RELP Vocoder 의 음질향상에 관한 연구

On Improving the Quality of RELP Vocoder

* 오 성 근 (Oh, S.K.)
** 은 종 관 (Un, C.K.)

요약

지금까지 알려진 여러가지 음성부호화 방식들 중 4.8에서 9.6 kbits/s 사이의 전송속도에서 제일 좋은 성능을 갖는 것은 Residual - Excited Linear Prediction (RELP) 방식이다. RELP 부호화 방식은 전송속도가 낮을때, 합성음이 거칠거나 금속성의 잡음을 갖는 단첨이 있다. 본 논문에서는 이러한 단점을 보완하여 음질을 개선하는 세가지의 방법들을 제안하며, 그들은 다음과 같다. 첫째는 여러개의 baseband를 이용한 spectral folding 방법이고, 둘째는 spectral folding 방법과 pulsed excitation 방법을 조합한 방법이며, 마지막 방법은 여러개의 baseband를 사용한 spectral folding 방법과 pulsed excitation 방법을 조합한 방법이다.

이 방법들을 사용하여 RELP vocoder의 음질을 많이 개선할 수 있으며, 9.6 kbits/s 근처의 전송속도에서 사용하기 위한 첫번째 방법과 세번째 방법은 spectral folding 이나 nonlinear distortion 방법에서 문제가 되는 roughness 나 tonal noise 를 거의 인지 할 수 없으며, 세번째 방법이 첫번째 방법보다 우수하다. 두 번째 방법은 4.8 kbits/s근처의 전송속도에 적합하며, 기존의 'RELP 방식들에 비해 많은 음질향상을 가져왔다. 제안한 세가지 방법들을 같은 조건에서 비교할 때 세번째 방법이 가장 우수하며, 이 경우 합성음은 원음과 거의 흡사하다.

Abstract

Residual-excited linear prediction (RELP) vocoding is known to be one of the best approaches to speech coding in the range of 4.8 to 9.6 kbits/s. One problem associated with the RELP vocoder is that it often produces some roughness and tonal noise as the

* 한국과학기술원 전기 및 전자공학과 박사과정

** 한국과학기술원 전기 및 전자공학과 교수

transmission rate becomes lower. In this paper, we investigate three methods to improve its quality、These include the multiband spectral folding method, the method of using both the spectrally folded signal and the pulsed excitation signal, and the method of using both the multiband spectrally folded signal and the pulsed excitation signal.

Using these methods, we can improve the speech quality in the RELP vocoder. The first and third methods which are used at bit rate of about 9.6 kbits/s, can perceive little the roughness and tonal noise that are produced by spectral folding and non-linear distortion methods. Then the third method is superior to first one. The second method which is used at 4.8 kbits/s about, produces better performance than methods studied previously. Under identical conditions, the last one yields the best performance. It produces nearly identical synthetic speech as the original.

I. INTRODUCTION

Linear predictive coding (LPC) technique is being widely used for narrow-band speech communications. It is well known that, although the speech quality of this LPC vocoder is relatively good at the bit rate of 2.4 kbits/s, it has some perceptible buzziness that makes the synthetic speech somewhat unpleasant, and it is very susceptible to acoustical distortion and background noise. The cause of the buzziness is known to be due to the monotonous excitation signal and the binary decision of the excitation source. This problem may be avoided by using an improved excitation signal.

There exist several techniques of overcoming the shortcomings of the pitch-excited LPC vocoder [1] – [3]. One method is the residual-excited linear prediction (RELP) technique. This has been known to be one of the most promising candidates for medium-low rate (4.8 to 9.6 kbits/s) speech communication because of its relatively good quality and robustness to noise and distortion. However, the synthetic speech of the RELP vocoder often has some roughness and tonal noise. This problem can be alleviated by properly designing the full-band reconstruction system and the baseband residual coder.

Since the RELP technique was first proposed [2], various methods for reconstructing the full-band excitation signal have been proposed [4] - [10]. These may be classified into three categories; nonlinear distortion, spectral duplication and hybrid excitation methods. First, non linear distortion methods include rectification [2] and waveform clipping [4], [5] methods which have been most widely used because they are simple and easy to implement. In general, the use of a nonlinear distortion method results in roughness of the synthetic speech. Recently, Un and Lee proposed an improved RELP system with split-band coding. It has been reported that the synthetic speech of this system has little roughness [6].

