MIXED MULTI-BAND EXCITATION CODER USING FREQUENCY DOMAIN MIXTURE FUNCTION (FDMF) FOR A LOW BIT-RATE SPEECH CODING

Woo-Jin Han, Sung-Joo Kim, Yung-Hwan Oh
Computer Science Dept.
Korea Advanced Institute of Science and Technology
Ku-song dong, Yu-song ku, Taejun, 305-701 Korea.
E-mail: hwjketel@bulsai.kaist.ac.kr

ABSTRACT

This paper describes the Mixed Multi-Band Excitation coder used for a low bit-rate speech coding. In MBE coders, there are significant differences of the fine structure between the original and the synthetic spectrum. They are mainly due to the exclusive partition of voiced and unvoiced regions in frequency domain and the decision procedure based on the experimental threshold. The MMBE uses frequency domain mixture function (FDMF) to overcome these drawbacks of the MBE coder. Also, two analysis methods, which do not need any decision procedure based on a threshold, are presented. The performance evaluation results show that the 2.6kbps MMBE coder reduces the average spectral distortion by a clear margin comparing to the 2.9kbps MBE coder. The computational load of the proposed coder is sufficiently small for a real-time implementation on the modern DSP chip.

1. INTRODUCTION

In a low bit-rate speech coding, speech is usually modelled by the time-varying filter output of the excitation signal considered as a speech source [1]. There have been considerable researches that represent the excitation signal accurately. Among them, Multi-Band Excitation (MBE) [2] coder can achieve high quality synthetic speech below 4.0kbps. The MBE model represents an input speech as the multiplication of a spectral envelope and an excitation spectrum. The excitation spectrum contains both V/UV components. The V/UV decisions are made over each harmonic of the fundamental frequency. As a result, the synthetic speech frame can have both V/UV components. But, there is notable distortion between the original and the synthetic spectrum because of its hard decision of the V/UV components.

HSX (harmonic stochastic excitation) [3] model was proposed to improve MBE model. Instead of having a binary V/UV decision by frequency subbands as in MBE, it uses a voicing level that is a continuous function of frequency. The voicing level is estimated by comparing the normalized autocorrelation function of each subband with the experimental threshold. Both the MBE coder and the HSX coder can be easily affected by the various environments because their speech models have the estimation procedure based on given thresholds.

This paper proposes MMBE speech model to overcome the drawbacks of both the MBE and the HSX coder. In MMBE speech model, excitation spectrum is represented by a frequency domain mixture function (FDMF), instead of the binary decision of the MBE coder. We also present a robust analysis method, which does not need any heuristic procedure based on thresholds. Section 2 of this paper describes the MMBE speech model. Section 3 presents two methods used to estimate the FDMF. A 2.6kbps MMBE coder is presented and evaluated in section 4.

2. MMBE SPEECH MODEL

MBE speech model represents the excitation spectrum as the sum of the periodic spectrum and the noise spectrum that do not exist together in the same harmonic band [2][4]. Let $\hat{E}(n)$ represent the excitation spectrum that is a discrete Fourier transform of the excitation signal. Then, $\hat{E}(n)$ can be written as:

$$\hat{E}(n) = V(n)(1-u(n)) + U(n)u(n) \qquad (1)$$

where $V(n)$ and $U(n)$ are the periodic spectrum and the noise spectrum, respectively. The V/UV decision function, $u(n)$, can be defined as

$$u(n) = \begin{cases} 0 & \text{if } \varepsilon_m \le \theta \\ 1 & \text{otherwise} \end{cases} \quad \text{for } a_m \le n \le b_m \qquad (2)$$

where $[a_m, b_m]$ is the interval around the m_{th} harmonic and θ is the experimental threshold. The prediction error of m_{th} harmonic, ε_m, is estimated assuming that the given harmonic band is declared voiced [2].

The major difference between the traditional MBE models and the proposed MMBE model is the way the V/UV information for each speech frame is represented. MMBE model uses a frequency domain mixture function (FDMF), $p(n)$, which is defined as the ratio of the V/UV components at the given frequency n. Unlike the V/UV decision function of MBE, FDMF has a real-valued function of frequency which range is the interval [0,1]. In

