the estimation errors for the model parameters would seriously degrade the speech quality. In general, the estimation errors are caused by the effects of various types of speech signal or background noise.

In this paper, we propose a novel sinusoidal speech coder, spectrally mixed excitation (SMX) vocoder, in which we focus on robust parameter estimation. The proposed SMX vocoder represents the speech signal with both spectral envelop and mixed excitation in frequency domain. The excitation spectrum is represented by a mixture of harmonic and random spectra at each harmonic frequency bin. The core parameters for SMX are 1) pitch used for the sampling of harmonic frequencies, 2) voicing levels for each spectral band to control the spectral mixture, 3) linear predictive coding (LPC) coefficients to represent baseline-spectral envelop, and 4) residual-spectrum magnitude at each harmonic frequency in order to compensate for the inaccuracy of the LP-spectral envelop.

Accurate pitch determination is considered to be most important when developing the SMX coder since the other model parameters, such as the voicing levels and spectral magnitudes, are based on it. The spectro-temporal autocorrelation method where the autocorrelation values in time and frequency domains are simultaneously applied, is proposed for robust pitch determination.

The SMX splits a spectral band into four sub-bands and different voicing levels are assigned to each band. The voicing levels control the mixture ratio of the harmonic and random spectra for each corresponding sub-band. By these parameters we can reduce buzzy or hoarse sound which can be frequently produced by low-bit rate speech coders. In order to determine robust voicing level parameters, we apply the frequency shifting method in which each sub-band is shifted to the origin and then the normalized autocorrelation values are computed.

The original spectral envelop is decomposed into LP-spectral envelop and residual-spectrum magnitude. The residual-spectrum magnitude is defined by the difference between the original and LP-spectral envelopes at each harmonic frequency. The use of residual-spectrum magnitude is found useful for enhancing the highly pitched speech.

2. SMX VOCODER STRUCTURE

The SMX vocoder is mainly composed of the following components; the pitch determination, voicing levels decision, LP-analysis, residual spectrum derivation, and speech synthesis parts. The overall encoder structure is shown in Figure 1.

2.1 ROBUST PITCH DETERMINATION

In SMX, pitch determination is done in three steps as shown in Figure 2. The first step is to expand the formant bandwidth as a preprocessing to reduce the effect of the first formant in pitch determination. The following is an integer pitch determination procedure using the spectro-temporal autocorrelation (STA) method which is effective in reducing gross pitch errors. The last is fractional pitch search to refine the pitch resolution.

In the process of pitch determination, pitch errors may occur when an integer multiple of the fundamental frequency is near the first formant location. In order to alleviate this problem, many algorithms have used the inverse filtering approach. However, this method would remove not only the first formant but also the harmonic structures in residual signal especially for the highly pitched speech signals. Formant bandwidth expansion [4] is applied before pitch determination for the reason that it is useful to remove the effect of first formant while keeping the harmonic structures.