Second, spectral duplication method includes spectral folding and spectral translation. It has been known that the synthetic speech by the spectral duplication method has some tonal and metallic noise because of spectral regularity [7]. Viswanathan et al. proposed prewhitening and random perturbation of the spectrally folded excitation signal. It has been reported that the synthetic speech with the first method exhibits slight pinging sound and with the second method slight roughness [8].

Lastly, the hybrid excitation methods include the pitch implantation method [9], the hybrid synthesizer method [10] and so forth. It has been reported that these methods produce some static-like noise but no audible roughness.

In this paper, we are concerned with the methods for reconstructing the full-band excitation signal for improving the synthesized speech quality in the RELP vocoder. Three improved methods for reconstructing the full-band excitation signal are proposed and evaluated by computer simulation. Details of the proposed methods follow.

II. RELP WITH IMPROVED EXCITATION

We now describe three methods for reconstructing the full-band spectrum of the excitation signal, which can improve the speech quality in the RELP vocoder. These methods, which may be used in RELP vocoding in the range of 4.8 to 9.6 kbits/s, take mainly the form of hybrid excitation, and do not require the process of nonlinear distortion, spectral flattening or gain adjustment. They include the multiband spectral folding method, the method of using both the spectrally folded signal and pulsed excitation signal, and the method of using both the multiband spectrally folded signal and pulsed excitation signal.

In the multiband spectral folding method (Hereafter, this method will be called the method I), we transmit two or three bands of the baseband signal, and then use the received bands to reconstruct the full-band excitation signal. In this method, multiple bands of the residual signal are obtained from the full-band residual signal. To implement the integar-band spectral folding, the following constraints are required:

$$f_{Li} = n\omega_i, \text{ (n: integer)} \qquad (1\text{-}1)$$

$$f_{Hi} = (n+1)\,\omega_i, \qquad (1\text{-b})$$

$$f_{s/2} = L\,\omega_i\,, \quad (L:\ \text{integer}) \qquad (1\text{-c})$$

$$f_{L_i} = M\,\omega_{(i-1)} + f_{L_{(i-1)}}, \qquad (2)$$
$$(f_{L_1} = 0 \text{ and } M \geq 2),$$

where f_{L_i} f_{H_i} and ω_i are the low and high edge frequencies and the bandwidth of the i-th baseband, respectively, and f_s is the sampling frequency of the input signal.

The baseband signals obtained are then down-sampled by the ratio of $L_i = f_{s/2\omega_i}$, and encoded.

At the receiver, each baseband is decoded and translated into bands corresponding to those at the transmitter. Then, the full-band reconstruction is done as follows:

(i) Each baseband is down-sampled and up-sampled by the equal ratio $L_i = f_s/2\omega_i$.

(ii) The band-passed signal is obtained by passing each spectrally folded signal through a band-pass filter with cutoff frequencies as follows:

$$f_{RL_i} = f_{L_i}, \qquad (3\text{-a})$$

$$f_{RH_i} = f_{L_i} + 2\omega_i\,. \qquad (3\text{-b})$$

where f_{RL_i} and f_{RH_i} are the low and high edge frequencies of the reconstruction filter for the i-th baseband, respectively. In this process, if the high edge frequency of the reconstruction filter for the highest baseband, f_{RH_N}, is not equal to $f_s/2$, i.e., $f_{RH_N} < f_s/2$, f_{RH_N} must be set to $f_s/2$.

(iii) Finally, the individually reconstructed bands are added to construct the full-band excitation signal.

The reconstructed full-band spectrum using this multiband spectral folding is shown in Fig. 1

The second method is a hybrid method that uses both the spectrally folded signal and pulsed excitation signal (This method will be called the method II). In this hybrid method, we adjust the bandwidth occupied by the pulsed excitation signal in the frequency domain. The spectral gap between the baseband and the upperband of pulsed excitation is filled with the folded spectrum of the baseband. The full-band reconstruction is done as follows:

(i) The pulsed excitation signal is highpass-filtered with cutoff frequencies f_c and $f_s/2$.

(ii) The baseband is spectrally folded by downsampling and upsampling the baseband residual.

