AN EFFICIENT SEGMENT-BASED SPEECH COMPRESSION TECHNIQUE FOR HAND-HELD TTS SYSTEMS

Chang-Heon Lee, Sung-Kyo Jung, Thomas Eriksson*, Won-Suk Jun** and Hong-Goo Kang*

Dept. of Electrical and Electronic Eng., Yonsei University, Korea
*Dept. of Signals and Systems, Chalmers University of Technology, Sweden
**R&D Center, Voiceware Co., Ltd
[leech, hgkang]@dsp.yonsei.ac.kr

10.21437/Interspeech.2006-60

Abstract

This paper proposes a novel segment-based speech coding algorithm to efficiently compress the database for concatenative text-to-speech (TTS) systems. To achieve a high compression ratio and meet the fundamental requirements of concatenative TTS synthesizers, i.e. partial segment decoding and random access capability, we adopt a modified analysis-by-synthesis scheme. The spectral coefficients are quantized by a length-based interpolation method and excitation signals are modeled with both non-predictive and predictive approaches. Considering that pitch pulse waveforms of a specific speaker show low intra-variation, the conventional adaptive codebook for pitch prediction is replaced by a speaker dependent pitch-pulse codebook. By applying the proposed algorithm to a hand-held Korean TTS system, we verify that the proposed coder provides a compression ratio of about 1/13, a low complexity of around 1.2 WMOPS, and random access capability.

Index Terms: TTS synthesizer, segment-based speech coding, speaker-dependent pitch pulse codebook, hybrid coding structure

1. Introduction

Modern state-of-the-art text-to-speech (TTS) systems generate synthesized speech by concatenating phonetically labeled speech segments [1]. To achieve high quality synthetic speech, those systems need a large corpus of one speaker. Therefore, an efficient compression of the speech database can lead to synthesis of high quality speech, given the limited system memory. The compression approach is also very efficient for the application of the TTS systems designed for hand-held devices having limitation on memory usage. In addition, the decoding process should be realized with very low complexity to implement it in real-time. Our TTS system for hand-held devices requires a memory size of around 8.0 MB. Since the size of 8 kHz uncompressed speech database including segments is about 70 MB and a unit information occupies around 1.5 MB in our TTS system, a compression ratio more than 1/12 is required. In addition, to implement the TTS synthesizer in real-time it should have a low decoding complexity, e.g. less than 1.5 WMOPS(weighted million operations per second)[2]. Speech compression algorithms for the TTS applications could be different from the conventional algorithms used for speech communication systems due to several characteristics of TTS systems such as random access capability but without requiring tight bound for encoding complexity.

Vrecken et al. proposed a compression technique for the database of multi-band resynthesis overlap add (MBROLA) [3] synthesizer by using several stochastic codebooks [4]. However, it should have problems related to random access capability because of large amount of error at the beginning of synthesized voiced segments. In [5] a speaker dependent compression technique was proposed considering the fact the database for TTS systems is recorded by a single speaker, thus a more flexible design was possible because its encoding complexity could be ideally unlimited. Assuming that the pitch pulse waveforms of a specific speaker have low intra-variation, a non-predictive coding algorithm with the speaker dependent pitch pulse codebook was employed for random access capability. In addition, to improve the coding efficiency, a hybrid structure combining non-predictive and predictive coding methods was proposed.

This paper further enhances the performance in term of quality and complexity by taking new functional modules for spectral quantization and excitation modeling. Assuming that a signal has low spectral variations within a synthesized segment, a length-based interpolation method is used to efficiently quantize spectral parameters. To improve the coding efficiency by reducing the required bits for non-predictive frames, we propose new coding types based on a location of the first pitch pulse. The first pitch pulse needed for constructing the memory buffer for the adaptive codebook in successive predictive frames is modeled with a newly designed speaker dependent pitch pulse codebook. For an efficient gain quantization in predictive frames, a safety-net method is utilized.

To evaluate the performance of the proposed algorithm in terms of quality and complexity, we measure the perceptual evaluation of speech quality (PESQ) scores [6] and WMOPS[2], respectively. The experiments confirm that the proposed coder provides low complexity and random access capability. The proposed algorithm is also successfully implemented into a Korean TTS system.

