ROBUST MULTIBAND EXCITATION CODING OF SPEECH BASED ON VARIABLE ANALYSIS FRAME SIZES

Eric W.M. Yu and Cheung-Fat Chan

Department of Electronic Engineering
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong.
Email: eewmeyu@cityu.edu.hk eecfchan@cityu.edu.hk

ABSTRACT

A robust technique for the coding of multiband excitation (MBE) model parameters from a non-stationary speech segment is proposed in this paper. The non-stationary speech segment which has an abrupt increase in its signal energy with respect to the time is divided into 2 quasi-stationary speech segments. A variable analysis frame size technique is proposed to analyze the lower energy portion and the higher energy portion separately. A high quality fixed 1.6 kbps variable frame size MBE linear predictive (MBELP) speech coder was developed.

1. INTRODUCTION

In MBE speech coding, the MBE model parameters are conventionally obtained by computation of the errors between the original short-term speech spectrum and a pitch dependent periodic spectrum [1]. The extracted model parameters are accurate provided that the speech segment contains a stationary signal. When a fixed size analysis window is sliding through the time, some of the captured speech segments will contain an abrupt change of signal energy with respect to the time due to the non-stationary nature of speech. An example of abrupt increase in speech signal energy is shown in Figure 1. As shown in Figure 2, explicit harmonic structure cannot be observed through the short-term magnitude spectrum of a speech segment which is captured by the window $w_k(m)$ over the abrupt transition. By means of the conventional MBE analysis procedure which relies on the pitch dependent periodic spectrum for error computation, the accuracy of the MBE model will be decreased and eventually affecting the speech quality. The effect will be more obvious in fixed analysis frame size low bit rate speech coding since the frame shift is longer and the temporal resolution was reduced consequently. In this paper, our attention is focused on improving the speech coder robustness to the type of abrupt transition in Figure 1. A variable analysis frame size technique is proposed to obtained reliable and accurate MBE model parameters when the windowed speech signal transits abruptly from low energy to high energy with respect to the time. Robust speech analysis and synthesis procedure are developed to improve the speech quality of a 1.6 kbps MBELP speech coder.

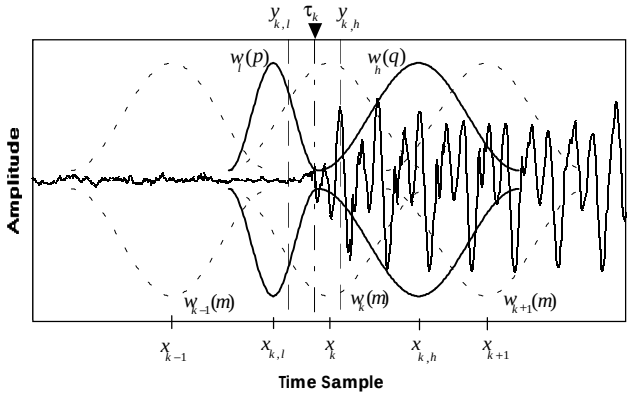

Figure 1. *Speech waveform (95 ms) and the windowing schemes*

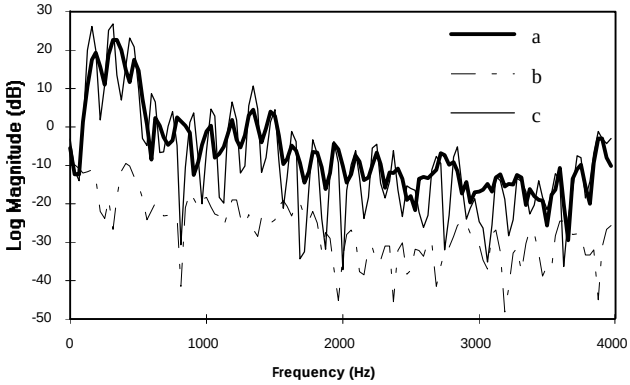

Figure 2. *Spectra of* (a) *the kth speech segment in Figure 1* (b) *the signal* $s_l(p)$ *and* (c) *the signal* $s_h(q)$