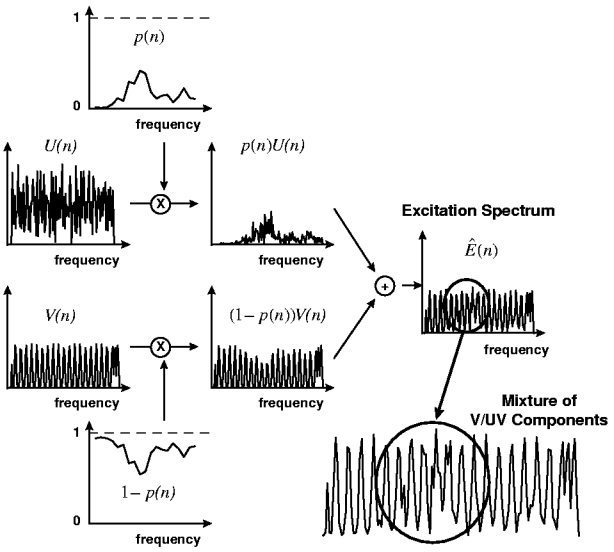


Figure 1: Synthesis procedure of the excitation spectrum in MMBE speech model

MMBE model, the excitation spectrum is represented by the FDMF as

$$\hat{E}(n) = V(n)(1-p(n)) + U(n)p(n) \quad \text{where } 0 \le p(n) \le 1 \tag{3}$$

Figure 1 shows the synthesis procedure of the excitation spectrum in MMBE model. MMBE model allows the V/UV components to be mixed together in the same frequency region. An example, which shows (a) original signal of the word "mu-kung-hwa", (b) V/UV decision function of MBE, and (c) FDMF of MMBE, is illustrated in Figure 2. In this Figure, Black regions in (b) mean the voiced component, whereas the degree of darkness in (c) means the ratio of voiced component.

3. DETERMINATION OF FDMF

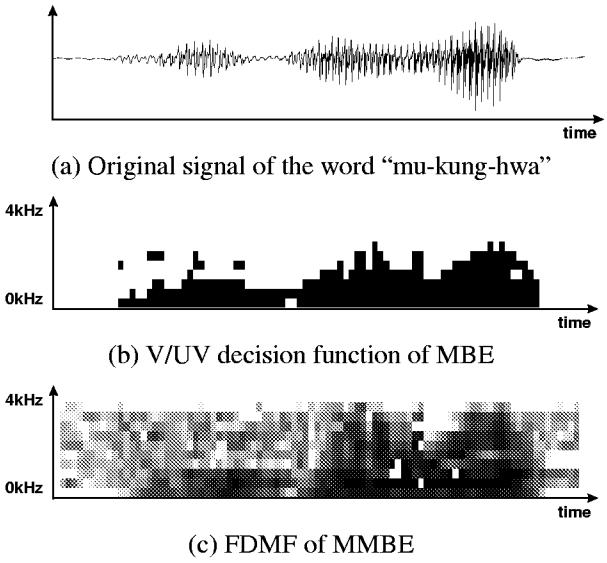
The MMBE model parameter set includes the spectral envelope, fundamental frequency, and the FDMF for each speech frame. The processing flow is similar to that of the original MBE coder [2]. Therefore, we only represent the methods for estimating the FDMF in this paper.

3.1. Codebook Search Method

At first, we simplify that the FDMF is constant in the certain interval, $[a_i, b_i]$, with a value of p_i. This enables us to write the mean square error between the original and the synthetic spectrum in $[a_i, b_i]$ as

$$E_i = \sum_{n=a_i}^{b_i} \left| S(n) - A(n)\{(1-p_i)V(n) + p_i U(n)\} \right|^2 \tag{4}$$

where $S(n)$ and $A(n)$ are the original speech spectrum and the spectral envelope function, respectively. For the



(a) Original signal of the word "mu-kung-hwa"

(b) V/UV decision function of MBE

(c) FDMF of MMBE

Figure 2: An example of the analysis

given i, the optimal value of p_i can be computed by setting the derivative of E_i with respect to a value of p_i to zero. As a result, the optimal value of p_i is calculated as

$$p_i = \frac{\sum_{n=a_i}^{b_i} A(n)(U(n)-V(n))(S(n)-A(n)V(n))}{\sum_{n=a_i}^{b_i} A(n)^2 (U(n)-V(n))^2} \tag{5}$$

As you see, the value of $A(n)$, $V(n)$, and $U(n)$ must be known for evaluating the equation (5). The spectral envelope function, $A(n)$, can be modelled effectively by spectral amplitude samples [2] or linear predictive analysis [4]. An efficient method for obtaining a good approximation for the periodic spectrum, $V(n)$, is to compute samples of the Fourier transform of the window function and centre it around the each multiple of fundamental frequency. But, the noise spectrum, $U(n)$, can not be modelled easily because it does not have a fixed shape.

In this paper, we use a noise spectrum codebook that has the overlapped structure for predicting the noise spectrum. The overlapped structure can reduce the storage requirement significantly. The noise spectrum codebook is the Fourier transform of white gaussian noise with a length L. If the length of the interval $[a_i, b_i]$ is M, the codebook is composed of L-M+1 codewords because of its overlapped structure. The structure of the codebook has shown in Figure 3.

The optimal noise spectrum is chosen from the noise spectrum codebook to minimize the error of equation (4). And then, we can calculate a value of p_i by equation (5). In this manner, FDMF can be estimated by applying the previous two equations to the entire frequency regions in an A-b-S way.

