A CASCADED ALGEBRAIC CODEBOOK STRUCTURE TO IMPOVE THE PERFORMANCE OF SPEECH CODER

Sung-Kyo Jung, Kyoung-Tae Kim, Hong-Goo Kang and Dae-Hee Youn

MCSP Lab., School of Electrical & Electronic Eng.,
Yonsei University, Seoul, KOREA

{skjung,kktae,hgkang}@mcsp.yonsei.ac.kr, dhyoun@yonsei.ac.kr

Abstract

This paper presents a cascade structure of an algebraic codebook to improve the performance of low bit-rate speech coder. A codeword of an algebraic codebook consists of a set of pulse amplitudes and positions. In general, the amplitude of each pulse is constrained to be either +1 or –1 due to the limitations of bit-rate and complexity. Thus, the performance of the codebook is varied depending on the characteristic of input target vectors. In this paper, we extend the algebraic codebook structure to two stages in order to provide flexible pulse combinations. While all pulses, *M*, are simultaneously selected in a classical one-stage algebraic codebook, the cascade structure searches the pulses with a two step procedure, i.e., *L* pulses at the first stage and (*M-L*) pulses at the second stage. Experiments confirm that our algorithm provides higher quality than the conventional scheme when the total number of pulses is same. In case of assigning 24 pulses per 8-ms subframe, a segmental SNR between target and synthesized signal increases 1.04 dB. In addition, at the same environment, the complexity of fixed codebook search is reduced by about 32 %.

1. INTRODUCTION

The analysis-by-synthesis principle leads to provide good synthesized quality in low bit-rate speech coding. The multipulse excitation (MPE) coding [1], code-excited linear prediction (CELP) coding [2] are the key technologies for the technique. However, the closed-loop procedure to decide optimal excitation code word requires a lot of computational complexity. The algebraic CELP [3] algorithm has been widely used in many speech coding standards, due to its low complexity and high quality in analysis-by-synthesis processing. In general, the algebraic codebook consists of a set of pulse amplitude and position combinations. The zero-amplitude pulses and non-zero amplitude pulses are assigned to respective positions of the combination. Many researches have been focused on improving the performance of the algebraic CELP in terms of quality [4][5] as well as complexity. A joint optimization procedure of the amplitudes and positions was proposed for searching algebraic multi-pulse codebooks in [4]. In [5], an algebraic code word with multiple magnitudes was used to improve the performance of ACELP. The performance of the schemes might be different depending on the characteristics of the speech segments because the structures of the code words are restricted to the number of pulses and their amplitudes.

In this paper, we propose an algebraic codebook structure that is flexible on magnitudes while maintaining low bit-rate and low complexity. By extending the algebraic codebook to two-stage structure, we approximate the target signal of the fixed codebook. Thus we will have two algebraic codebooks that have two different overall gains. Then, we compare the performance of the cascaded algebraic codebook to conventional codebook by assigning the same number of pulses. While *M* pulses are simultaneously selected in a classical algebraic codebook, the cascaded structure finds the optimal combination of the code vectors by separating it into *L* and (*M-L*) pulses, respectively. The proposed approach is more efficient in terms of flexibility. If the variance of the target signal is low such as unvoiced, the pulse position of the first and the second stage might be different and consequently its behavior is pretty similar to the single stage approach. If the variance of the target signal is high such as onset, the proposed method is good for approximating the second dominant peak, which could not be modeled well by the single-stage approach. For the cascaded codebook, the gain information of each stage should be transmitted to the decoder. It increases the overall bit-rates, however, since the two gains of the cascaded codebook are highly correlated, the gain of the second stage can be efficiently quantized with a small number of bits. Simulation results confirm that the cascaded algebraic codebook approach shows higher performance than the single-stage codebook scheme.

2. CONVENTIONAL ALGEBRAIC CODEBOOK

The objective of the algebraic codebook is to find the positions and amplitudes of the multiple pulses in order to minimize the difference between target and synthesized signal. To avoid the huge computation due to a convolution operation, the algebraic codebook uses an interleaved single-pulse permutation (ISPP) structure. In addition, each code vector contains several nonzero pulses with amplitudes of +1 or –1 to be selected from different groups of pulse locations [3].

