A Novel Rate Selection Algorithm for Transcoding CELP-type Codec and SMV

Dalwon Jang, Seongho Seo, Sunil Lee, and Chang D. Yoo

Department of Electrical Engineering and Computer Science
Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
mcjang@eeinfo.kaist.ac.kr

Abstract

In this paper, we propose an efficient rate selection algorithm that can be used to transcode speech encoded by any code excited linear prediction (CELP)-type codec into a format compatible with selectable mode vocoder (SMV) via direct parameter transformation. The proposed algorithm performs rate selection using the CELP parameters. Simulation results show that while maintaining similar overall bit-rate compared to the rate selection algorithm of SMV, the proposed algorithm requires less computational load than that of SMV and does not degrade the quality of the transcoded speech.

1. Introduction

Today, various communication networks are being developed. Each communication network uses different speech coding standard based on different requirements. To communicate using two different speech codecs, the decoder of one codec and encoder of the other should be placed in tandem. This method is called the tandem transcoding algorithm. Although simple, the tandem transcoding algorithm has several problems—degradation in speech quality, high computational complexity and long delay. To address these problems, various transcoding algorithms that directly transform speech parameters have been proposed [1][2][3].

Selectable Mode Vocoder (SMV)[4] was selected by the Telecommnication Industry Association (TIA) and the 3th Generation Partnership Project 2 (3GPP2) as a new speech coding standard for code division multiple access (CDMA) application. For efficient use of bandwidth, SMV determines the transmission rate of each frame based on the network-controlled operating mode and the attribute of the frame. The performance of rate selection algorithm has a great effect on the over all performance of SMV codec.

The proposed algorithm is only applicable to a CELP-type codec since the algorithm uses CELP parameters as inputs. CELP-type codec transmits 5 different kinds of parameters, which are converted to bit stream: line spectral frequency (LSF), fixed codevector (FCB), fixed code-book gain (FCB gain), adaptive code-book, which is represented by pitch delay, and adaptive code-book gain (ACB gain). These parameters are used in the proposed rate selection algorithm.

To use the rate selection algorithm provided by the SMV in transcoding the parameters from a CELP-type codec to the SMV leads to inefficiency in terms of computation and delay. The transcoder should maximally utilize the information carried by the CELP parameters and avoid duplicating any encoding procedures of the CELP-type codec. Any information obtained by the LP analysis and open-loop pitch detection procedures is carried by the LSP and adaptive codebook parameters, thus the two procedures are unnecessary in the transcoder[2][3]. However, the two have an important role in the rate selection algorithm of SMV. Thus a rate selection algorithm that uses CELP parameters and does not require LP analysis and open-loop pitch detection is required.

This paper is organized as follows. In Section 2, operations of SMV, especially its rate selection algorithm are briefly described. In Section 3, proposed new rate selection algorithm is explained in detail. Section 4 provides the simulation results on the performance of the proposed rate selection algorithm. G.723.1[5] and G.729A[6] are selected as source speech codecs and used in simulation. Finally, Section 5 concludes the paper.

2. SMV Speech Coding Standard

The SMV is based on the extended-CELP (eX-CELP)[7] algorithm and takes its input as speech signal sampled at the rate of 8kHz. Since the frame length of SMV is 20ms which corresponds to 160samples and the length of the look-ahead is 10ms, total algorithmic delay of SMV is 30ms.

The SMV is based on four codecs (encoder/decoder) operating at the rates of 8.55kbps, 4.0kbps, 2.0kbps and 0.8kbps. These codecs are called Rate 1(full-rate), Rate 1/2(half-rate), Rate 1/4(quarter-rate), and Rate 1/8(eighth-rate), respectively. SMV also selects a frame type as either type-0 or type-1 for each Rate 1 and Rate 1/2 frame. Input speech frame is determined as type-1 when it contains stationary voiced speech, otherwise input frame is declared as type-0. The encoding and decoding method of type-0 frame is different from those of type-1. Thus, for satisfactory performance, appropriate selections of encoding rate and frame type are essential.

The SMV has 4 network-controlled operating modes: Mode 0 (premium mode), Mode 1 (standard mode), Mode 2 (economy mode), and Mode 3 (capacity-saving mode). The different modes allow a tradeoff between average bit rate (ABR) and speech quality. Prior to the rate selection, input speech frame is classified into one of 6 categories - silence, noise-like, unvoiced, onset, non-stationary voiced, and stationary voiced. Tables 1 and 2 show the rates allowed for various modes and frame classes. The network determines the operating mode, and the SMV is responsible for determining the frame class and type. Based on the mode and frame class, the most appropriate rate is determined.