2. Proposed algorithm for TTS synthesis

Since a TTS system concatenates phonetically-labeled speech segments in a random access and partial manner, the proposed speech coder should be able to independently synthesize each segment. Assuming that spectral information changes slowly within one segment, the spectral coefficients are quantized by using a length-based interpolation method similar to the method introduced in [7]. To meet the random access requirement, excitation signals at the beginning of the segment including the first pitch-pulse are quantized using a non-predictive coding scheme. To further improve the coding efficiency, the proposed coder also utilizes a predictive coding method when they are applicable, i.e. after the look-back

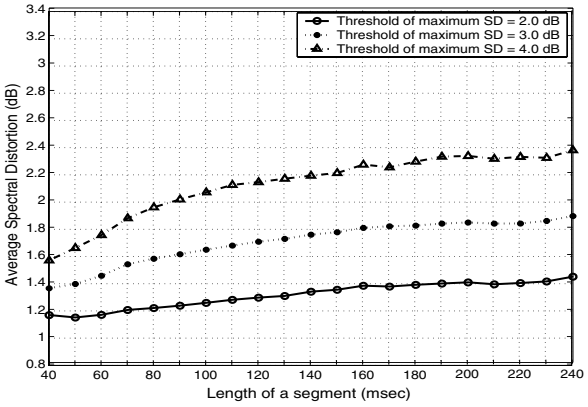

Figure 1: Average spectral distortion (dB) for various maximum threshold values.

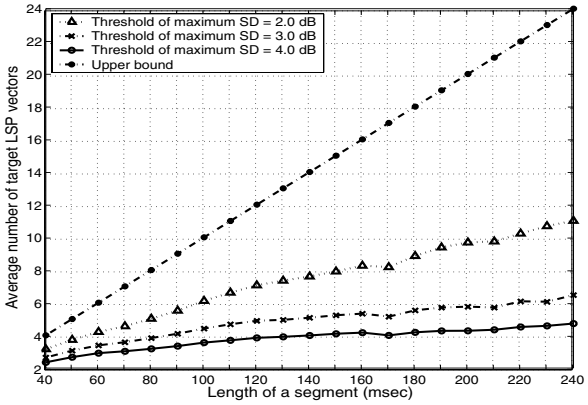

Figure 2: Average number of determined target LSP vectors within one segment for various maximum threshold values.

memory buffer is full due to the non-predictive coding.

2.1. Spectral Information

Since spectral parameters show high correlations among successive time frames, the redundancies caused by spectral quantization can be increased if they are quantized at every frame interval. A good approach to reduce the redundancy is using the temporal decomposition (TD) method [7]. The TD method adopts a rate-distortion criterion to efficiently quantize spectral parameters. Eq.(1) represents the temporal decomposition using a length-based interpolation method. In an off-line processing like our system, interpolation functions can be obtained by utilizing all LSP vectors in training corpus.

$$\hat{a}_i(n+n_k) = \phi_{i,N}(n)a_{i,k} + \{1-\phi_{i,N}(n)\}a_{i,k+1}, \quad 0 \le n \le n_{k+1}-n_k = N,\ 1 \le i \le p, \tag{1}$$

where n_k is the location of the kth target LSP vector, p is the LSP order, $a_{i,k}$ is the ith LSP coefficient of the kth target LSP vector, and $\phi_{i,N}(n)$ is the N-length interpolation function for the ith LSP coefficient.

The target LSP vectors are determined based on a criterion of minimizing the number of bits, while constraining maximum spectral distortion to be below a given threshold. To determine the target LSP vectors to be quantized within one segment, we modify the algorithm as shown in eq.(2). Let $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_K]$ denote candidates of the target LSP vectors within one segment, which are obtained from linear predictive analysis at every 10 ms. By using the criterion of eq.(2), we determine locations of the target LSP vectors, $\mathbf{L} = [l(1), l(2), \cdots, l(M)]$, where M is the number of selected target LSP vectors within one segment.