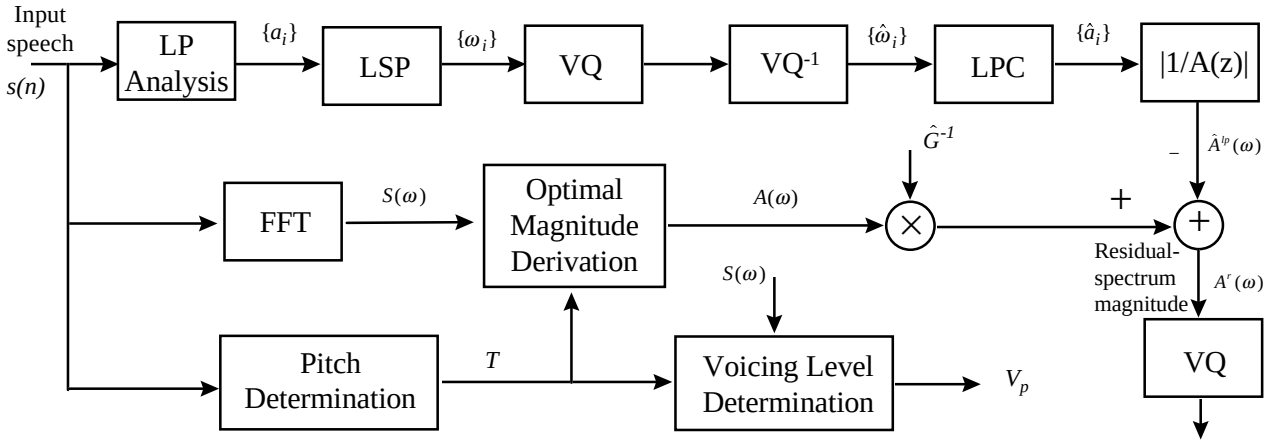

Figure 1. Block diagram of SMX vocoder.

It takes advantage of the interpolated signal $s_f(n)$ between the speech and residual signals as follows:

$$S_f(z) = \frac{A(z)}{A(z/\gamma)} S(z), \quad (1)$$

where $S(z)$ and $A(z)$ are the z-transforms of the speech signal $s(n)$ and the inverse filter, respectively, and γ is a weighting factor of interpolation. If $\gamma = 1$, the filtered signal is the same to the original speech signal. On the other hand, $\gamma = 0$ makes the filtered signal equal to the LPC residual of $s(n)$. It is not difficult to see that $S_f(z)$ is the interpolated spectrum between the original and residual spectra when $0 \le \gamma \le 1$. In our experiments, we used $\gamma = 0.8$.

To diminish gross pitch errors such as pitch doubling or halving, the STA method is proposed. STA is defined by the weighted summation of the temporal and spectral autocorrelation values.

The temporal autocorrelation (TA) method, simply called as the autocorrelation method, is widely used for pitch determination in time domain. Given an interpolated speech signal $s_f(n)$, the TA at a candidate pitch τ is defined as

$$R^T(\tau) = \frac{\sum_{n=0}^{N-\tau-1} \tilde{s}_f(n)\tilde{s}_f(n+\tau)}{\sqrt{\sum_{n=0}^{N-\tau-1} \tilde{s}_f^{\,2}(n) \sum_{n=0}^{N-\tau-1} \tilde{s}_f^{\,2}(n+\tau)}}, \quad (2)$$

where $\tilde{s}_f(n)$ is zero-mean speech signal of $s_f(n)$, and N is the number of samples for pitch determination. Even though the TA method produces exact pitch values in most cases, it may result in doubled pitch error when the speech signal is highly periodic with short pitch period. For the given true pitch interval τ, TA may have high autocorrelation not only at the lag τ, but also at the integer multiples of τ such as 2τ, 3τ, etc.

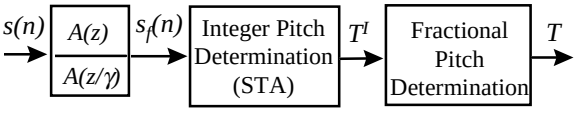

Figure 2. Block diagram of pitch determination.

To compensate for the doubled pitch error, the spectral autocorrelation (SA) method [5] is introduced, and it is shown that the SA method is very useful for removing doubled pitch error in MBE vocoder. Given a magnitude spectrum $S(\omega)$ of the speech signal $s(n)$, the SA at a candidate pitch τ is obtained by

$$R^S(\tau) = \frac{\int_0^{\pi-\omega_\tau} \tilde{S}_f(\omega)\tilde{S}_f(\omega+\omega_\tau)d\omega}{\sqrt{\int_0^{\pi-\omega_\tau} \tilde{S}_f^{\,2}(\omega)d\omega \int_0^{\pi-\omega_\tau} \tilde{S}_f^{\,2}(\omega+\omega_\tau)d\omega}}, \quad (3)$$

where $\omega_\tau = 2\pi/\tau$ and $\tilde{S}_f(\omega)$ is a zero-mean spectrum of $S_f(\omega)$. In spite of its effectiveness in avoiding pitch doubling, the SA method would cause pitch halving, i.e. the original pitch τ could be wrongly found in integer divisions such as $\tau/2$, $\tau/3$, etc.

