# A NOVEL WI DECODER FOR THE SEGMENTED FRAME DECODING IN THE TEXT-TO-SPEECH SYNTHESIZER

Kyungjin Byun, Nak-Woong Eum and Hee-Bum Jung
*Electronics and Telecommunications Research Institute (ETRI), Daejeon, 305-700, Korea*

Keywords: Waveform interpolation, Speech coding, Speech synthesizer, Text-to-speech.

Abstract: The implementation of a high quality text-to-speech (TTS) requires huge storage space for a large number of speech segments, because current TTS synthesizers are mostly based on a technique known as synthesis by concatenation. In order to compress the database in the TTS system, the use of speech coders would be an efficient solution. Waveform interpolation (WI) has been shown to be an efficient speech coding algorithm to provide high quality speech at low bit rates. However, the speech coder used in a TTS system has to be different from the one used in communication applications because the decoder in the TTS system should have an ability to decode segmented frames. In this paper, we propose a novel WI decoder scheme that can handle the segmented frame decoding. The decoder can reconstruct a good quality speech even at the concatenation boundary, which is effective for the TTS system based on a synthesis by concatenation.

## 1 INTRODUCTION

In recent years, various speech coding algorithms have been widely used in many applications such as mobile communication systems and digital storage systems for the speech signal to be represented in lower bit rates while maintaining its quality. The WI coding algorithm has been known as one of the good coding algorithms producing the good quality speech even at the below 4 kbps rates. Most speech coders operate on narrow bandwidth limited to 200 - 3400 Hz. However, as mobile systems are evolving from speech-dominated services to multimedia ones, the advent of the wideband coder becomes highly desirable because it is able to provide higher quality speech. The wideband speech coder extends the audio bandwidth to 50 - 7000 Hz in order to achieve the high quality both in the sense of speech intelligibility and naturalness.

Although WI speech coder is classified into parametric one, it is able to provide high perceptual quality at low bit rates (Kleijn & Haagen, 1995). Most literatures for the WI coding have been concentrated on the narrow band speech coding (Kleijn, 1993; Gottesman & Gersho, 2001), but recent researches (Ritz, et al. 2002, 2003) show the potential of applying a WI algorithm to wideband speech signal. In the WI coding, there are four parameters to be transmitted. They are the linear prediction (LP) parameter, the pitch value, the power and the characteristic waveform (CW). The CW parameter is decomposed into a slowly evolving waveform (SEW) and a rapidly evolving waveform (REW). Since the SEW and REW have very distinctive requirements they should be quantized separately to enhance the coding efficiency.

On the other hand, the implementation of a high quality TTS requires huge storage space for a large number of speech segments, because most TTS synthesizers are based on a technique known as synthesis by concatenation. In order to compress the database in the TTS system, which usually consists of wideband speech, the use of speech coders is an efficient solution (Vercken, et al., 1997). However, the speech coder used in a TTS system has some differences compared to the one used in communication applications. In the communication systems, the speech coder continuously performs the encoding and decoding procedure to process continuous speech signals. Once the speech coder operates, it maintains its normal operating mode. Therefore, the speech coder always can keep the parameters of the previous frame and maintains filter memories required for performing the next frame.

However, the decoding scheme for the TTS system has to decode segmented frames of the arbitrary intervals to reconstruct segmented speech

Byun K., Eum N. and Jung H. (2010).
A NOVEL WI DECODER FOR THE SEGMENTED FRAME DECODING IN THE TEXT-TO-SPEECH SYNTHESIZER.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 151-154
DOI: 10.5220/0002978901510154

signals, which will be used as an input to the TTS system. In this case, the reconstructed speech signal generated by a common speech coder is seriously degraded especially at the initial parts of the signals because the decoder cannot have previous parameters at the start frame. In order to cope with this problem, we propose a novel WI decoder scheme that can handle the segmented frame decoding. Therefore, it can reconstruct a good quality speech even for the segmented frame decoding, which is effective for the TTS system based on a synthesis by concatenation.