$$l(m) = \underset{l(m-1)<k\le K}{\arg\max} \{D_{\max}(\mathbf{a}_k, \mathbf{a}_{l(m-1)}) \le d_{\text{thres}}\}, \tag{2}$$

where $l(1) = 1$, $D_{\max}(\mathbf{a}_k, \mathbf{a}_{l(m-1)})$ is the maximum spectral distortion (SD) between the interpolated and the original LSP vectors in the interval $\langle n_k, n_{l(m-1)} \rangle$, and d_{thres} denotes a threshold value defining the maximum SD. The interpolated LSP vectors are obtained from eq.(1). Eventually all the selected 10-th order LSP vectors for 8kHz speech are quantized by a 24-bit split vector quantization method [8]. Using 49216 female speech segments sampled at 8kHz of the *Voiceware* database [9], we measure the average spectral distortion for different threshold values, d_{thres}. Fig.1 shows the average spectral distortion depending on various maximum threshold values. The proposed quantization method keeps consistency on spectral distortions although the length of segment increases. To estimate required bits for the proposed quantization scheme, the average number of selected target LSP vectors within one segment are computed by varying threshold values. Fig.2 shows the results. In this figure, 'Upper bound' means the quantization method of LSP parameters extracted at every 10ms interval. For example, in case of using the threshold value of 2.0dB, the number of target LSP vectors to be quantized is nearly half of the number of 'Upper bound'. The quantization scheme with lower threshold can have better performance in terms of spectral distortion, but it needs more bits due to an increasing number of target LSP vectors. Thus, trade-off is necessary.

2.2. Excitation Signal Modeling

To fulfill the requirement of random access capability, excitation signals should be independently quantized on a segment-by-segment basis. Conventional speech coders utilize the adaptive and stochastic codebooks to model excitation signals. Since the adaptive codebook efficiently models pitch-pulse waveforms in voiced speech regions, it plays a very important role in speech quality.

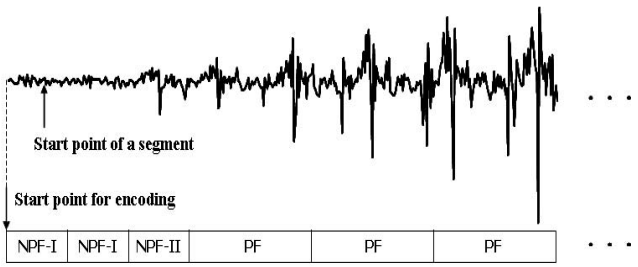

Figure 3: Proposed coding structure for excitation signals

Table 1: Weighted segmental SNR (dB) and codebook memory size (kbyte) for different quantization schemes of the first pitch pulse.

Quantization scheme	Weighted seg. SNR (dB)	Memory size (kbyte)
11-bit VQ	6.38	163.8
12-bit VQ	6.82	327.7
13-bit VQ	8.01	655.4

The adaptive codebook consists of excitation information of the previous frame, which recursively affects the overall speech quality. However, in our application, since each segment starts to be modeled without any knowledge of previous excitation information, the adaptive codebook causes the quality degradation of synthetic speech.

To improve the performance while keeping bit-rate as low as possible, we propose three types of coding schemes : non-predictive unvoiced signal modeling(5ms NPF-I type), the first pitch-pulse modeling(5ms NPF-II type), and predictive signal modeling(10ms PF type). Fig.3 shows the proposed coding structure for excitation signals. Before modeling the first pitch pulse, only a stochastic codebook is used for modeling non-predictive unvoiced signal. In the procedure of modeling the first pitch-pulse, a speaker-dependent pitch-pulse codebook replaces the conventional adaptive codebook for pitch prediction. To further improve the coding efficiency, the proposed coder flexibly combines non-predictive and predictive type methods.

2.2.1. The first pitch pulse modeling with a speaker dependent pitch-pulse codebook

We propose a speaker-dependent pitch-pulse codebook to model the first pitch pulse. To reduce redundancy and improve the compactness of the codebook, we design the codebook with peak-aligned pitch pulse waveforms. The peak point of the first pitch pulse is determined by using a pitch marking information given from the TTS system. After determining the peak point, M, the first pitch pulse of 5 ms long is extracted within one segment. Considering that the minimum pitch period is generally 2.5 ms, the length of the first pitch pulse is determined to include only one pitch pulse.

Based on the generalized Lloyd algorithm (GLA) [10], the codebook is designed by using peak-aligned pitch pulses in the training corpus. Table 1 represents performance of the proposed codebook in terms of weighted segmental SNR (dB) [11] and required memory size (kbyte). Since our TTS system has limitation on an allowable run-time memory size of less than 500 kbyte, we utilize the 12-bit VQ method to model the first pitch pulse. To enhance the modeling accuracy, the remaining residuals are quantized by using a stochastic codebook.