A general criterion to determine the optimal code word is minimizing a perceptually weighted error between the target signal $x_w(n)$ and the synthesized signal $\hat{x}_w(n)$ as follows:

$$\begin{aligned} E_\xi &= \sum_{n=0}^{N-1}\left(x_w(n)-\hat{x}_w(n)\right)^2 \\ &= \sum_{n=0}^{N-1}\left(x_w(n)-g_c\sum_{k=1}^{M}s_k h_w(n-m_k)\right)^2 \end{aligned} \tag{1}$$

This work was supported in part by Electronics and Telecommunications Research Institute (ETRI).

where N is the subframe size and M is the number of the pulses to be selected. $h_w(n)$ is the impulse response of the weighted synthesis filter, and g_c is the gain of the algebraic codeword. s_k and m_k are the sign and position of k-th pulse, respectively. As we see from the equation (1), each pulse to the total residual energy is equivalent across the subframe. Thus, the quality degradation may be often observed in transient segments or in high-pitched speech segments where long-term prediction could not flatten the characteristics of the residual signal. In those regions, multiple peaks would be simultaneously existed in the subframe and their amplitude variations would be high.

3. CASCADED ALGEBRAIC CODEBOOK

In a two-stage algebraic codebook approach, the search procedure of M pulses is composed of two steps, a search procedure of L pulses at the first stage and that of (M-L) pulses at the second stage. The target vector of the first-stage codebook, $x_1(n)$, consists of the weighted speech after subtracting the zero-input response of the weighted synthesis filter and pitch contribution. The target signal for the next-stage codebook, $x_2(n)$, is obtained by subtracting the contribution of the first-stage codebook from the target vector for the first stage. Figure 1 shows the block diagram of the cascaded algebraic codebook.

The cascaded algebraic codebook can be represented as a weighted sum of the code vector from each stage. Thus, the synthesized signal, $\hat{x}_w(n)$, becomes:

$$\begin{aligned}\hat{x}_w(n) = g_1 \sum_{k=1}^{L} s_{1,k}\ h_w(n - m_{1,k}) \\ + g_2 \sum_{k=1}^{M-L} s_{2,k}\ h_w(n - m_{2,k})\end{aligned} \tag{2}$$

where g_1 and g_2 are the codebook gains of each stage. $s_{1,k}$ and $m_{1,k}$ are the sign and position of k-th pulse in the first stage, $s_{2,k}$ and $m_{2,k}$ are the sign and position of k-th pulse in the second stage.

If we cascade the search procedure to M stage and only one optimal pulse is searched in each stage, it becomes a multi-pulse coding scheme, i.e.,

$$\hat{x}_w(n) = \sum_{k=1}^{M} g_k\ h_w(n - p_k) \tag{3}$$

where p_k represents k-th position and g_k represents the amplitude of the k-th pulse [1].

This analysis tells us that the proposed method is a kind of midway process between algebraic and multi-pulse codebook structure. Let's consider the pros and cons of the algebraic and multi-pulse codebook structures. The performance of the multi-pulse structure would be better than that of the algebraic codebook in high bit-rates because of its flexibility on choosing the amplitudes of the pulses. For low bit-rates, the algebraic codebook would be a better choice. In unvoiced regions where the amplitudes of the targets are somewhat constant within a processing frame, the algebraic codebook structure would be a better choice. However, in onset regions where the dynamic range of the targets is relatively high, the multi-pulse scheme might be the better choice. From the observation, we knew the quality of the proposed method should be varied depending on the characteristics of the processing frame, the number of total pulses, and the number of processing stages. The complexity of the proposed algorithm should be also varied depending on the number of total pulses and that of the processing stages. Overall, we may conclude that the proposed multi-stage approach is more flexible than the single-stage approach.

