A NEW 2-KBIT/S SPEECH CODER BASED ON NORMALIZED PITCH WAVEFORM

Yuusuke Hiwasaki and Kazunori Mano
NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo 180, Japan

ABSTRACT

Speech coding at very low bitrate is useful for purposes such as voice communication over computer networks. However, speech coding at around 2.0 kbit/s is difficult for CELP coders while maintaining a high quality. In this paper, a speech coding model called 'normalized pitch waveform' and its quantization scheme are presented, aiming for effective compression coding of the 'voiced' speech. Listening tests has proven that an efficient and high quality coding has been achieved at bitrate 2.0 kbit/s, less than half of the FS1016. Furthermore, this paper discusses the disadvantage of the normalized pitch waveform and presents an alternative method of using non-normalized pitch waveform. Encoding of a transitional 'mixed' state between the 'voiced' and the 'unvoiced' state is discussed for further improvements.

1. INTRODUCTION

The rapid increase in use of computer networks resulted in needs for multimedia oriented computer communication. With improving quality of speech coder operating at 2-4 kbit/s, voice communication over computer network has become possible. Now, the need for low bitrate coder with ability to edit and modify is growing in those environment.

Recent studies in speech coders have paid special attention to bitrate reduction of 'voiced' speech. Since 'voiced' speech is relatively stationary over a period of time, pitch wise coding has been found effective. This has an advantage that the model is suitable for speed change and the pitch manipulation of the output speech.

There is an approach called *waveform interpolation* and coding using this model has been reported in [1], [2], [3], and [4]. In most of these studies, either the coding of the cycles and the interpolation is done in the frequency-domain, which leads to greater complexities.

This paper discusses a new coding model on normalized pitch waveform, a pitch wise coding method in time-domain, and the performance of the speech coder using this model, operating at 2.0 kbit/s. The quantization scheme is also described. Then another coding method using the non-normalized pitch waveforms is discussed and we compare its effect to normalized pitch waveform quantizing scheme. The use of 'mixed' state, apart from 'voiced' and 'unvoiced' state, is discussed.

2. OVERVIEW OF THE CODER

2.1. The encoder

The coder basically processes the input speech once in every 25ms, and operates linear predictive analysis and inverse filters the input signal with LPC parameters to acquire the residual signal. This 25ms frame is then categorized 'voiced' or 'unvoiced', depending on the periodicity of the speech. The periodicity is detected by the following:

$$\begin{cases} k_1/2 + \rho_{max} > \theta & ; \quad \text{voiced}, \\ k_1/2 + \rho_{max} < \theta & ; \quad \text{unvoiced}, \end{cases} \tag{1}$$

where ρ_{max} is the maximum value of the modified correlation ρ, k_1 is the first coefficients of the PARCOR parameters, and θ is the threshold value ranging between 0.5 and 0.8.

When a frame is detected as 'voiced', the encoder extracts a single pitch cycle from LPC inverse filtered residual signal. Then the length of the waveform (the pitch period) is normalized to a fixed vector length. This waveform is rotated and aligned to a standard pulse to adapt the phase. We call this single pitch cycle of residual as 'normalized pitch waveform'. LPC parameters, pitch, pitch cycle excitation vectors, and gain vector are quantized.

It has been reported that in 'unvoiced' speech, human ear is insensitive to the structure of the excitation of the LPC synthesis filter, as long as the power is updated frequently[5]. Applying this, for the 'unvoiced' frame, the encoder quantizes the LPC parameters and average power of 5ms sub-frame. Since only LPC parameters and a set of average power are quantized, the 'unvoiced' frame can sufficiently be quantized at fewer bits than that of 'voiced' one. Figure 1 presents a brief block diagram of the encoder.