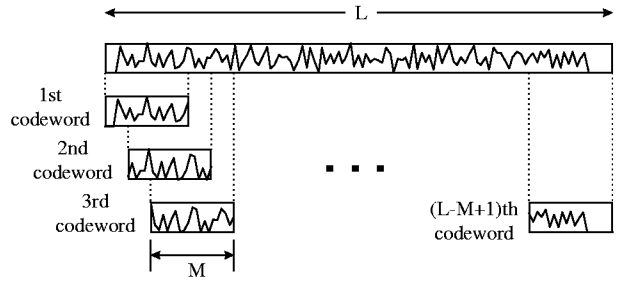


Figure 3: Overlapped structure of the noise spectrum codebook

By the way, equation (4) and (5) are defined over the interval [a_i, b_i]. This interval can be set by the interval around the harmonic as that of MBE coders, or the interval of fixed length as that of subband coders. We adopt the interval of fixed length for the simplicity of quantization. In this case, the analysis interval [a_i, b_i] is defined as

$$a_i = \left\lfloor \frac{N(i-1)}{2F} \right\rfloor, \; b_i = \left\lfloor \frac{Ni}{2F} \right\rfloor - 1 \quad \text{for } 1 \le i \le F \tag{6}$$

where N and F are the size of DFT and the number of intervals, respectively. In practice, we use N=256 and F=4.

3.2. Complexity Reduction Method

The codebook search method can estimate FDMF minimizing the mean square error between the original and the synthetic spectrum. It requires, however, a large computational complexity for searching the entire codebook. We propose another method to reduce the computational requirement of the codebook search method.

To reduce the computational complexity, we assume that the mean of the original and the synthetic spectrum is equal in the interval [a_i, b_i] if we use the optimal value of p_i and $A(n)$. From this assumption, equation (7) is derived.

$$\sum_{n=a_i}^{b_i} S(n) = \sum_{n=a_i}^{b_i} A(n)\{(1-p_i)V(n) + p_i U(n)\} \tag{7}$$

Then, a optimal value of p_i can be calculated from equation (7) as

$$p_i = \frac{\sum_{n=a_i}^{b_i} (S(n) - A(n)V(n))}{\sum_{n=a_i}^{b_i} A(n)U(n) - \sum_{n=a_i}^{b_i} A(n)V(n)} \tag{8}$$

In general, the noise spectrum, $U(n)$, can be modelled by the Fourier transform of the normalized white gaussian noise. Therefore, we assume that the mean of $U(n)$ in the interval [a_i, b_i] is always constant with a value of c. Equation (9) can be derived from this assumption.

$$\sum_{n=a_i}^{b_i} A(n)U(n) \approx c \sum_{n=a_i}^{b_i} A(n) \tag{9}$$

We can make a value of c one by normalizing the power of the noise spectrum. As a result, equation (8) can be simplified by applying equation (9) as

$$p_i = \frac{\sum_{n=a_i}^{b_i} (S(n) - A(n)V(n))}{\sum_{n=a_i}^{b_i} A(n)(1-V(n))} \tag{10}$$

The FDMF can be estimated by applying the equation (10) to all frequency intervals without any codebook search, so the computational complexity is greatly reduced comparing to the codebook search method. In practice, equation (10) can be implemented very efficiently with multiply-add operation in modern DSP chips, such as TMS320C30. It can be shown that the overall computational load of the MMBE coder is comparable to that of the MBE coder.

4. EXPERIMENTAL RESULTS

4.1 MMBE Coder Configuration

To illustrate the potential of the proposed speech model, we developed a MMBE speech coder. The bit allocation for an overall bit-rate of 2.6kbps (20ms frame length) is tabulated in Table 1. The fundamental frequency and excitation energy are quantized using logarithmic quantizers. The spectral envelope is represented by 10 LPC coefficients that are scalar quantized in the form of LSF parameters. If we apply the more efficient quantization method such as 2DdLSP [6], the overall bit-rate can be reduced significantly. We quantize the LSF coefficients by direct scalar quantization for simplicity's sake. The FDMF is vector quantized with 16 levels. We use LBG algorithm [6] for the vector quantization.

Table 1: Bit allocation for 2.6kbps MMBE coder.