2.2.2. Predictive modeling of excitation signals

After the front regions of each segment are modeled with two non-predictive coding schemes, we may employ a predictive coding method after them. To improve the coding efficiency, the conventional adaptive codebook is used for pitch prediction, and we propose a safety-net gain quantization scheme.

Fig. 4 depicts a block diagram of the proposed safety-net gain

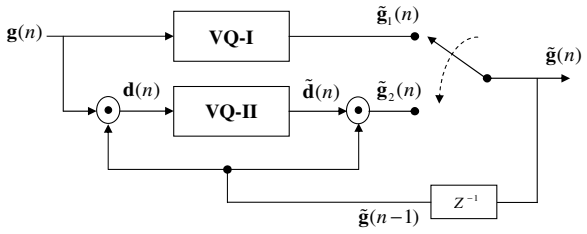

Figure 4: Block diagram of the proposed safety-net gain quantization method.

quantization method, where $\mathbf{g}(n) = [g_p(n),\, g_c(n)]^T$ is the unquantized gain vector consisting of pitch and stochastic codebook gains. VQ-I is a trained memoryless vector quantizer, and VQ-II is a trained memoryless differential quantizer. By using this quantization scheme, we can obtain two quantized gain candidates, $\tilde{\mathbf{g}}_1(n) = [\tilde{g}_{1,p}(n),\, \tilde{g}_{1,c}(n)]^T$ and $\tilde{\mathbf{g}}_2(n) = [\tilde{g}_{2,p}(n),\, \tilde{g}_{2,c}(n)]^T$.

$$\begin{aligned} d_p(n) &= g_p(n) - \tilde{g}_p(n-1), \\ d_c(n) &= g_c(n)/\tilde{g}_c(n-1), \\ \tilde{g}_{2,p}(n) &= \tilde{g}_p(n-1) + \tilde{d}_p(n), \\ \tilde{g}_{2,c}(n) &= \tilde{d}_c(n) \cdot \tilde{g}_c(n-1), \end{aligned} \tag{3}$$

where $\tilde{\mathbf{d}}(n) = [\tilde{d}_p(n),\, \tilde{d}_c(n)]^T$ is the quantized version of $\mathbf{d}(n) = [d_p(n),\, d_c(n)]^T$. By comparing two perceptually weighted error [12] related to $\tilde{\mathbf{g}}_1(n)$ and $\tilde{\mathbf{g}}_2(n)$, we select the better quantized gain, $\tilde{\mathbf{g}}(n)$ that has lower weighted error.

3. Implementation and Performance Evaluation

3.1. Implementation

Based on the results of Fig.1 and 2, we designed the quantizer with the threshold value, d_{thres}, of 2.0dB for spectral parameters. Assuming that a segment is N×10ms long, it needs 4 bits for a number of target LSP vectors, M, N-2 bits for positions of target LSPs, and 24×M bits for quantization of target LSPs. In addition, each segment needs 6 bits for a time-shift caused by the difference between start points as described in Fig.3, and 3 bits for a number of non-predictive frames such as NPF-I and NPF-II.

Table 2 describes a bit allocation for each type of excitation signal modeling. 'Pitch' in the NPF-II frame and in the PF frame are related to the speaker-dependent pitch pulse codebook and the conventional adaptive codebook, respectively, and 'FCB' denotes a fixed codebook for modeling random signals. The proposed algorithm used an ACELP structure for the fixed codebook. Since the speech quality of non-predictive frames mainly affects the overall quality of synthesized speech, we use the ACELP structure with 10 pulses [13] in non-predictive frames and with 4 pulse [12] in the predictive frame. As explained in Section 2.2.2, gain parameters are quantized by using the safety-net method with 1 bit for selection in the predictive frame.

3.2. Performance Evaluation

By using the *Voiceware*[9] Korean database sampled at 8 kHz, we show the performance of the proposed algorithm being applied

Table 2: Bit allocation for excitation modeling methods

Coding parameters	5ms NPF-I	5ms NPF-II	10ms PF 1st 5ms	10ms PF 2nd 5ms
Pitch	·	12	8	5
FCB	35	35	17	17
Gain	10	10	1+7	1+7
Total	45	57	33	30

to the *Voiceware* TTS synthesizer with 49216 speech segments. PESQ[1] [6] and WMOPS[2] are used as objective quality and complexity measurement. For experiments, we used 15 female and 15 male speech samples of 6 to 8sec long each, generated from the TTS system with 30 text files. Table 3 shows PESQ scores of the proposed algorithm with reference to the synthesized speech using original signals.