2.2. The decoder

In 'voiced' frame, the decoder rotate and aligns the current normalized pitch waveform vector to the previous one, and linearly interpolates them in the time-domain to acquire the intermediate waveform. The intermediate waveforms are up/down-sampled according to the interpolated pitch, and are joined together to generate the excitation signal for the LPC synthesis filter. The synthesis filter's coefficients are also interpolated in LSP domain, to acquire a smooth evolution of the formant. For 'unvoiced' frame, the decoder excites the LPC synthesis filter with the Gaussian noise,

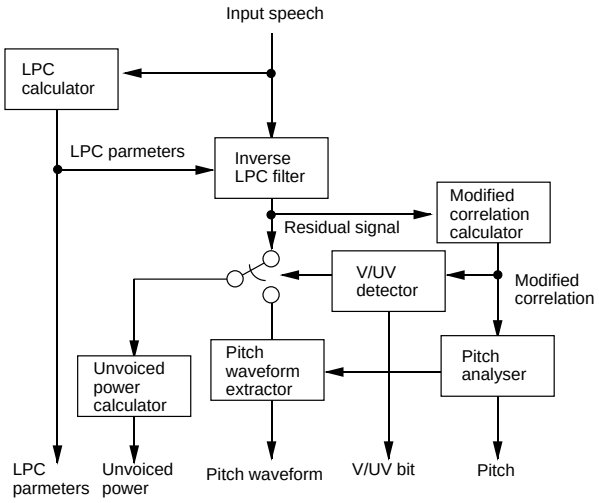

Figure 1. Block diagram of the encoder

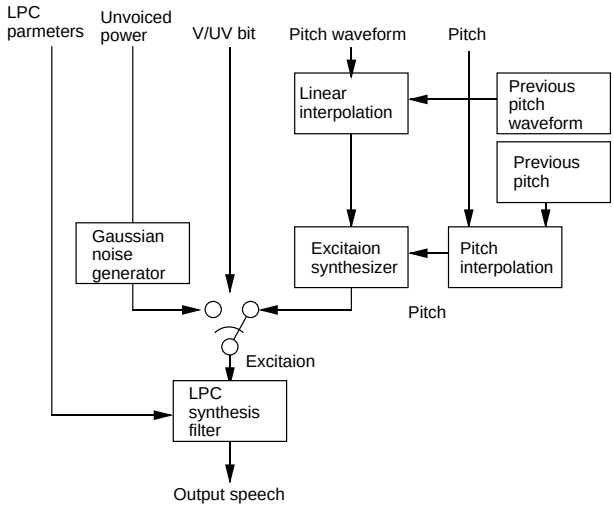

Figure 2. Block diagram of the decoder

and resulting sub-frame power is fixed to the corresponding value. Figure 2 presents a brief block diagram of the decoder.

2.3. Quantization of voiced speech using normalized pitch waveform

As described previously, residual signal is quantized pitch wise in the 'voiced' speech. The coder up/down-samples a pitch cycle (l samples) of residual signal to a fixed vector length of n, to normalize the pitch period. Then the coder quantizes the waveform using analysis-by-synthesis method like a CELP coder.

However, the distance calculation between target and the code vector of a single pitch cycle cannot be done in a manner of CELP coders. Traditionally in a CELP coder[6], the zero input response of the previous frame is subtracted from the target prior to the distance calculation. However, for quantization of single pitch waveform, the zero input subtraction cannot be applied. This is because the waveform are extracted every 25ms and are totally independent from the previous one. On the other hand, when exciting the LPC filter, the trailing zero input response of the current pitch cycle makes an effect to the next neighboring pitch cycle. Therefore, the effect of the zero input response of the code vector cannot be ignored.

To solve this problem, we decided to use the following pitch wise distance calculation measure. Since waveforms evolve slowly over a period of time, the zero input response of the current pitch cycle is similar to the previous neighboring one. For the distance calculation of a single pitch cycle, taking account of the zero input response of the current pitch cycle seems a fair compromise. Here, the target is not the remaining of input speech subtracted by the zero input response of the previous, but is defined as the following $\tilde{\boldsymbol{x}}$:

$$\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{H}}\tilde{\boldsymbol{r}}, \tag{2}$$

where $\tilde{\boldsymbol{r}}$ is an unquantized normalized pitch waveform, $\tilde{\boldsymbol{H}}$ is a matrix representing the normalized impulse response of the LPC synthesis filter $\tilde{\boldsymbol{h}}$.