2. A REVIEW OF FIXED FRAME SIZE MBELP ANALYSIS AND SYNTHESIS

In fixed rate block coding of speech signal, a short speech segment of about 20 to 40 ms is conventionally windowed for analysis and speech model parameters are extracted to represent the speech segment. The fixed size analysis window is sliding through the time in a fixed interval, say

N samples, and proceeding with the order as shown in Figure 1 by the windows $w_{k-1}(m)$, $w_k(m)$, $w_{k+1}(m)$ and so on, where k is the frame index and $-\infty < k < \infty$. The windows are overlapped by (L-N) samples, where L is the window length. For a discrete-time speech signal $s(n)$, where $-\infty < n < \infty$, a segment $\{s(Nk+m)\}_{m=-L/2}^{L/2-1}$ will be captured by the window $w_k(m)$. The spectrum of windowed speech signal $\{s(Nk+m)w_k(m)\}_{m=-L/2}^{L/2-1}$ will then be obtained by a L-point discrete Fourier transform. MBE analysis procedure is then applied to extract the MBE model parameter set Γ_k. The MBE model parameters are pitch, V/UV information and band magnitudes. The band magnitudes are quantized through the corresponding 10th order LPC spectrum. A fast embedded 2-dimensional differential line spectrum pair (2DDLSP) quantization scheme is applied to convert the PARCOR coefficients to LSP parameters and quantize the prediction errors simultaneously [2].

The MBE speech synthesis procedure can be divided into the voiced speech synthesis and the unvoiced speech synthesis. For the voiced speech synthesis, the information associated with the parameter sets Γ_k and Γ_{k+1} is used for voiced band magnitude interpolation between the time samples $x_k = (Nk + n_o)$ and $x_{k+1} = (Nk + N + n_o)$ so that a bank of harmonic oscillators can be applied to synthesize voiced speech signal between x_k and x_{k+1}. The value of n_o indicates the delay in samples. Note that n_o is set to zero in Figure 1 only for the sake of illustration. For the unvoiced speech synthesis, a windowed white noise of length equals to the analysis window size L is filtered according to the unvoiced band magnitudes obtained from the corresponding set of model parameters. For the model parameter set Γ_k, the filtered noise segment is centred on time sample x_k. Continue sequence of unvoiced speech is reconstructed by an overlap-add procedure between adjacent segments of filtered noise. With proper alignment of the time samples, the reproduced speech signal is obtained by the sum between the reconstructed voiced speech and unvoiced speech.

Due to the non-stationary characteristics of speech signal, it is possible for a fixed size analysis window to capture a segment of speech which is non-stationary. Eventually, the accuracy of extracted model parameters will be decreased.

3. VARIABLE FRAME SIZE MBE ANALYSIS AND SYNTHESIS

The proposed variable analysis frame size technique includes method for the detection of abrupt increase in signal energy and the corresponding procedure in speech analysis and synthesis.

3.1 Detection of Abrupt Transition

For each speech frame which follows a low energy speech frame, we detect the occurrence of the abrupt increase in signal energy and determine the time instant of the occurrence. The abrupt transition of speech segment $\{s(Nk+m)\}_{m=-L/2}^{L/2-1}$ can be detected by tracking of the energy ratio $\alpha(k, j)$ with respect to the time sample j, where $j = 0,1,\ldots,L-1$. The energy ratio is defined as

$$\alpha(k,j) = \frac{\frac{1}{L-j}\sum_{i=j}^{L-1} s^2(Nk - \frac{L}{2} + i)}{\frac{1}{j}\sum_{i=0}^{j-1} s^2(Nk - \frac{L}{2} + i)} \quad (1)$$

Transition occurs at the time index

$$\tau_k = \arg\max_{\forall j}\{\alpha(k,j)\} \quad (2)$$

provided that $\alpha(k,\tau_k)$ is greater than a pre-defined threshold.