Hence, the STA method is designed to be robust against gross pitch errors such as pitch doubling or halving. Given a pitch candidate τ, the STA is defined by

$$R(\tau) = \beta R^T(\tau) + (1-\beta)R^S(\tau), \quad (4)$$

where β is a weighting factor in the range of 0 and 1. $R(\tau)$ is computed over a 40 sec. of sample speech where $\beta = 0.5$ yields the lowest pitch error as shown in Figure 3.

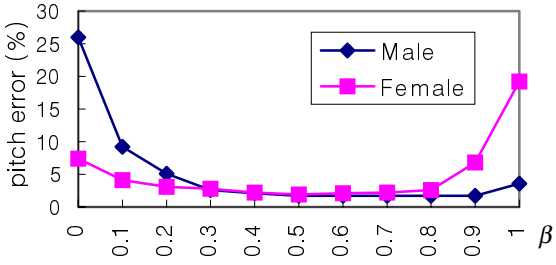

Figure 3. Pitch error comparison in accordance with the weighting factor β.

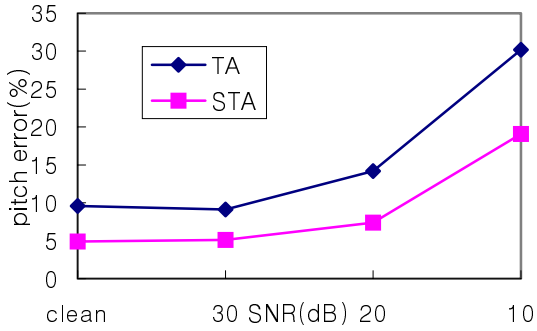

Figure 4. Pitch error comparisons between the temporal autocorrelation (TA) and the spectro-temporal autocorrelation (STA) methods.

Finally, the optimal integer pitch T^I is determined as

$$T^I = \arg\max_{\tau}\{R(\tau)\}. \qquad (5)$$

The performance of the integer pitch determination algorithm is evaluated both for the clean and noisy speech corrupted by the vehicle and babble noises. Figure 4 shows the experimental results which clearly indicate the superiority of the proposed method compared to the conventional TA method.

It is generally known that the fractional pitch is crucial for coding highly pitched speech usually observed in female and children's voice. In SMX, fractional pitch determination is done by the spectral analysis-by-synthesis method proposed by Griffin [1]. We use a half sample period for the fractional pitch resolution.

2.2 VOICING LEVEL DETERMINATION

Traditionally, sinusoidal speech coders assign voicing levels for each harmonic component. In the proposed structure, the voicing level for each harmonic is selected in the range between 0 and 1. Our first attempt to determine the voicing level was based on the bandpass filtering approach. Input speech was bandpass filtered, the normalized autocorrelation value was obtained for each sub-band, and finally the voicing level was determined. However, the above bandpass filtering method would result in unreliable voicing level due to the fluctuation of the autocorrelation value in higher sub-bands as shown in Figure 5(c).

For the purpose of improving the voicing level determination performance in higher spectral sub-bands, frequency shifting (FS) method is proposed. Given a spectrum, $S(\omega)$, the frequency-shifted spectrum $S_b(\omega)$ of the b-th spectral sub-band is as follows:

$$S_b(\omega) = S(\omega - f_b\omega_0), \quad 0 \le b \le B-1, \qquad (6)$$

where B is the total number of spectral sub-bands, ω_0 is the fundamental frequency and f_b is the lowest harmonic frequency of the b-th spectral sub-band, i.e. $f_b = \lfloor \pi b/(\omega_0 B) + 0.5 \rfloor$. By applying Goertzel's inverse discrete Fourier transform (IDFT) to the shifted power spectrum, it is possible to efficiently obtain more reliable autocorrelation for the higher spectral sub-bands, which is shown in Figure 5(d). The overall procedure is shown in Figure 6.