(iii) The spectrally folded baseband residual (full-band residual) is lowpass-filtered with cutoff frequency f_c.

(iv) The signals obtained from the steps (i) and (iii) are added for reconstruction of the full-band excitation signal.

The resulting spectrum is shown in Fig. 2.

In the last method, we propose a method in which we combine the multiband spectrally-folded signal for the lowerband with the pulsed excitation signal for the upperband to

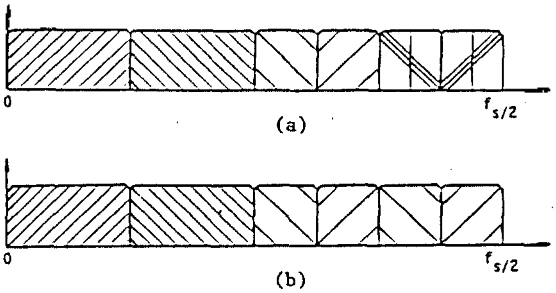

Fig. 1. Reconstructed full-band spectra by the method I.
(a) The case of 3 base bands
(b) The case of 2 base bands

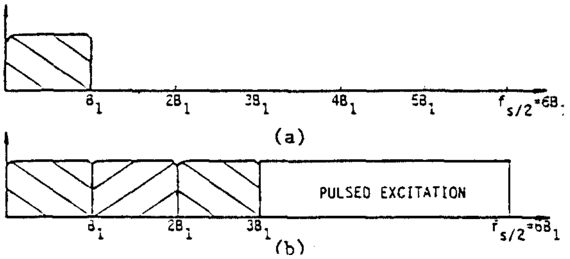

Fig. 2. The hybrid excitation method of combining spectrally folded signal and pulsed excitation signal.
(a) Baseband spectrum
(b) Reconstructed full-band spectrum

obtain a full-band excitation spectrum (This method will be called the method III). The procedure of reconstructing the full-band excitation signal is the same as for the method I except that the upperband is filled with the high-passed pulsed excitation signal in accordance with the step (i) of the method II. The resulting full-band spectrum is shown in Fig. 3. A block diagram of the RELP algorithm using the method III is shown in Fig. 4.

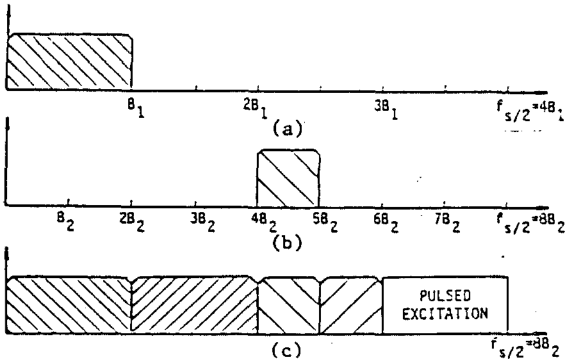

Fig. 3. The method of combining multiband spectral folding and pulsed excitation.
(a) Spectrum of the first baseband
(b) Spectrum of the second baseband
(c) Reconstructed full-band spectrum

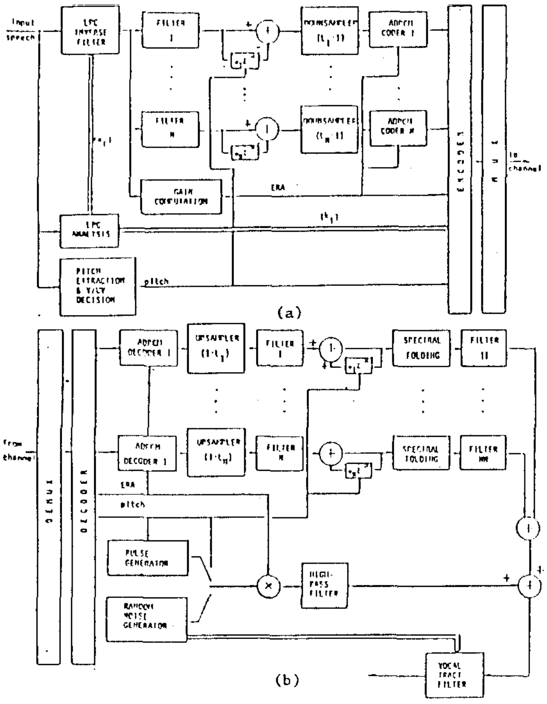

Fig 4. Block diagram of RELP vocoder with combined excitation of spectrally folded multiband signal and high-passed pulsed excitation signal.
(a) Transmitter
(b) Receiver

