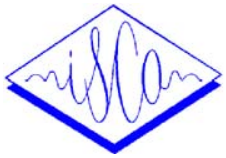

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

JOINT ESTIMATION OF PITCH, BAND MAGNITUDES AND V/UV DECISIONS FOR MBE VOCODER

Yong Duk Cho[1], Hong Kook Kim[2], Moo Young Kim[3], and Sang Ryong Kim[4]

Human & Computer Interaction Lab., Samsung Advanced Institute of Technology
San 14, Nongseo-Ri, Kiheung-Eup, Yongin City, Kyungki-Do, 449-712, Korea
{ydcho[1], kimhk[2], moo[3], srkim[4]}@saitgw.sait.samsung.co.kr

10.21437/Eurospeech.1997-190

ABSTRACT

The multiband excitation (MBE) vocoder represents speech signal with a pitch, band magnitudes, and a voice / unvoice (V/UV) decision for each spectral band. In the conventional MBE model, model parameters are sequentially estimated in two steps. The pitch and band magnitudes are firstly estimated on the assumption of voiced speech model by the analysis-by-synthesis (AbS) in frequency domain, and then the V/UVs are decided. However, the synthetic spectrum by the above assumption may have large spectral distortion if the speech frame is strongly unvoiced such as transient region.

In this paper, we propose joint estimation method which estimates and decides all the model parameters in AbS loop. For this, voiced or unvoiced speech models for each band are used during the analysis procedure. After estimating the parameters with the two speech models, a model for each band is selected so as to produce smaller spectral estimation error. By analyzing the short time spectrum and the long time spectrogram, it is shown that the reproduced speech of the proposed model is superior to that of the conventional one. In addition, through informal listening test we also confirm the superiority of the proposed model.

1. INTRODUCTION

In low bit rate speech coding, sinusoidal speech coders such as multiband excitation (MBE) vocoder [1] and sinusoidal transform coder (STC) [2] are getting more and more interest, and widely known that they reproduce highly qualified speech signals. Particularly the MBE vocoder has been adopted at INMARSAT-M, APCO/NASTD/Fed Project 25, and INMARSAT Mini-M [3] [4].

The MBE vocoder represents speech signals with a pitch, band magnitudes and voice/unvoice (V/UV) decisions. The pitch and band magnitudes are estimated by analysis-by-synthesis (AbS) which assumes all bands are voiced and minimizes the estimation error between the original and synthesized spectra. And then, the estimated band magnitudes and the pitch are used for the V/UV decision of each harmonic band, and the decision is performed by comparing the spectral estimation error with a predetermined or adaptive threshold. From the inspection of the MBE model, we can see some mismatches of synthesized spectra between the analysis and synthesis procedures. In analysis procedure, the speech spectrum is synthesized by assuming all bands are voiced. However, this results in large estimation error of model parameters when lots of bands are unvoiced.

In this paper, we propose a new method for the parameter estimation in the MBE vocoder. This method estimates speech signal not only with voiced spectrum but also with unvoiced spectrum. For this, we first propose a new V/UV decision method. The two spectral estimation errors are computed with the assumption of the voiced or unvoiced spectra, and these errors are compared with each other. The V/UV decision is done to have smaller spectral estimation error. This V/UV decision is combined into the closed loop of AbS. In other words, all the MBE model parameters are jointly estimated. Using this joint estimation, we can obtain highly accurate estimation of model parameters and better quality of reproduced speech than that of the conventional MBE model. Additionally we can eliminate voicing thresholds in the V/UV decision because the decision is done by the voiced and unvoiced spectra.

2. SEQUENTIAL OPTIMIZED MBE MODEL

In MBE vocoder [1], speech signal is modeled by using pitch τ and harmonic band magnitudes $\{A_m, m=1,\cdots,M(\tau)\}$, where $M(\tau)$ is the total number of harmonics depending on τ, as

$$\left|\hat{S}_w(\omega)\right| = A_m \left|E_w(\omega)\right|. \qquad (1)$$

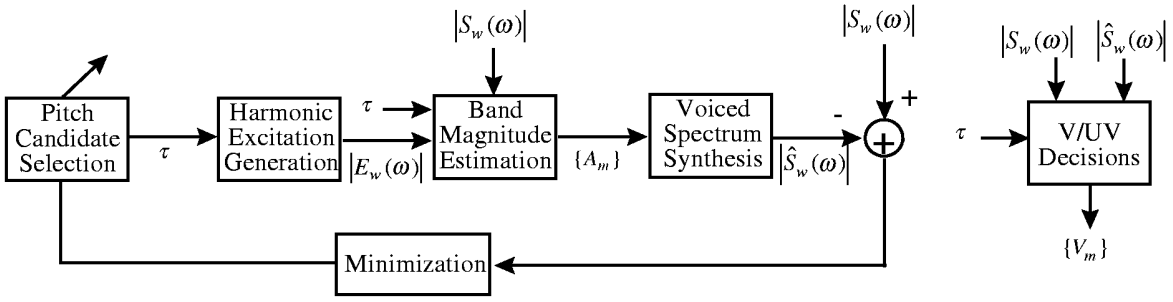

Figure 1. Block diagram of the sequential optimized MBE model.