Here, the elements of $\tilde{\boldsymbol{H}}$ matrix is the lower triangular matrix which elements are constituted from normalized impulse response of the LPC filter $\tilde{\boldsymbol{h}}$ $(\tilde{h}_0, \cdots, \tilde{h}_n)$. This $\tilde{\boldsymbol{h}}$ is calculated from up/down-sampling the pitch length (l samples) of the impulse response $\boldsymbol{h}$ $(h_0, \cdots, h_l)$ to vector length of n, following the same procedure as the normalized pitch waveform. Length of the target vector $\tilde{\boldsymbol{x}}$ is extended to m where $\tilde{x}_i (n < i < m)$ are trailing zero input response of $\tilde{\boldsymbol{H}}$. This calculates the effect of trailing zero input response of current pitch cycle to the next neighboring one, and selects a code as close as possible as to the original one. In ordinary CELP coders, $\boldsymbol{H}$ matrix is a lower triangular square $(n \times n)$ matrix, but now this matrix is extended to $(m \times n)$ matrix to acquire the target vector. This $\tilde{\boldsymbol{H}}$ matrix is presented as bellow:

$$\tilde{\boldsymbol{H}} = \left(\begin{array}{ccccccc}
\tilde{h_0} & & & & & & 0 \\
\tilde{h_1} & \tilde{h_0} & & & & & \\
\tilde{h_2} & \tilde{h_1} & \tilde{h_0} & & & & \\
\vdots & & & \ddots & & & \\
\vdots & & & & \tilde{h_0} & & \\
\tilde{h_{n-2}} & & & & \tilde{h_1} & \tilde{h_0} & \\
\tilde{h_{n-1}} & \tilde{h_{n-2}} & \cdots & \cdots & \tilde{h_2} & \tilde{h_1} & \tilde{h_0} \\
\hline
 & \tilde{h_{n-1}} & \tilde{h_{n-2}} & \cdots & \cdots & \tilde{h_2} & \tilde{h_1} \\
 & & \tilde{h_{n-1}} & & & & \tilde{h_2} \\
 & & & \ddots & & & \vdots \\
0 & & & & \tilde{h_{n-1}} & \cdots & \tilde{h_{m-n}}
\end{array}\right) \tag{3}$$

The elements above the horizontal line in Equation 3 is as same as the traditional $\boldsymbol{H}$ matrix, except that the elements are from normalized impulse response.

The general block diagram of normalized pitch waveform quantizer is shown in Figure 3. A series of codebook structure can be employed, such as quantization of the difference

from the previous pitch cycles, use of pulse and noise mixed excitation, and so on. In our implementation, we simply designed a two-stage quantizer using two trained noise codebooks. The distance calculation between excitation code vector and the normalized pitch waveform is performed as follows:

$$D = \left\| \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{H}}(g_0 \boldsymbol{c}_0 + g_1 \boldsymbol{c}_1) \right\|^2, \quad (4)$$

where $\boldsymbol{c}_0$ and $\boldsymbol{c}_1$ are the normalized pitch code vectors with length of n, and the g_0 and g_1 are the gain factors of the code vectors respectively.

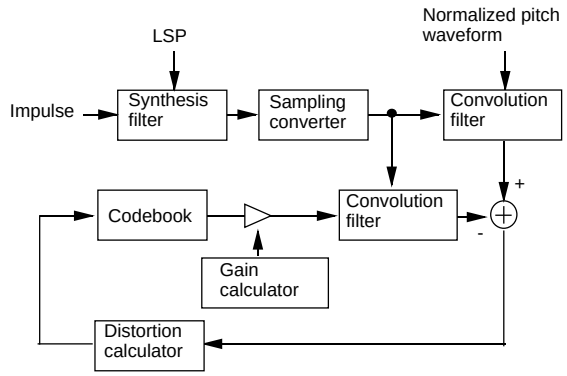

Figure 3. Block diagram of normalized pitch waveform quantizer

2.4. The performance of the proposed coder

A mean opinion score experiment was performed for the proposed coder at a bitrate of 2.0 kbit/s. For each coder or condition, a set of 10 IRS-filtered speech was evaluated by 16 non-experts. Table 1 compares the MOS result converted to equivalent Q-values. The result shows that the quality is better than that of FS 1016 at 4.8 kbit/s[7] and nearly equivalent to MNRU 20dB.