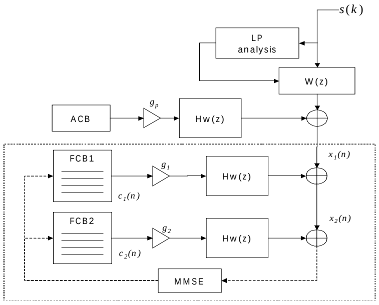

Figure 1: *Block diagram of cascaded algebraic codebook.*

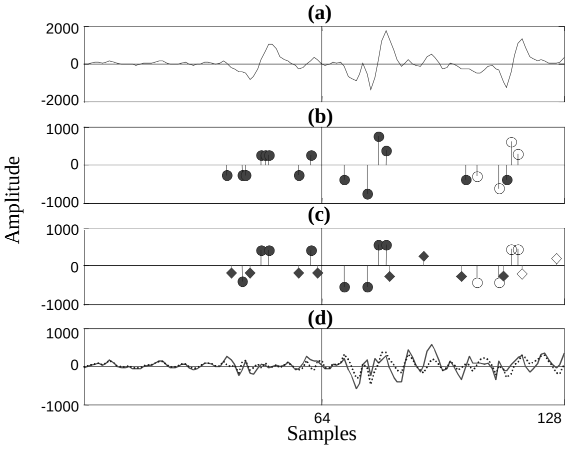

Figure 2: *Example of the code vectors obtained from two codebook approaches. The number of pulses to be searched each subframe is 8 and subframe size is set to 64 samples. (a) Target signal for fixed codebook. (b) Code vector of the classical codebook. ('●'-marked pulses: the code vectors including the gain term, '○'-marked pulses : the pulses obtained by pitch sharpening process). (c) Code vector of the cascaded codebook. ('●'-marked pulses: the code vectors of the first-stage codebook, '◆'-marked pulses: the code vectors of the second-stage codebook, '○'-marked and '◇'-marked pulses: the pulses obtained by pitch sharpening process). (d) Error signals of two codebook approaches in weighted domain. (solid line: error signal of the classical codebook, dotted line: error signal of the cascaded codebook).*

Figure 2 shows an example of two code vectors obtained from the different codebook approaches. To approximate the target sequence as shown in Figure 2 (a), the different excitation is used depending on the codebook approach. The amplitudes of the cascaded codebook are more flexible than those of the classical approach. The error signal is lower than the single-stage approach. In case of the cascaded codebook structure, the gain information of each stage should be transmitted to the decoder. However, we may not transmit the gain for the second stage if we consider the correlation between the first and the second-stage gain. Figure 3 shows the distribution of the ratio of second stage gain to first stage gain. It clearly shows that two gains are highly correlated. Thus, instead of quantizing the gain of the second stage, we can quantize the ratio of first and second stage gain with a small number of bits.

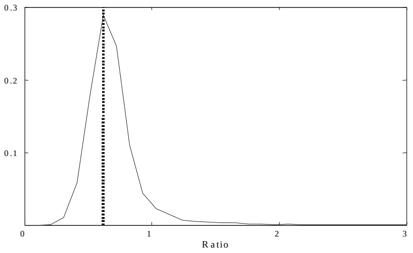

Figure 3: *Histogram of the ratio between two codebook gains. The highest probability of the ratio is 0.6566.*

4. SIMULATION RESULTS

In this paper, we only consider two-stage structure to verify the performance of the proposed algorithm. To flexibly perform the performance comparison between the cascaded codebook and the classical method, we modified the frame structure of the ITU-T G.729 [7]. We set the frame size and the subframe size to 16 ms and 8 ms, respectively, because it results the number of samples to power of two. In each frame, we extract the linear predictive coefficients from the autocorrelation method with a 30 ms asymmetric window as used in the G.729 coder. Then, a pitch lag is obtained once per frame using an open-loop pitch analysis in weighted speech domain. The search procedure to obtain the optimum adaptive codebook index is identical to that of the G.729 coder. In order to approximate the non-periodic component of excitation signal, we redesigned the structure of the algebraic codebook considering both classical and cascaded codebook scheme. The fixed codebook structure is designed based on ISPP method as shown in Table 1. The 64 positions in the code vector are divided into 4 tracks of interleaved positions, with 16 positions in each track. Depending on the number of total pulses, the different code vectors are constructed by placing a certain number of signed pulses in each track. (from 1 to 6 pulses per track). Table 2 shows the pulse combinations of classical codebook and cascaded codebook that we used for the simulation. When total 4 pulses are assigned to the cascaded codebook, the first and third tracks, T_0 and T_2, are used for searching the first pulse and the second and fourth tracks for the other pulse. When the four or more pulses are searched in each stage, we divided the multiple pulses in groups of 4 and selected the best-optimized pulse combination using depth-first tree search strategy [7].