The $|E_w(\omega)|$ of (1) is assumed as the spectrum of voiced excitation which is represented as $|E_w(\omega)|=|E(\omega)*W(\omega)|$, where $E(\omega)=\delta(\omega-mF_s/\tau)$ and $W(\omega)$ mean the spectra of the periodic excitation and window, respectively (F_s is sampling frequency). The band magnitudes and pitch are estimated by minimizing the following error measure $\xi(\tau)$ in AbS method defined by

$$\xi(\tau)=\frac{\sum_{m=1}^{M(\tau)}\int_{a_m}^{b_m}\left(|S_w(\omega)|-|\hat{S}_w(\omega)|\right)^2 d\omega}{(1-B\tau)\sum_{m=1}^{M(\tau)}\int_{a_m}^{b_m}|S_w(\omega)|^2 d\omega}, \qquad (2)$$

where a_m and b_m are lower and upper frequency bounds for the m-th harmonic band, respectively, $w(n)$ is a window, and $1\text{-}B\tau$ is the correction factor of the error measure due to pitch biasing.

After estimating pitch and band magnitudes, MBE vocoder calculates normalized spectral estimation error ξ_m for each band as

$$\xi_m(\tau)=\frac{\int_{a_m}^{b_m}\left(|S_w(\omega)|-|\hat{S}_w(\omega)|\right)^2 d\omega}{\int_{a_m}^{b_m}|S_w(\omega)|^2 d\omega}, m=1,\cdots,M(\tau). \quad (3)$$

Subsequently, the MBE vocoder determines the V/UV of m-th harmonic band using a threshold θ as

$$V_m=\begin{cases} Voice(V), & if \quad \xi_m \le \theta, \\ Unvoice(UV), & otherwise. \end{cases} \qquad (4)$$

In practice for the V/UV decisions, Improved MBE (IMBE) vocoder [5] uses several thresholds which are defined by using integer pitch search error, band energy, and (3). The analysis procedure of the sequential MBE model is shown in Figure 1.

From the above description, we can notice that the MBE model analyzes speech signal on the basis of voiced speech. That is, if the quantity of the error measure is large, the AbS procedure determines that the pitch candidate is incorrect, and then classifies a band into unvoiced. But this model has some problems. First, unvoiced bands which have large spectral estimation errors may cause incorrect pitch estimation. Second, the V/UV decision procedure requires a threshold which is hard to obtain because it may be varied according to the signal characteristics, the speaking environment, and so on. In the case of IMBE, to define several thresholds, pitch smoothing process by dynamic programming is used, but the process requires large codec delay and computational complexity. Lastly, the finally synthesized spectrum of the AbS is different from that of decoder. The AbS process in the encoder synthesizes voiced spectrum, but the decoder synthesizes voiced and unvoiced spectrum.

3. JOINT OPTIMIZED MBE MODEL

The proposed MBE model jointly estimates the pitch, band magnitudes, and V/UV decisions in AbS method. For the estimation of a spectrum in the speech analysis, the proposed MBE model uses voiced and unvoiced spectra while the sequential model does only voiced spectrum. The voiced spectrum $|\hat{S}_w^v(\omega)|$ is modeled with band magnitude $\{A_m^v\}$, periodic pulse train $\delta(\omega-mF_s/\tau)$ and window spectrum $W(\omega)$ as

$$\left|\hat{S}_w^v(\omega)\right| = A_m^v\left|E_w^v(\omega)\right| = A_m^v\left|\delta(\omega-mF_s/\tau)*W(\omega)\right|. \quad (5)$$

On the other hand, the unvoiced spectrum $|\hat{S}_w^{uv}(\omega)|$ is modeled as

$$\left|\hat{S}_w^{uv}(\omega)\right| = A_m^{uv}\left|E_w^{uv}(\omega)\right| = A_m^{uv}\left|R(\omega)*W(\omega)\right|, \quad (6)$$