The outline of this paper is as follows. Firstly, in section 2, we describe the overview of WI speech coding algorithm. In section 3, we propose the novel WI decoder scheme for the segmented frame decoding with the good quality of the reconstructed speech. Then, the experimental results are discussed in section 4. Finally, conclusions are made in section 5.

## 2 WI SPEECH CODING

The WI coding algorithm has been extensively and steadily developed since it was first introduced by Kleijn (Kleijn & Haagen, 1995). The encoder block diagram of the WI speech coder is shown in Fig. 1.

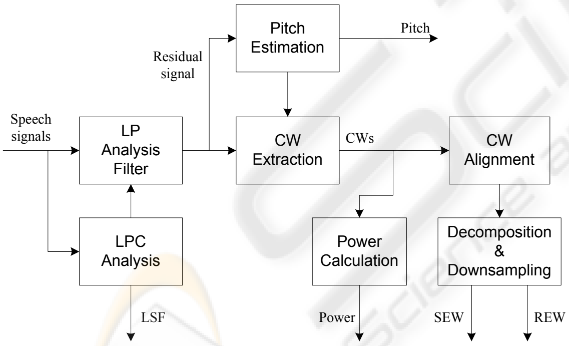

Figure 1: The encoder block diagram of WI speech coder.

The WI coder firstly performs LP analysis once per frame for the input speech. The LP parameter set is converted into the line spectrum frequency (LSF) for the efficient quantization and usually vector quantized using various quantization techniques. The pitch estimation is performed in the linear prediction residual domain. In the WI paradigm, the accuracy of this pitch estimator is very crucial to the performance of the coder. After the pitch is estimated, WI coder extracts pitch-cycle waveforms which are known as CWs from the residual signal at a constant rate. These CWs are used to form a two-dimensional waveform which evolves on a pitch synchronous nature. The CWs are usually represented using the discrete time Fourier series (DTFS) as follows:

$$u(n,\phi) = \sum_{k=1}^{\lfloor P(n)/2 \rfloor} \left[ a_k(n)\cos(k\phi) + b_k(n)\sin(k\phi) \right] \quad (1)$$

where $\phi = \phi(m) = 2\pi m / P(n)$, $0 \le m < P(n)$ , $a_k$ and $b_k$ are the DTFS coefficients, and $P(n)$ is the pitch value. The CWs are used to construct a two dimensional surface $u(n,\phi)$ to display the shape of the discrete time waveform along the phase $\phi$ axis and the evolution of the shape along the discrete-time $n$ axis.

The extraction procedure performed in the LP residual domain provides a DTFS description for every extracted CW. Since these CWs are generally not time-aligned, the smoothness of the surface in the time direction should be maximized. This can be accomplished by aligning the extracted CW with the previously extracted CW by introducing a circular time shift to the current one. Since the DTFS description of the CW enables to regard the CW as a single cycle of a periodic signal, the circular time shift is equivalent to adding a linear phase to the DTFS coefficients. The CWs are then normalized by their power, which is quantized separately. The main motivation of this normalization is to separate the power and the shape in CWs so that they can be quantized separately to achieve higher coding efficiency.

This two-dimensional surface is decomposed into two independent components, i.e., SEW and REW, via low pass filtering prior to quantization. The SEW and the REW are down sampled and quantized separately. The SEW component represents mostly the periodic (voiced) component while the REW corresponds mainly to the noise-like (unvoiced) component of the speech signal. Because these components have different perceptual properties, they can be exploited to increase coding efficiency in the compression. In other words, the SEW component requires only a low update rate but has to be described with a reasonably high accuracy whereas the REW component requires a higher transmission rate but even a rough description is perceptually accurate. This property of the CW suggests that low pass filtering of the CW surface leads to a slowly evolving waveform. The rapidly evolving part of the signal can be obtained by simply subtracting the corresponding SEW from the CW.