Table 1: *ACELP excitation codebook.*

Track	Sign	Positions
T_0	±1	0,4,8,12,16,20,24,28,32,36,40,44,48,52,56,60
T_1	±1	1,5,9,13,17,21,25,29,33,37,41,45,49,53,57,61
T_2	±1	2,6,10,14,18,22,26,30,34,38,42,46,50,54,58,62
T_3	±1	3,7,11,15,19,23,27,31,35,39,43,47,51,55,59,63

Table 2: *Pulse combinations of two codebook approaches.*

Number of total pulses	Classical codebook	Cascaded codebook	
		1^{st}-stage	2^{nd}-stage
M = 4	4	2	2
M = 8	8	4	4
M = 12	12	8	4
M = 16	16	8	8
M = 20	20	12	8
M = 24	24	12	12

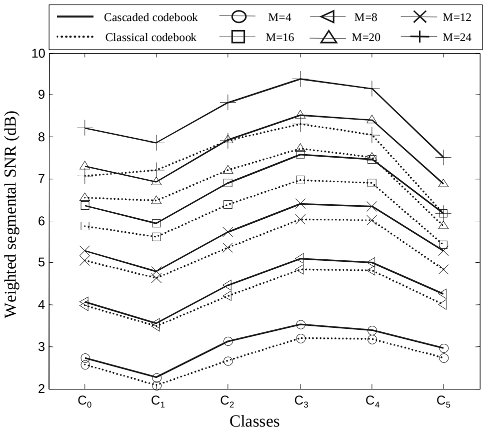

Figure 4: *Weighted segSNR results of the classical codebook and the cascaded codebook depending on the characteristics of speech segments.*

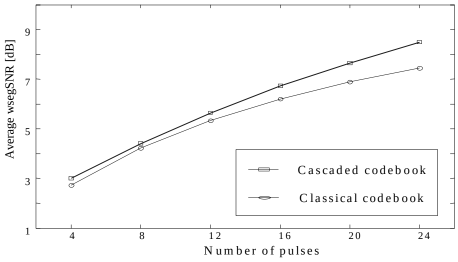

Figure 5: *Average wsegSNR according to the number of pulses assigned in each track.*

4.1. Objective measures

In order to compare the performance of two codebook approaches, we used a weighted segmental signal-to-noise ratio (wsegSNR) and ITU-T P.862 perceptual evaluation of speech quality (PESQ) [8].

To only evaluate the performances of codebook schemes, the wsegSNR was defined between the target signal and the contribution of algebraic codebook. In order to investigate the wsegSNR values depending on the characteristics of speech segments, each frame was classified into six categories such as silence (C_0), noise-like unvoiced (C_1), unvoiced (C_2), onset (C_3), non-stationary voiced (C_4), and stationary voiced (C_5). We employed the frame classification module used in the selectable mode vocoder (SMV) [6]. On the other hand, the PESQ scores were measured between input speech signal and output speech of the coders. The test materials were obtained from NTT multi-lingual database and included 16 Korean sentences pronounced by four female and four male speakers. Figure 4 shows the wsegSNR of the classical codebook and the cascaded codebook. The performances of both codebook schemes are improved as more pulses are assigned. In case of the cascaded codebook, we could see more improvement of wsegSNR in perceptually important segments such as onset region, non-stationary voiced region, and stationary voiced region. Figure 5 shows the average wsegSNR by varying the number of pulses assigned per track. When total 24 pulses are allocated to each subframe, an improvement of around 1.04 dB is achieved by the cascaded codebook over the conventional codebook. When the more pulses are assigned, the performance of the cascaded codebook is much better than that of the classical codebook. For the high bit-rates, the cascaded codebook is more efficient structure to model the various speech segments.

Table 3 describes the results of PESQ measurement. Comparing to the results of the classical codebook, the cascaded codebook approach shows higher quality when the total number of pulse are same. In particular, the performance improvement of the cascaded codebook in female voices is more significant. Since the pitch interval of the female voices is usually shorter than that of the male voices, female voices have higher probability that there exist two or more peaks in each processing frame. If the dynamic range of the peaks is high, the classical algebraic codebook does not model them correctly because it has only one overall gain. However, the cascaded structure has more flexibility in accurately modeling the region.