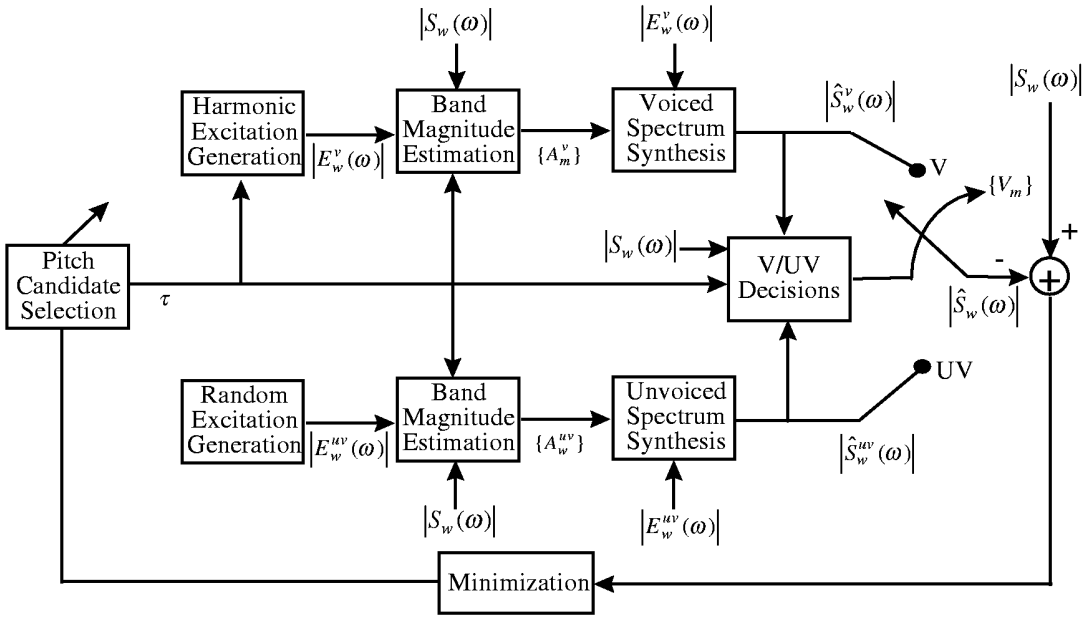

Figure 2. Block diagram of the joint optimized MBE model.

where $R(\omega)$ is random spectrum whose expected power is 1. For measuring the accuracy of the voiced and unvoiced models, the spectral estimation errors ε_m^v and ε_m^{uv} which correspond to voiced and unvoiced models, respectively, are calculated between the original and synthesized spectra for the m-th harmonic band as follows:

$$\varepsilon_m^k = \int_{a_m}^{b_m} \left(|S_m(\omega)| - |\hat{S}_m^k(\omega)| \right)^2 d\omega, \quad for \ \ k = V \ or \ UV. \quad (7)$$

If a spectral estimation error of voiced band is smaller than that of unvoiced one, the band is determined to be voiced. Otherwise, the band is classified into unvoiced. That is, the V/UV decision of the m-th spectral band is

$$V_m = \begin{cases} V, & if \ \varepsilon_m^v < \varepsilon_m^v, \\ UV, & otherwise. \end{cases} \quad (8)$$

This model does not require any thresholds for the V/UV decision of each harmonic band. The m-th band magnitude can be obtained by the minimization procedure of (7) and V_m of (8) such that

$$A_m^{V_m} = \frac{\int_{a_m}^{b_m} |S_w(\omega)| |E_w^{V_m}(\omega)| d\omega}{\int_{a_m}^{b_m} |E_w^{V_m}(\omega)|^2 d\omega}. \quad (9)$$

According to the V/UV decision of each band, synthetic spectrum $|\hat{S}_w(\omega)|$ of (1) becomes $|\hat{S}_w^v(\omega)|$ if the m-th harmonic band is voiced, otherwise $|\hat{S}_w^{uv}(\omega)|$. The pitch is determined by the AbS with (2) by using the synthetic spectrum. The joint estimation procedure of pitch, band magnitudes, and V/UV decisions are shown in Figure 2.

4. EXPERIMENTS AND DISCUSSIONS

For the above two models, the sequential optimized MBE and the joint optimized MBE, the performance of each model is evaluated with the power spectrum of synthetic speech. The speech synthesis at the decoder follows the IMBE model [5]. The phases of the spectrum are not coded but predicted using the frequencies at frame boundaries while preserving the smoothness at the boundaries. For the comparision of the quality of the reconstructed speeches, we depicted the power spectra as shown in Figure 3. We selected a transient region from a speech signal because it has highly complex mixture of voiced and unvoiced signals. The figure explains that the joint optimized MBE model produces better spectrum than the sequential MBE model.

Next, we also have compared spectrograms of long time speech as shown in Figure 4. From the spectrograms, we also can observe that the proposed MBE model reproduces better harmonic and noise structure in reconstructed speech. Informal listening test also confirms the superiority of the proposed model.