3.2 Analysis

By considering the speech segment $\{s(Nk+m)\}_{m=-L/2}^{L/2-1}$ which has been captured by the window $w_k(m)$, $-L/2 \le m < L/2$, if no abrupt increase of speech energy was detected, the parameter set Γ_k will be extracted from the windowed speech segment $\{s(Nk+m)w_k(m)\}_{m=-L/2}^{L/2-1}$ as described in Section 2. Supposing an abrupt transition was detected, with the aid of transition time index τ_k, we split the original fixed size analysis window $w_k(m)$ into a window $w_l(p)$, $-L/2 \le p < -L/2+\tau_k$, for the lower energy portion of the speech segment and a window $w_h(q)$, $-L/2+\tau_k \le q < L/2+\tau_k$, for the higher energy portion of the speech segment as shown in Figure 1. The window $w_l(p)$ is applied to the low energy portion to form the speech segment

$$s_l(p) = s(Nk+p)w_l(p) \quad (3)$$
$$-L/2 \le p < -L/2+\tau_k$$

while the window $w_h(q)$ is applied to the high energy portion to form the speech segment

$$s_h(q) = s(Nk+q)w_h(q) \quad (4)$$
$$-L/2+\tau_k \le q < L/2+\tau_k$$

As shown in Figure 2, the short-term magnitude spectra obtained individually from the windowed speech segments $s_l(p)$ and $s_h(q)$ provide explicit information of the harmonic structure. Besides performing the discrete Fourier transforms and the subsequent MBE analysis individually on $s_l(p)$ and $s_h(q)$, we propose a more efficient approach for coding of the parameters of these 2

speech segments through some reasonable assumptions. For the speech signal $s_l(p)$, the excitation is assumed to be the same as the previous speech segment $\{s(Nk-N+m)w_{k-1}(m)\}_{m=-L/2}^{L/2-1}$. Therefore, the V/UV information and pitch of the parameter set Γ_{k-1} can be used to represent the excitation of the speech signal $s_l(p)$. The band magnitudes with respect to $s_l(p)$ are quantized using 10th order LPC spectrum. We utilize the autocorrelation sequence of $s_l(p)$ to compute the corresponding 10th order PARCOR coefficients through the LeRoux-Gueguen method [3]. The PARCOR coefficients are quantized by the embedded 2DDLSP scheme as applied in [2]. For the speech signal $s_h(q)$, the whole set of model parameters is assumed to be the same as Γ_{k+1} which is the parameter set of the following speech segment $\{s(Nk+N+m)w_{k+1}(m)\}_{m=-L/2}^{L/2-1}$. If we denote the parameter sets of speech segments $s_l(p)$ and $s_h(q)$ by $\Gamma_{k,l}$ and $\Gamma_{k,h}$, respectively, the pitch and V/UV information of the parameter set Γ_{k-1} together with the 10th order LPC spectrum of $s_l(p)$ constitute the parameter set $\Gamma_{k,l}$ while the parameter set $\Gamma_{k,h}$ is equivalent to Γ_{k+1}. Consequently, we only have to encode the time index of the abrupt transition and the LPC spectrum of $s_l(p)$ from the parameter set $\Gamma_{k,l}$. There is no need to encode $\Gamma_{k,h}$ since $\Gamma_{k,h}=\Gamma_{k+1}$. The time index of the abrupt transition and the LPC spectrum of $s_l(p)$ are grouped into a parameter set $\overline{\Gamma}_k$. The parameter set $\overline{\Gamma}_k$ occupies the time slot which would be used by Γ_k if there was no abrupt transition detected. For fixed rate speech coding, the number of bits required by $\overline{\Gamma}_k$ is adjusted to be the same as Γ_k. Details of the bit allocation are provided in Section 4.

3.3 Synthesis

The synthesis procedure is the same as that described in Section 2 if no abrupt transition was detected. Otherwise, we have to cater for the varied window sizes and positions in speech synthesis. Prior to the synthesis of voiced and unvoiced speech, we have to recover the parameter sets $\Gamma_{k,l}$ and $\Gamma_{k,h}$. The parameter set $\Gamma_{k,l}$ is recovered by the combination of the pitch and V/UV information obtained from the parameter set Γ_{k-1} together with the band magnitudes and abrupt transition time index τ_k obtained from the parameter set $\overline{\Gamma}_k$. The parameter set $\Gamma_{k,h}$ is recovered by using the relationship $\Gamma_{k,h}=\Gamma_{k+1}$. For the voiced speech synthesis, the parameter sets Γ_{k-1} and $\Gamma_{k,l}$ are used for voiced band magnitude interpolation between the time samples $x_{k-1}=(Nk-N+n_o)$ and