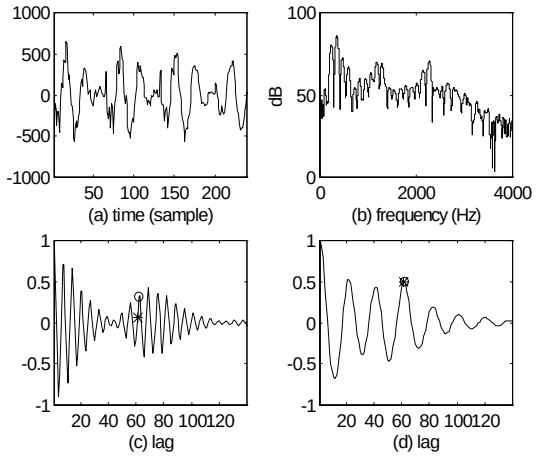

Figure 5. (a) original speech signal, (b) magnitude spectrum, (c) normalized autocorrelation of bandpass filtered signal, and (d) normalized autocorrelation of frequency shifted signal for a spectral sub-band (2000 - 3000 Hz).

The autocorrelation values obtained by the bandpass filtering and the FS methods are shown in Figure 5 where the true pitch period is 62 samples. If the pitch estimation is correct, the autocorrelation values of the bandpass filtering and FS methods are the same as marked with "∘". However, as we can see from the points marked with "∗", when fine pitch determination error occurs due to background noise or algorithmic inaccuracy, the autocorrelation value is far from the true one in case of the bandpass filtering method while it is nearly the same in the case of the FS method.

2.3 SPECTRAL REPRESENTATION AND QUANTIZATION

The envelop for the magnitude spectrum sampled at each harmonic frequency can be characterized by the LPC coefficients and the corresponding gain. However these two sets of parameters are not sufficient in reconstructing high-qualified speech, even though used without quantization. Large spectral differences around formants, especially for the first formant, are mainly responsible for this. In order to make more detailed spectral representation, we also use the residual spectrum as an additional coding parameter. The residual spectrum is derived by the difference between the original and the LP-spectral envelops for each harmonic component. It is known that lower harmonics are far more important than the higher ones [6]. On this basis, the LPC parameters and the lower six residual-spectrum magnitudes are quantized by a 14th-order linked-split vector quantizer (LSVQ) [7] and a 6th-order split VQ, respectively.

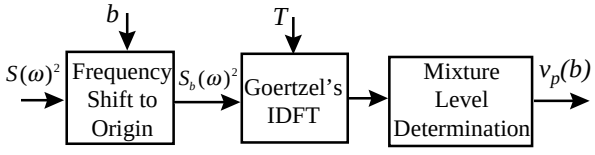

Figure 6. Block diagram of voicing level determination by frequency shifting method.

2.4 SPEECH SYNTHESIS WITH MIXED EXCITATION

With the SMX vocoder, we can represent the synthesized spectrum $\hat{S}(\omega)$ as

$$\hat{S}(\omega) = \hat{A}(\omega)\hat{E}(\omega), \quad (7)$$

where $\hat{A}(\omega)$ and $\hat{E}(\omega)$ are the decoded magnitude and excitation spectra, respectively. The spectral magnitude can be obtained from the quantized LP-spectrum $\hat{A}^{lp}(\omega)$, quantized residual spectrum $\hat{A}^{r}(\omega)$ and quantized gain $\hat{G}$ as follows:

$$\hat{A}(\omega) = G\left\{\hat{A}^{lp}(\omega) + \hat{A}^{r}(\omega)\right\}. \quad (8)$$