Table 1. Result of listening test

Coding method	Q_{eq} [dB]
Original	33.06
G.726 (32kbit/s)	23.67
G.726 (24kbit/s)	18.68
FS1016	18.26
Proposed	19.55
LPC-10e	12.58

There is a tendency that the proposed coder has a slightly buzzy quality, but this results in clear speech, compared to noisy character audible in other CELP based coders.

3. AN ALTERNATIVE USE OF NON-NORMALIZED PITCH WAVEFORM

3.1. Problems of normalized pitch waveform

The above quantization of normalized pitch waveform is very effective for high pitched speech generally by female speakers. However, for the lower-pitched speech by male speakers, the output speech tends to lack the power in the higher frequency. We suspected that this results from the codebook training. Through low-pass filtered up-sampling of a pitch cycle, the normalized waveform has narrower bandwidth of spectrum power. Training such a vector naturally results in a codebook with limited frequency range. By de-normalizing the code vector may induce the above result.

Another problem is that the coder has a relatively high complexity. Since the the vector length n is usually longer than that of pitch of an ordinary speech, the pitch waveform vector is usually up-sampled for the normalization. Moreover, as the distance calculation is done with extended $\boldsymbol{H}$ matrix, the convolution is more complex than an ordinary CELP coder.

3.2. Quantizing non-normalized pitch waveforms

To solve the above problems, we used non-normalized pitch waveforms. While the above stated method normalizes the pitch of a single pitch excitation, the pitch excitation is now quantized in the real time- domain. Thus, the distance calculation is done in the same domain.

In the codebook search, code vectors are truncated beyond the pitch length l, and only the pitch length is evaluated. The target is calculated as the following, using non-normalized $\boldsymbol{r}$ vector.

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{r} \quad (5)$$

Still, $\boldsymbol{H}$ is an extended rectangular matrix similar to Equation 3. Here, the elements are constituted from non-normalized impulse response of LPC filter $\boldsymbol{h}$ $(h_0, \cdots, h_l)$. The distance calculation for the codebook search is done similarly to Equation 4.

This method has the advantage that it reduces the complexity of the coder because the normalization steps has been abbreviated. Furthermore, the pitch is used for defining code vector and this contributes to a greater quantization efficiency.

A two-stage quantizer has been implemented using non-normalized pitch waveform and listening tests were performed. The results showed an improvement of nearly 1dB respect to the equivalent Q-value.

4. 'MIXED' STATE

Encoding speech with only two ('voiced' and 'unvoiced') state tends to prove itself noisy. The abrupt switching between 'voice' and 'unvoiced' state, and the occasional determination error are the reasons for this. A 25ms frame may be long enough to contain both 'voiced' and 'unvoiced' part of speech.

We introduced the third 'mixed' state to the coder, a transitional state between the 'voiced' and the 'unvoiced'. The basic idea is to create a state in which periodical and aperiodic components of speech co-exist. As a tentative approach, we decided to take the following strategy.

The periodicity decision is made according to the follow-

ing:

$$\begin{cases} \theta_{high} \leq k_1/2 + \rho_{max} & ; \text{ voiced,} \\ \theta_{low} \leq k_1/2 + \rho_{max} < \theta_{high} & ; \text{ mixed,} \\ k_1/2 + \rho_{max} < \theta_{low} & ; \text{ unvoiced,} \end{cases} \quad (6)$$

where θ_{low} and θ_{high} are the lower and higher thresholds respectively, ranging around the value of the previously shown θ (in Equation 1). A rule based decision has been designed, so the 'mixed' frame is always placed between 'voiced' and 'unvoiced' frame, as a transitional frame. Since mixed state is always placed prior or next to the 'voiced' frame, the information of the neighboring 'voiced' frame can be used to encode the periodic component of the mixed state.

The coder searches the position of the pitch waveform which is similarly shaped to the one extracted from the neighboring 'voiced' frame. When matched, the neighboring pitch waveform is multiplied by ideal gain and subtracted from the residual signal at that position.