$$y_{k,l}=(Nk-L(1+\gamma)/2+\tau_k+n_o) \tag{5}$$

as shown in Figure 1. The information associated with $\Gamma_{k,l}$ and $\Gamma_{k,h}$ is used for voiced band magnitude interpolation between the time samples $y_{k,l}$ and

$$y_{k,h}=(Nk-L(1-\gamma)/2+\tau_k+n_o) \tag{6}$$

Subsequently, the information associated with $\Gamma_{k,h}$ and Γ_{k+1} is used for voiced band magnitude interpolation between the time samples $y_{k,h}$ and x_{k+1}. The parameter γ in (5) and (6) controls the gradient of the voiced band magnitude linear interpolation from the low energy portion to the high energy portion. In fixed window size analysis, we have $\gamma=N/L$. As shown in (5) and (6), a lower value $\gamma=0.25$ is adopted in order to reproduce the abrupt increase in signal energy closely. For the unvoiced speech synthesis, a windowed white noise of length equals to the length of the window $w_l(p)$ is filtered according to the unvoiced band magnitudes obtained from the parameter set $\Gamma_{k,l}$. The filtered noise segment is centred on the time sample

$$x_{k,l}=(Nk-L/2+\tau_k/2+n_o) \tag{7}$$

as shown in Figure 1 and is joined with the noise segment associated with Γ_{k-1} by the overlap-add method. Afterwards, a windowed white noise of length equals to the length of the window $w_h(p)$ is filtered according to the information obtained from the parameter set $\Gamma_{k,h}$. The filtered noise segment is centred on the time sample

$$x_{k,h}=(Nk+\tau_k+n_o) \tag{8}$$

and is joined with the noise segment associated with $\Gamma_{k,l}$. Similarly, a windowed white noise of length equals to the length of the window $w_{k+1}(m)$ is filtered according to the information obtained from the parameter set Γ_{k+1}. The filtered noise segment is centred on the time sample x_{k+1} and is joined with the noise segment associated with $\Gamma_{k,h}$.

4. IMPLEMENTATION OF A FIXED 1.6 KBPS VARIABLE FRAME SIZE MBELP SPEECH CODER

The proposed variable analysis frame size technique was applied to develop a 1.6 kbps MBELP speech coder which produces intelligible speech with good quality. The proposed speech coder operates in either the steady mode or the transient mode according to the result of abrupt transition detection as described in Section 3.1. The continuous-time speech signal is sampled in 8 kHz. In the steady mode, the proposed speech coder has a fixed analysis window length of 32 ms and a fixed analysis window shift of 25 ms. For a fixed 1.6 kbps speech coder, we can use 40 bits per frame to quantize the model parameters in both steady and transient modes. The bit allocation per frame of the 1.6 kbps variable analysis

frame size MBELP speech coder is shown in Table 1. We allocated 8 bits for both pitch quantization and operation mode indication. The first $(2^8 - 1)$ quantization levels are used for pitch quantization while the last quantization level is used to indicate transient mode. In the steady mode, pitch is quantized where the pitch search range is from 76.4 Hz to 400 Hz. The band magnitudes are quantized by 30 bits. The corresponding 10th order spectral envelope will be quantized by 24 bits while the gain will be quantized by 6 bits. In order to quantize the V/UV information with only 2 bits, the conventional binary sequence for V/UV information quantization was replaced by a 2-bit V/UV transition frequency index [2]. The V/UV transition frequency is obtained by an analysis-by-synthesis procedure where the error between the original spectrum and the synthetic spectrum associated with the simplified V/UV mixture function are closed loop minimized. If the segment of speech is found having an abrupt increase in signal energy, the speech analysis procedure will follow Section 3.2 where pitch is not quantized. As an indication of transient mode, all the 8 bits which are originally used for pitch quantization will be filled with binary 1. In the transient mode, we only have to quantize the parameter set $\overline{\Gamma}_k$ as described in Section 3.2. The 2 bits originally used for quantization of V/UV transition frequency index will be used for quantization of the abrupt transition time index τ_k. The 10th order spectral envelope and gain of the windowed signal $s_l(p)$ will be quantized by 24 bits and 6 bits, respectively.