## 3 A NOVEL WI DECODER SCHEME

In the conventional WI decoder used in the communication areas, the received parameters are the LP coefficients, the pitch value, the power of the CW, the SEW and REW magnitude spectrum. The decoder can obtain a continuous CW surface by interpolating the successive SEW and REW and then recombining them. After performing the power de-normalization and subsequent realignment, the two-dimensional CW surfaces are converted back into the one-dimensional residual signal using a CW and a pitch value at every sample point, which can be obtained by linear interpolation. This conversion process also requires the phase track estimated from the pitch value at each sample point. The reconstructed one-dimensional residual signal is used to excite the linear predictive synthesis filter to obtain the final output speech signal.

However, for the decoder used in the TTS system, the final reconstructed signals are obtained by decoding the segmented frames and concatenating them. Therefore, if the medium part of the concatenated speech is decoded by using the conventional decoder, the final reconstructed speech will be seriously distorted especially at the concatenation boundary. For the segmented frame decoding, if the decoder can use the previous parameters at the start frame, the distortion due to mentioned above can be dramatically decreased. Therefore, in this paper we present the novel decoder scheme for the segmented frame decoding to reduce the distortion at the concatenation boundary by utilizing the previous parameters. The block diagram of the novel decoding scheme is shown in Fig. 2.

As shown in Fig. 2, in order to decrease the distortion for the segmented frame decoding, the decoder utilizes all the previous parameters; the (n-1) frame's LSFs, pitch, CW power, SEW and REW magnitudes, where the current frame number is n. In the decoder, the (n-1) frame's CWs are required for processing the initial state at the start frame. Since the CWs are generated by combining the SEWs and REWs, both the (n-2) frame's SEWs and REWs are necessary to generate the (n-1) frame's CWs. For the continuous frame decoding, the previous CWs are always available at the time of the current frame processing since they are preserved in the decoder during the decoding procedure. However, for the segmented frame decoding, since the previous CWs are not available at the start frame, the decoder should make the previous CWs using the (n-1) and (n-2) frame's SEWs and REWs.

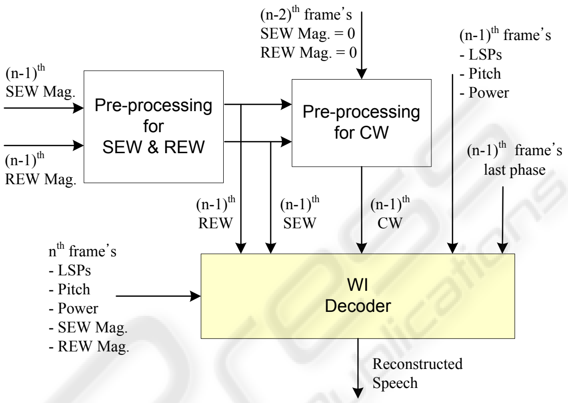

Figure 2: A novel WI decoding scheme.

In addition to these five previous parameters, the phase instant is also exploited in the proposed decoding scheme for the segmented frame decoding. The phase instant is calculated at the phase estimation. The phase instant is used for obtaining the one-dimensional residual signal from the two-dimensional characteristic waveform. During the procedure the phase instant at every sample is calculated and the last one is stored for the next frame processing. In the proposed scheme, the previous phase instant at the start frame is also exploited with the five previous parameters mentioned above. Adopting the phase instant for the segmented frame decoding dramatically improves the performance of the final reconstructed speech segments.

## 4 EXPERIMENTAL RESULTS

The waveforms in the Fig. 3 are generated by decoding the segmented frames and then concatenating two reconstructed speech signals. In this figure, the solid vertical line indicates the concatenation boundary. Waveform (a) is generated by the continuous frame decoding; hence this waveform can be considered as the original waveform compared to the waveforms generated by the segmented frame decoding.