Parameters	No of Bits/Frame	Bit-rate (kb/s)
Pitch	8	0.4
10 LSF Coef.	34	1.7
Energy	6	0.25
FDMF	4	0.2
Total	52	2.6

4.2 SD Objective Measurement

A spectral distortion (SD) measure is an attempt to measure the distortion between the original and the synthetic spectrum objectively. The SD is given by

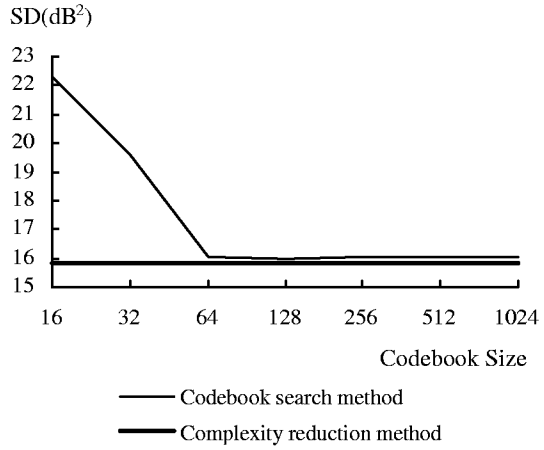


Figure 4: Evaluation of two proposed methods

$$SD = \frac{2}{N}\sum_{n=1}^{\frac{N}{2}}\left[10\log_{10}|S(n)| - 10\log_{10}|\hat{S}(n)|\right]^2 \ (\mathrm{dB}^2) \quad (11)$$

where N is the size of DFT, $S(n)$ is the original spectrum and $\hat{S}(n)$ is the synthetic spectrum [7].

4.3 Evaluation of Two Proposed Methods

Figure 3 shows the average SD of 10 Korean sentences for two proposed methods, that are the codebook search method and the complexity reduction method. The codebook size L=16, 32, 64, 128, 256, 512 and 1024 was used. In Figure 4, it is indicated that the performance of two proposed methods is very similar if L is larger than 64. Also, the complexity reduction method is slightly better than the codebook search method even if L is sufficiently large. From these results, it is easily concluded that the complexity reduction method works well in spite of its small computational complexity.

4.4 Performance Comparison with MBE Coder

We compared the 2.6kbps MMBE coder using the complexity reduction method with a 2.9kbps MBE coder on the aspect of the spectral distortion. The MBE coder used the same configuration with the MMBE coder except the V/UV information. In the MBE coder, the V/UV information is represented by the V/UV decision function instead of the FDMF. We assigned 10 bits to quantize it. SDs of 10 Korean sentences for two coders are illustrated in Figure 5. The test result shows that the proposed coder reduces average SD of all test sentences comparing to the MBE coder.

5. CONCLUSION

In this paper, Mixed Multi-Band Excitation (MMBE) speech model is proposed. Two analysis methods, which do not need any procedure based on thresholds, are also

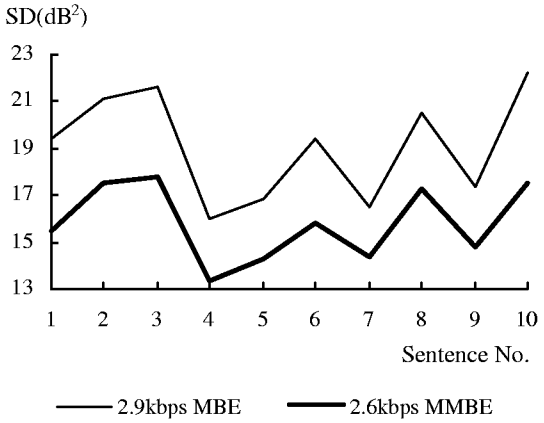


Figure 5: Performance comparison of the 2.9kbps MBE coder and the 2.6kbps MMBE coder

presented. In the proposed model, the excitation spectrum can be represented by the frequency domain mixture function (FDMF), which allows the V/UV components to be mixed over all frequency regions. Test results shows the proposed coder has smaller spectral distortion than that of the MBE coder even at a lower bit-rate. The computational load of the proposed coder is comparable to the MBE coder.

REFERENCES

[1] Sadaoki Furui, Digital Speech Processing, Synthesis, and Recognition, Marcel Dekker, Inc., 1992.

[2] Daniel W. Griffin and Jae S. Lim, "Multiband Excitation Vocoder," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-36, pp. 1223-1235, Aug. 1988.

[3] C. Laflamme, R. Salami, R. Matmti, J-P. Adoul, "Harmonic-Stochastic Excitation (HSX) Speech Coding Below 4Kbit/s," *proc. of ICASSP*, pp. 204-207, 1996.

[4] D. Rowe, W. Cowley, A. Perkis, "A Multiband Excitation Linear Predictive Hybrid Speech Coder," *proc. of Eurospeech*, pp. 239-242, Genova, Italy, Sept. 1991.

[5] Chih-Chung Kuo, Fu-Rong Jean, and Hsiao-Chuan Wang, "Low Bit-Rate Quantization of LSP Parameters Using Two-Dimensional Differential Coding," *proc. of ICASSP*, pp. 97-100, 1992.

[6] Gray, R. M., "Vector Quantization," *IEEE ASSP Magazine*, 1, 2, pp. 4-29, 1984.

[7] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*, John wiley and sons, 1994.