On the other hand, the synthesized excitation spectrum $\hat{E}(\omega)$ is computed by

$$\hat{E}(\omega) = \sqrt{\hat{V}_p(\omega)}P(\omega) + \sqrt{1-\hat{V}_p(\omega)}R(\omega), \quad (9)$$

where $P(\omega)$ is a harmonic spectrum to model voiced harmonic structure, $R(\omega)$ is a random spectrum to model unvoiced or noisy spectrum, and $\hat{V}_p(\omega)$ is a decoded voicing level for each harmonic frequency ω. Since the voicing levels are available only for each spectral sub-band, some interpolation technique is needed to get the levels for every harmonic frequency. As an interpolation technique, simple linear interpolation is used within a frame. Smooth temporal voicing level change is also archived using the inter-frame linear interpolation method.

3. IMPLEMENTATION

The SMX vocoder is implemented at 4 kbit/s with 20 ms frame size and 15 ms lookahead, and the bit assignment is shown in Table 1.

From the informal listening test, it is found that the speech quality of the proposed coder is comparable to that of 8 kbit/s CS-ACELP [8] over noisy environment, nonetheless, slightly worse over clean speech signal.

Table 1. Bit assignment of the 4 kbit/s SMX vocoder.

Parameters	Bits
Pitch	8
Voicing rate	12
Gain	6
LSP	36
Residual-spectrum magnitude	16
CRC	2
Total	80 / 20 ms

4. CONCLUSIONS

In this paper, SMX vocoder is proposed which is featured by spectro-temporal autocorrelation method for robust pitch determination, frequency shifting technique for robust voicing level measurement, spectral envelop quantization with LSVQ and its residual-spectrum magnitude coding for elaborated spectral magnitude, and speech synthesis for the mixed excitation source. The SMX is implemented at 4 kbit/s with 20 ms frame size and 15 ms lookahead. The speech quality of the proposed coder is compared with CS-ACELP, and shows comparable and slightly worse performances over noisy and clean speech signals, respectively.

ACKNOWLEDGEMENT

The authors thank Dr. Nam Soo Kim and Dr. Hong Kook Kim of SAIT for their invaluable reviews and comments on this paper.

REFERENCES

[1] D. Griffin and J. S. Lim, "Multiband Excitation Vocoder," *IEEE Trans. on ASSP*, Vol. 36, No. 8, pp. 1223-1235, Aug. 1988.

[2] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. on ASSP*, Vol. 34, No. 4, pp. 744-754, Aug. 1986.

[3] B. Kleijn, "Encoding Speech Using Prototype Waveforms," *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 4, pp. 386-399, Oct. 1993.

[4] R. Matmti, M. Jelinek, and J. P. Adoul, "Low Bit Rate Speech Coding Using an Improved HSX Model," Proc. *EuroSpeech*, pp. 1287-1290, Rhodes, Greece, Sep. 1997.

[5] Y. D. Cho, H. K. Kim, M. Y. Kim, S. R. Kim, "Pitch Estimation using Spectral Covariance Method for MBE Vocoder," *1997 IEEE Workshop on Speech Coding*, pp. 21-22, Pocono Manor, PA, Sep. 1997.

[6] I. Atkinson, S. Yeldener, A. Kondoz, "High Quality Split Band LPC Vocoder Operating at Low Bit Rates," Proc. *IEEE ICASSP*, Munich, Germany, Apr. 1997.

[7] M. Y. Kim, N. K. Ha, S. R. Kim, "Linked_Split Vector Quantization of LPC Parameters," Proc. *IEEE ICASSP*, Vol. 2, pp. 741-744, Atlanta, GA, 1996.

[8] R. Salami, C. Laflamme, J. P. Adoul, D. Massaloux, "A Toll Quality 8 kb/s Speech Codec for the Personal Communications System (PCS)," *IEEE Trans. on Vehicular Technology*, Vol. 43, No. 3, Aug. 1994.