III. SIMULATION RESULTS AND DISCUSSION

We now present the results of computer simulation. In our simulation, we adopted the pitch-predictive ADPCM coding method for residual coding because it produced subjectively the best performance. We also used finite impulse response (FIR) digital filters for various filtering processes, which have passband ripple less than 0.5 dB, stopband attenuation greater than 55 dB and the filter length of 128 coefficients.

We first studied the multiband spectral folding method (i.e., the method I) and the method of combining both the multiband spectrally folded signal and pulsed excitation signal (i.e., the method III) for 9.6 kbits/s RELP vocoding. In the simulation, we studied two different systems based on the methods I and III. One takes the lower baseband of 0 to 900 Hz and the upper baseband of 1800 to 2100 Hz, and another takes the lwoer baseband of 0 to 600 Hz and the upper baseband of 1200 to 1800 Hz. Also, a RELP system with a single baseband of 0 to 1200 Hz that utilized the spectral folding method was simulated for comparison. In this simulation each baseband residual signal was encoded by a 3-bit pitch-predictive ADPCM coder.

According to simulation results, the spectrum reconstructed by the method I has less regularity and is whiter than that reconstructed by the spectral folding method. According to informal listening tests, the tonal noise is significantly reduced with the RELP vocoder using the method I. We also compared the two systems using the method I. The first system having two bands of 0 to 900 Hz and 1800 to 2100 Hz yields better speech quality than the second system. In simultation of the method III the multiband spectrally folded signal placed in the range of 0 to 2400 Hz is combined with the highpass filtered (2400 to

3600 Hz) pulsed excitation signal. Fig. 5 shows the spectrum of a two-band residual signal and the spectrum reconstructed by the method III. Fig. 6 shows the original and synthetic speech waveforms. According to informal listening tests, the synthetic speech with the method III yields no tonal noise and roughness. Here, we compare the methods I and III. Although the tonal noise has been reduced significantly with the method I, this method produces still a little tonal noise due to spectral regularity resulting from integer multiple baseband widths between upper and lower bands. By introducing the pulsed excitation signal into highest baseband, however, the method III elliminates the spectral regularity perfectly. Thus the method III is superior to the method I and the resulting synthetic speech is nearly identical to that of the original speech.

Next, the method of combining the spectrally folded signal and the pulsed excitation signal (i.e., the method II) has been studied for a 4.8 kbits/s RELP vocoder. We compared this method with the methods of using spectral folding, center clipping and hybrid excitation [9]. In simulation of the method II, we used only one baseband of 0 to 600 Hz and a 2-bit pitch-predictive ADPCM for residual coding.

Fig. 7 shows the spectra of the baseband signal of 0 to 600 Hz and the reconstructed excitation signal using the method II. We can notice that the spectrum reconstructed by the method II does not have the spectral regularity. Also, Fig. 8 shows the waveforms of original and synthetic speech by the method II. According to our listening tests, the nonlinear distertion method introduce considerable roughness in synthetic speech, and the spectral folding method produces some tonal noise, while the proposed method II gives slight tonal noise and buzz-like noise. The method II, however, produced subjectively the best performance. Also, this method is subjectively superior to the hybrid excitation method which introduces the static-like noise and buzz-like noise.

IV. CONCLUSION

In this work we have studied three methods of reconstructing a full-band excitation signal to improve the speech quality of the RELP vocoder. These include the multiband spectral folding method, the method of using both the spectrally folded and pulsed excitation signals, and the method of using both the multiband spectrally folded and pulsed excitation signals. According to the simulation results, the multiband spectral folding method is superior to the spectral folding and the nonlinear distortion methods, and the resulting synthetic speech has little tonal noise. Comparing the method II to the nonlinear distortion, spectral folding and hybrid excitation methods, it has less tonal noise and buzz-like noise. Finally, comparing the method III to the method I, the former yields better quality of synthetic speech. The synthetic speech of the method III has no tonal noise, and is nearly identical to the original speech at a bit rate of 9.6 kbits/s. In addition, we make a conclusion. Having the identical bandwidth constraint, the method III yields the best performance. Also the method I is superior to the method II at high hit rates. The method II, however, is the most suitable thing at low bit rates.