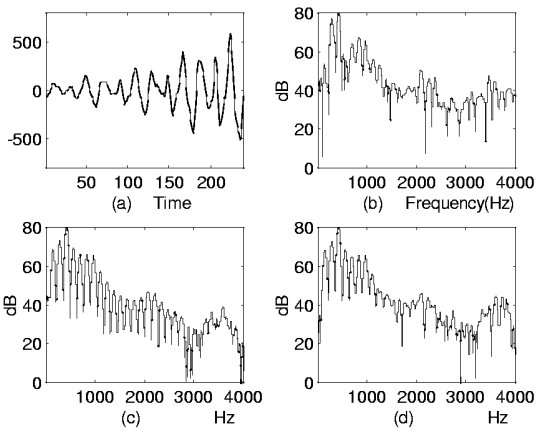

Figure 3. Given (a) an original waveform, the comparisions of (b) original power spectrum, (c) the power spectrum of the sequential optimized MBE model, and (d) the power spectrum of the joint optimized MBE model.

5. CONCLUSION

This paper proposes a joint optimized MBE model whose synthetic speech results in higher quality when compared with that of the conventional MBE model. In order to show the improved performance of the joint optimized MBE model, the short time spectrum, long time spectrogram and informal listening test result are used. It can be observed that the joint optimized MBE model produces more accurate model parameters rather than the conventional model does.

REFERENCES

[1] W. Griffin and J. S. Lim, "Multiband Excitation Vocoder," *IEEE Trans. ASSP*, Vol. ASSP-36, pp. 1223-1235, Aug. 1988.

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

[3] R. V. Cox, "Speech Coding Standards," *Speech Coding and Synthesis*, Ed. by W. B. Kleijn and K. K. Paliwal, Elservier, 1995.

[4] S. Dimolitsas, et. al. "Transmission Performance Evaluation of Voice Encoding Technology for the INMARSAT Mini-M System," *International Journal of Satellite Communications*, Vol. 14, pp. 381-387, 1996.

[5] DVSI, "APCO Project 25 Vocoder Description," Ver 1.3, July 1993.

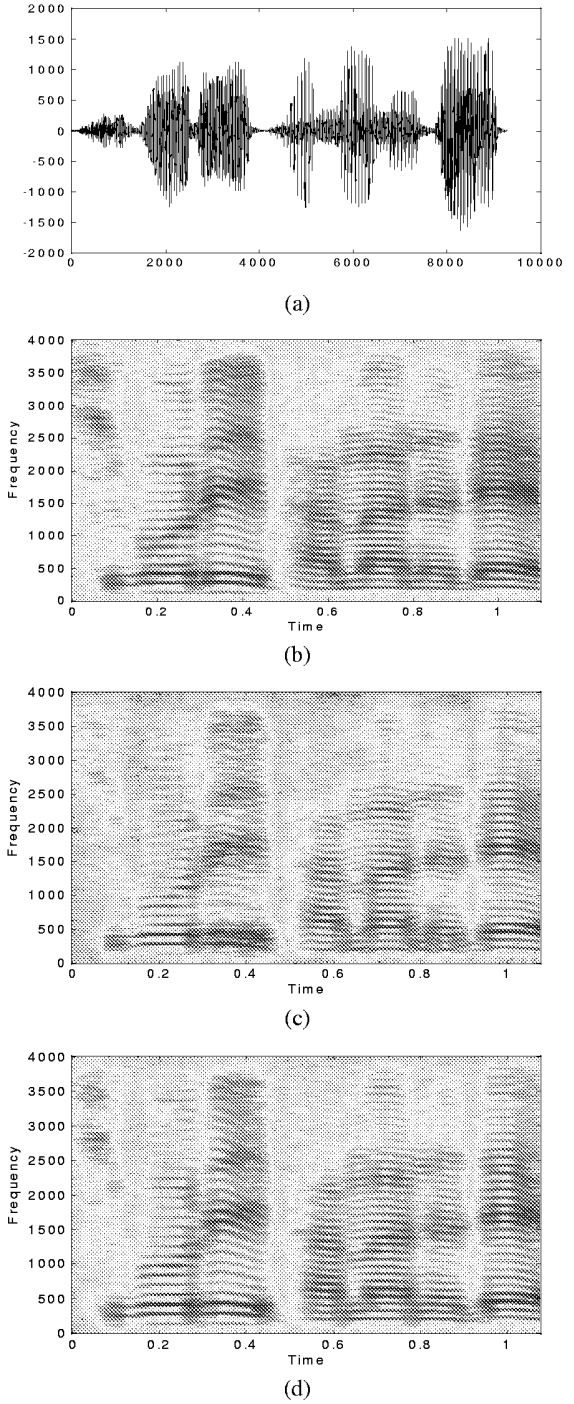

Figure 4. Given (a) an original waveform of a Korean male speaker, the comparisions of (b) original spectrogram, (c) the spectrogram of the sequential optimized MBE model, and (d) the spectrogram of the joint optimized MBE model.