Table 3: *PESQ results of two codebook approaches.*

No. of pulses	Classical codebook			Cascaded codebook		
	Female	Male	Avg.	Female	Male	Avg.
M=4	3.454	3.706	3.580	3.505	3.728	3.616
M=8	3.669	3.886	3.777	3.707	3.895	3.801
M=12	3.812	3.984	3.898	3.843	3.984	3.914
M=16	3.860	4.026	3.944	3.932	4.044	3.988
M=20	3.890	4.052	3.971	3.976	4.095	4.035
M=24	3.915	4.073	3.994	4.008	4.124	4.066

Table 4: *Computation comparison in terms of addition and multiplication. K (=16) is the number of candidates per track and p (from 2 to 6) is the number of pulses to be searched each track.*

Types	Operations	
	Additions	Multiplications
SS	$A_{SS}^{p} = \sum_{t=1}^{4} (2p^2 + (t-1)p)K^t$	$M_{TS}^{p} + (p-1)K$
TS	$2 \times A_{SS}^{p-1}$ or $A_{SS}^{p-1} + A_{SS}^{p-2}$	$M_{TS}^{p} = p(K^2 + K^3 + 4K^4)$

4.2. Complexity check

In order to estimate the reduction amount of computational complexity, we measured the numbers of additions and multiplications only used during the codebook search procedure.

Table 4 shows the numbers of multiplications and additions required in the pulse-position search routine. The total number of iterations to find the possible pulse combinations is same in single-stage (*SS*) and two-stage (*TS*) approaches. However, the operations to search the possible would be different. Figure 6 shows the reduction ratio of the computations obtained using the equations in Table 4. We compared the ratios by varying the number of pulses to be searched in each track. The range of the reduction ratio is varied from 18 % to 32 % depending on the number of pulses.

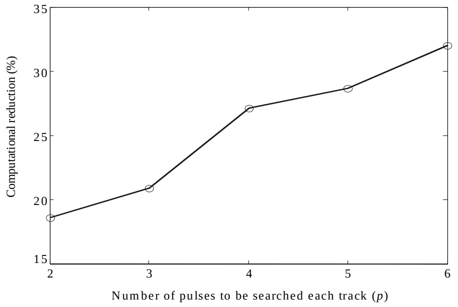

Figure 6: *Computational reduction of the proposed approach.*

5. CONCLUSIONS

In this paper, we propose a cascaded algebraic codebook structure in order to improve the quality of the decoded speech. While the classical algebraic codebook simultaneously searches the total of *M* pulses, the cascaded two-stage codebook searches the pulses in a sequential manner, i.e., *L* and (*M-L*) pulses are selected at the first and second stage, respectively. Experiments confirm that our algorithm shows higher quality than the conventional one when the same number of pulses is assigned. Especially, the performance improvement was significant in transient, non-stationary voiced regions, and in high-pitched voice segments.

REFERENCES

[1] B.S. Atal, and J.R. Remde "A new model of LPC excitation for producing natural sounding speech at low bit rates," in *Proc. Int. Conf. Acoust. Speech Sign, Process.*, pp. 614-617, 1982.

[2] B.S. Atal and M.R. Schroeder, "Stochastic coding of speech at very low bit rates," in *Proc. Int. Conf. Comm.*, pp. 1610-1613, 1984.

[3] J.-P. Adoul, P. Mabilleau, M Delprat, and S. Morisette, "Fast CELP coding based on algebraic codes," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, pp. 1957-1960, 1987.

[4] M.A. Ramírez and Max Gerken, "A multistage search of algebraic CELP codebooks," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, pp. 17-20, 1999.

[5] O. Halmi, H. Tolba, D. Guerchi, and D. O'Shaughnessy, "On improving the performance of analysis-by-synthesis coding using a multi-magnitude algebraic code-book excitation signal," in *Proc. Int. Conf. Spoken Language Process.*, pp. 1857-1860, 2002.

[6] 3GPP2, "*Selectable mode vocoder service option for wideband spread spectrum communication systems*," Dec. 2001.

[7] Richard V. Cox, "Three new speech coders from the ITU cover a range of applications," *IEEE Communications Magazine*, pp.40-47, Sep. 1997.

[8] 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.