Table 1. Bit Allocation

	NO. OF BITS	
PARAMETERS	Steady Mode	Transient Mode
Pitch	8	8 (11111111_b)
V/UV or Abrupt Transition Index	2 (V/UV)	2 (Abrupt Transition Index)
Gain	6	6
Spectral Envelope	24	24
TOTAL	40	40

5. RESULTS

Spectra and waveforms of the original speech signal windowed by $w_k(m)$ and the reproduced speech signal are shown in Figure 3 and 4, respectively. The spectrum and waveform of the speech signal reproduced by a same bit rate fixed analysis frame size MBELP speech coder are also shown for comparison. It is obvious that the synthetic spectrum and the event localization of the speech segment with abrupt energy increase is improved by the proposed technique. Informal listening shown that the quality of the proposed 1.6 kbps MBELP speech coder is higher than its fixed analysis frame size counterpart and is comparable with the 2.4 kbps MBELP speech coder.

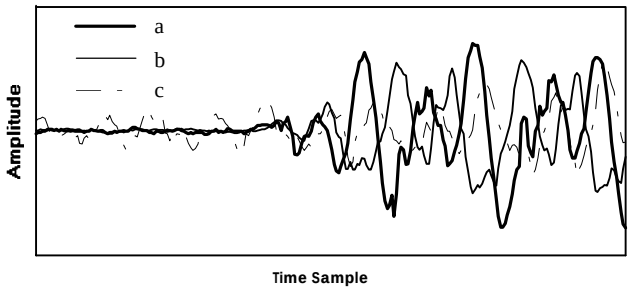

Figure 3. *Waveforms of* (a) *the kth speech segment in Figure 1* (b) *the reproduced speech through proposed 1.6 kbps MBELP coder and* (c) *the reproduced speech through a fixed frame size 1.6 kbps MBELP coder.*

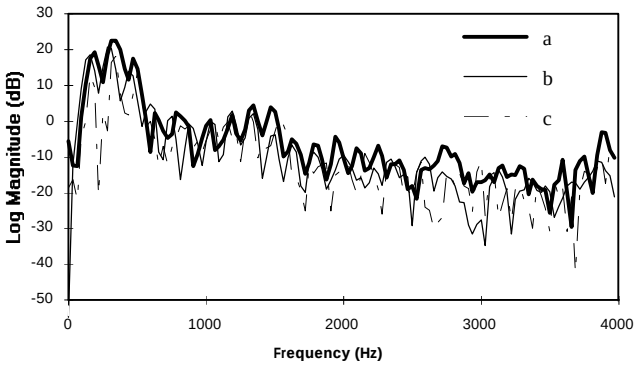

Figure 4. *Spectra of* (a) *the kth speech segment in Figure 1* (b) *the reproduced speech through proposed 1.6 kbps MBELP coder and* (c) *the reproduced speech through a fixed frame size 1.6 kbps MBELP coder.*

6. CONCLUSION

A robust variable analysis frame size technique was developed for the MBE analysis and synthesis of non-stationary speech segment which has an abrupt increase in energy with respect to the time. The proposed technique was applied successfully on the development of a high quality 1.6 kbps MBELP speech coder.

REFERENCES

[1] D.W. Griffin and J.S. Lim, "Multi-band excitation vocoder," *IEEE Trans. on ASSP*, vol.36, pp. 1223-1235, 1988.

[2] W.M.E. Yu and C.F. Chan, "Efficient multiband excitation linear predictive coding of speech at 1.6 kbps," *Proc. Eurospeech*, pp. 685-688, Sept. 1995.

[3] J. LeRoux and C. Gueguen, "A fixed point computation of partial correlation coefficients," *IEEE Trans. on ASSP*, vol. ASSP-25, pp. 257-259, 1977.