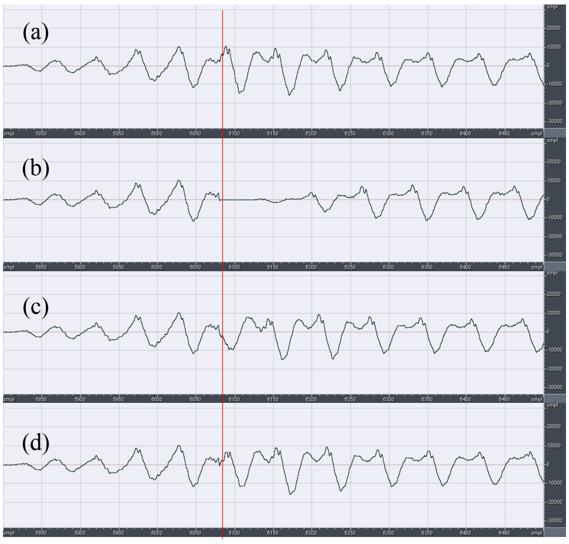

Figure 3: Reconstructed waveforms: (a) continuous frame decoding, (b) segmented frame decoding with no previous parameters, (c) with 5 previous parameters, (d) and adding phase information.

In the waveform (b) generated by using no previous parameters, the beginning part of the reconstructed speech generated from the second segmented frames is seriously distorted. The waveform (c) is for using all five previous parameters. The shape of this waveform is mostly similar to the waveform (a), but the phase distortion begins from the concatenation boundary. However, the waveform (d) utilizing the previous phase information together with all other parameters is almost same as the waveform (a) in both shape and the phase. Therefore, the SNR measurement between (a) and (d) is much higher than that between (a) and (c). The SNR performance is given in Table 1.

Table 1: Performance of the novel decoding scheme.

| Previous Parameters | | SNR | segSNR |
|---|---|---|---|
| No parameter | | 5.502 | 4.413 |
| 5 parameters | | 9.167 | 6.033 |
| 5 parameters and phase (quantization bits) | 4 bits | 20.187 | 15.772 |
| | 5 bits | 23.565 | 18.972 |
| | 6 bits | 23.158 | 24.103 |
| | 7 bits | 27.697 | 29.273 |

Although the absolute SNR value is not important meaning for the parametric coder like a WI coder, it still can be considered as a useful measure of the relative performances between two waveforms. In Table 1, we can find that the proposed scheme improves the performance of the reconstructed speech signals. Especially, the performance is dramatically improved together with the phase information. The previous phase instant value should be quantized in order to utilize it for the segmented frame decoding in the TTS system. The simulation result according to the bit allocation for the phase instant is also provided in this table.

## 5 CONCLUSIONS

In this paper, we propose the efficient WI decoder scheme for the segmented frame decoding, which would be used in the TTS synthesizer to reconstruct the speech signals from the compressed speech segments. The proposed decoding scheme is able to suppress the distortion at the concatenation boundary by utilizing the previous parameters. Moreover, the phase information is also adopted to additionally improve the quality of final reconstructed speech. It is found that the adoption of the phase instant makes it possible to improve the performance dramatically.

## ACKNOWLEDGEMENTS

This work was supported by the IT R&D program of MKE/IITA. [2006-S-048-01, Embedded DSP platform for audio and video signal processing]

## REFERENCES

Kleijn, W. B., Haagen, L. J., 1995. *Waveform interpolation for coding and synthesis: Speech coding and synthesis*. Elsevier Science B. V.

Kleijn, W. B., 1993. Encoding Speech Using Prototype Waveforms. *IEEE Tans. on Speech and Audio Processing*, 1(4), pp. 386-399.

Gottesman, O., Gersho, A., 2001. Enhanced Waveform Interpolative Coding at Low Bit-Rate. *IEEE Trans. on Speech and Audio Processing*, 9(8), pp. 786-798.

Ritz, C. H., Burnett, I. S., Lukasiak, J., 2002. Extending waveform interpolation to wideband speech coding. *Proc. IEEE workshop on Speech Coding*, pp. 32-34.

Ritz, C. H., Burnett, I. S., Lukasiak, J., 2003. Low bit rate wideband WI speech coding. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 804-807.

Vercken, O., Pierret, N., Dutoit, T., Pagel, V., Malfrere, F., 1997. New Techniques for the Compression of Synthesizer Databases. *Proc. IEEE Int. Symp. on Circuits and Systems*, pp. 2641-2644.