When all the voiced component is matched and subtracted, remaining signal can be considered as the aperiodic component of the speech. The average power of remaining signal is calculated by sub-frame, in the manner of the 'unvoiced' frame, as described above. The LPC parameters, average pitch and the average ideal gain of the periodic component, and for he aperiodic component the set of average power per sub-frame are quantized.

In decoder, the neighboring pitch waveform is repeated according to the transmitted pitch and power adapted Gaussian noise is composed to form the excitation. The resulting excitation signal is passed through the LPC synthesis filter to acquire the output speech.

5. DISCUSSION

Currently, the coder is designed as a variable rate codec and the bit-allocation is as Table 2. The pitch waveform is quantized with two excitation codebook and a gain codebook containing power information of the excitation.

Table 2. Bit allocation of the coder

	voiced	mixed	unvoiced
State	2	2	2
LPC	20	20	20
Pitch	7	7	–
Pitch waveform	21	–	–
Power	–	7	7
Total	50	36	29
Bitrate	2000bit/s	1440bit/s	1160bit/s

The state bits are for transmitting the state of the frame, either 'voiced', 'mixed', 'unvoiced', or 'silent'. The silent frame is excluded from this table, since the coder only transmits the 2 state bits, when speech is 'silenced'. By this bit-allocation, the average bitrate of 20 sets of Japanese speech became 1.325 kbit/s.

The above table shows that additional bits can be used for 'unvoiced' and 'mixed' frames. Particularly for 'mixed' speech, only a preliminary implementation was done. Still, the listening tests showed nearly 1dB improvement in equivalent Q-value, and we believe this encourages further exploitation of using 'mixed' states.

The quantization is only tested with two-stage trained codebook. The use of other quantization scheme may enhance the quantization efficiency. For example, use of previous pitch waveforms and encoding with mixed excitation of pulse and noise are the subjects of the future experiments.

The complexity is still a problem. Since the $\boldsymbol{H}$ matrix is extended as described before, the calculation steps for matrix convolution is large. The use of truncated impulse response $\boldsymbol{h}$ should be tested for lesser complexity, but the quality of output speech should be a trade off.

6. CONCLUSION

A new speech coder at bitrate of 2.0 kbit/s based on normalized pitch waveform model was presented. We found that this model is an effective compression method of 'voiced' speech. The distance calculation for quantization of excitation with single pitch length is done with extended impulse response matrix. The subjective quality of the proposed coder is found superior to that of FS 1016 and G.726 at 24 kbit/s. An alternative use non-normalized pitch waveforms is examined and this showed some improvement, particularly for the lower pitched speech. The use of 'mixed' state as a transitional state between 'voiced' and 'unvoiced', showed a promising result for further improvement.

ACKNOWLEDGMENT

The authors wish to thank Dr.Nobuhiko Kitawaki and Takao Kaneko for guiding our research. We also would like to thank the members of the speech coding group, especially Dr.Takehiro Moriya and Hitoshi Ohmuro for their helpful advice and valuable discussion.

REFERENCES

[1] W. B. Kleijn: "Encoding Speech Using Prototype Waveforms," IEEE Trans. on Speech and Audio Coding, Vol.1, No.4, pp.386–399, 1993.

[2] I. S. Burnett et. al: "A Mixed Prototype Waveform / CELP Coder for sub 3kb/s," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.II-175–178, 1993.

[3] G. Yang et. al: "Voiced Speech Coding at Very Low Bit Rates based on Forward_Backward Waveform Prediction (FBWP)," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.II-179–182, 1993.

[4] J. C. De Martin, et al.: "Mixed-Domain Coding of Speech at 3kb/s," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.216–219, 1996.

[5] G. Kubin, et al.: "Performance of Noise Excitation for Unvoiced Speech," Proc. IEEE Workshop on Speech Coding for Telecommunications, pp.35–36, 1993.

[6] M. R. Schroeder and B. S. Atal: "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.937–940, 1985.

[7] J.P.Campbell ,Jr, et al.: "The DoD 4.8 Kbps Standard (Proposed Federal Standard 1016)," Advances in Speech Coding, Chapter 4.1, Kluwer Academic Publishers, 1990.