Table 3: PESQ scores of the proposed algorithm

PESQ scores	Male	Female	Average
Mean	3.414	3.420	3.417
STD.	0.041	0.072	0.056

We measured the computational complexity of decoding procedure in terms of WMOPS, which reflects the complexity weight of 16- and 32-bit arithmetic operations in the fixed-point descriptions. Table 4 shows the complexity of the proposed synthesizer by classifying it with NPF-I, NPF-II, and PF frames. The proposed algorithm mainly needs table look up and synthesis filtering for the decoding process, which is trivial in terms of complexity. The post-processing means a short-term post-filtering procedure widely used in standard speech coders, which enhances spectral formant regions of synthesized speech signals. The result shows that the proposed synthesizer has low computational complexity, considering that ITU-T standard G.729 annex A [14] has complexity of around 1.62 WMOPS for total decoding procedures.

Table 4: Computational complexity of the proposed synthesizer

WMOPS	Frame structure		
	NPF-I type	NPF-II type	PF type
Decoding	0.530	0.587	0.788
Post-processing	0.501	0.501	0.501
Total	1.031	1.088	1.289

[1]The quality of synthesized speech is influenced by an overall performance of TTS systems. To estimate the speech quality caused by only a database compression, however, we measure PESQ scores with reference to the synthesized speech using original signals.

4. Conclusion

This paper proposed a speech coding and synthesis algorithm for TTS synthesizer database based on analysis-by-synthesis paradigm. We proposed the speaker dependent codebook to model pitch-pulse shapes. The rational behind the idea was that pitch-pulse shapes for one speaker might be restricted because of limited pronunciation and physical characteristics. Additionally, to increase coding efficiency, the mixed structure of non-predictive and predictive frame types were employed. From the performance verification tests with TTS Korean databases, we confirmed that the proposed coder had a compression ratio of about 1/13, very low complexity of around 1.2 WMOPS, and random access capability.

5. References

[1] Sami Lemmetty, *Review of speech synthesis technology*, Master's thesis, Helsinki university of technology, 1999.

[2] ITU-T Rec. G.191, "Software tools for speech and audio coding standardization," Nov. 2000.

[3] T. Dutoit, V. Pagel, N. Pierret, F. Bataille and Oliver van der Vrecken, "The MBROLA Project: Towards a set of high quality speech synthesizer free of use for noncommercial purposes," in *Proceedings of ICSLP 1996*, Philadelphia, 1996.

[4] Oliver van der Vrecken, N. Pierret, T. Dutoit, V. Pagel and F. Malfrere, "New techniques for the compression of synthesizer databases," in *1997 IEEE International Symposium on Circuits and Systems*, pp. 2641-2644, June 9-12, 1997, Hong Kong.

[5] Chang-Heon Lee, Sung-Kyo Jung and Hong-Goo Kang, "Applying a speaker-dependent speech compression technique to concatenative TTS synthesizers," in *IEEE Transactions on Speech and Audio Processing*, Accepted and to be published.

[6] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech coders," Feb. 2001.

[7] Ki-Seung Lee, "Temporal decomposition based on a rate-distortion criterion," in *IEEE Signal Processing Letters*, Vol.11, No.1, pp. 33-35, Jan. 2004.

[8] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24bits/frame," in *IEEE Trans. on Speech and Audio Processing*, Vol.1, Issue 1, pp. 3-14, Jan., 1993.

[9] http://www.voiceware.co.kr

[10] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.

[11] W. B. Kelijin, *Speech Coding and Synthesis*, Elsevier Science B. V., 1995.

[12] ITU-T Rec. G.729, "Coding of Speech at 8kbit/s Using CS-ACELP," 1996.

[13] ITU-T Rec. G.729, "Coding of Speech at 8kbit/s Using CS-ACELP Annex D and E," 1998.

[14] R. Salami, C. Laflamme, B. Bessette and J-P. Adoul, "ITU-T G.729 annex A : Reduced complexity 8kbit/s CS-ACELP codec for digital simultaneous voice and data," *IEEE Communication Magazine*, pp. 53-63, 1997.