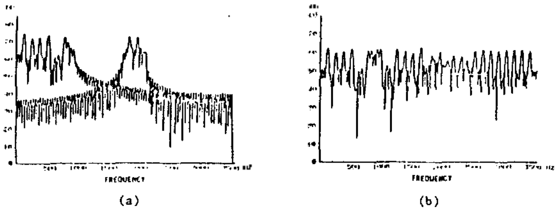

Fig. 5. Spectra of the two-band residual signal and the excitation signal reconstructed by the method II.
(a) Spectrum of the two-band residual signal (0 to 900 Hz and 1800 to 2100 Hz)
(b) Full-band spectrum reconstructed by the method III

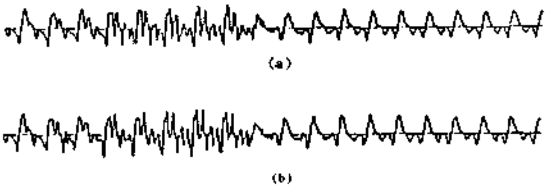

Fig. 6. Waveforms of (a) original speech and (b) synthetic speech of the RELP vocoding using the method III (Transmission rate: 9.6 kbits/s).

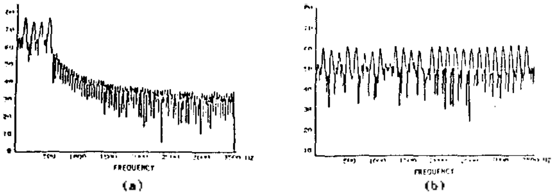

Fig. 7. Spectra of baseband signal and reconstructed excitation signal using the method II.
(a) Baseband spectrum
(b) Full-band spectrum reconstructed by the method II

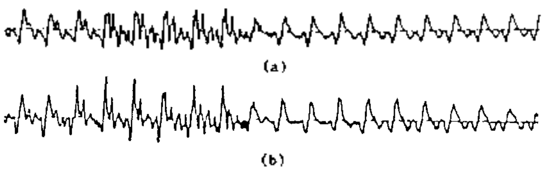

Fig. 8. Waveforms of (a) original speech and (b) synthetic speech of RELP vododer using the method II (Transmission rate: 4.8 kbits/s).

REFERENCES

1. Weinstein, C.J., "A linear prediction vocoder with voice excitation," in Proc. EASCON '75, pp. 30A-30G, Sept. 29-Oct. 1, 1975.
2. Un, C.K. and Magill, D.T., "The residual-excited linear prediction vocoder with transmission rate below 9.6 kbits/s," IEEE Trans. Commun., vol. COM-23, pp. 1466-1474, Dec. 1975.
3. Atal, B.S. and Remde, J.R., "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in Proc. 1982 IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 614-617, May 1982.
4. Un, C.K. and Lee, J.R., "On spectral flattening techniques in residual-excited linear prediction vocoding," in Proc. 1982 IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 216-219, Mar. 1982.
5. Sondhi, M.M., "New methods of pitch extraction," IEEE Trans. Audio Electroacoust., vol. AU-16, pp. 262-266, June 1968.
6. Un, C.K. and Lee, J.R., "A 9600 bit/s RELP vocoder split-band coding," IEEE '84, Amsterdam, Netherlands, pp. 1174-1178, May 1984.
7. Makhoul, J. and Berouti, M., "High-frequency regeneration in speech coding systems," in Proc. 1979 IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 428-431, Apr. 1979.
8. Viswanathan, V.R., Higgins, A.L. and Russel, W.H., "Design of a robust baseband LPC coder for speech transmission over 9.6 kbits/s noisy channels," IEEE Trans. Commun., vol. COM-30, no. 4, pp. 663-673, Apr. 1982.
9. Un, C.K. and Sung, W.Y., "A 4800 bps LPC vocoder with improved excitation," in Proc. 1980 IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 142-145, Apr. 1980.
10. Wong, D.Y., "On understanding the quality problems of LPC speech," in Proc. 1980 IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 725-728, Apr. 1980.