The rate selection of SMV is performed in two steps: cal-

Table 1: *SMV class-rate map(Mode 0)*

Frame class	Rate 1/8	Rate 1/4	Rate 1/2	Rate 1
Silence	√		√	
Noise-like			√	√
Unvoiced			√	√
Onset				√
Non-stat. voiced				√
Stat. voiced				√

Table 2: *SMV class-rate map(Mode 1,2,3)*

Frame class	Rate 1/8	Rate 1/4	Rate 1/2	Rate 1
Silence	√			
Noise-like		√	√	
Unvoiced		√	√	
Onset		√	√	√
Non-stat. voiced			√	√
Stat. voiced			√	√

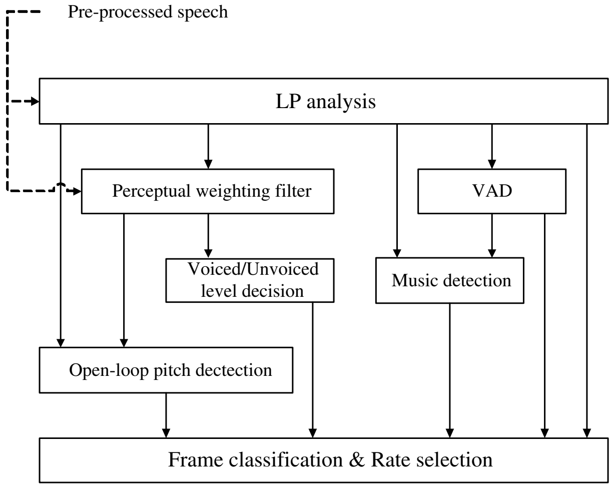
Pre-processed speech
LP analysis
Perceptual weighting filter
VAD
Voiced/Unvoiced level decision
Music detection
Open-loop pitch dectection
Frame classification & Rate selection

Figure 1: *Block diagram of rate selection algorithm of SMV*

culate speech parameters and then compares these to a fixed threshold to determine the frame class and rate. Figure 1 shows a simplified block diagram of the rate selection algorithm of SMV.

Among the blocks shown in Figure 1, linear prediction (LP) analysis block and the open-loop pitch detection block play an integral part of the encoder. The LP analysis block not only outputs line spectral frequencies, which is its main function, but as a by-product generates autocorrelation function, prediction error, reflection coefficient and few other parameters. These by-products are utilized by the rate selection process. The open-loop pitch detection block estimates the open-loop pitch delay, and in the process generates the maximum values of the autocorrelation of the excitation. The estimated pitch delay and the maximum values are used extensively in determining whether a frame is voiced or not. Overall, the two blocks mentioned play an indispensable role in the rate selection algorithm of SMV.

The voice activity detector (VAD) decides whether the input speech frame is silence or not. An input frame absent of any voice activities is classified as silence. The music detection determines whether the input frame is music or not. An input frame that is classified as music is always encoded at Rate 1. The voiced/unvoiced level decision computes the degree of voice in the frame and other useful parameters. The perceptual weighted filter generates weighted speech signal used in computing the pitch delay.

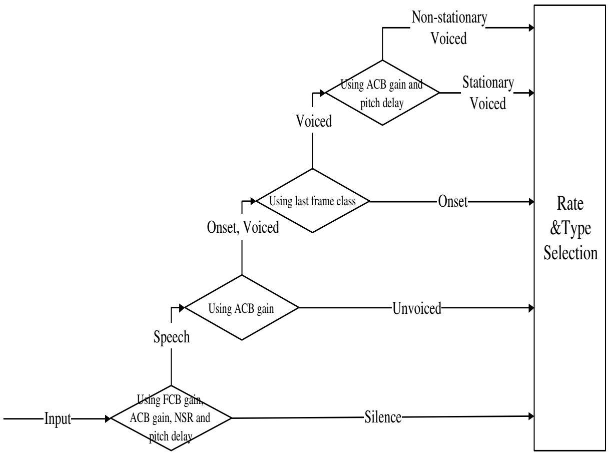

Figure 2: *Block diagram of frame classification*

3. New Rate Selection Algorithm

The goal of the proposed rate selection algorithm is to maintain ABR and speech quality at the level of that of the tandem with a lower computational complexity.

3.1. Frame Classification

For a given mode, the frame class determines the frame rate. For this reason, frame classification plays an essential role in rate selection. The SMV classifies each input frame into one of 6 classes. In the proposed algorithm, the number of frame classes is reduced to 5 classes to simplify the algorithm: silence, unvoiced, onset, non-stationary voiced and stationary voiced. Noise-like and unvoiced are combined into unvoiced. Figure 2 shows a simplified block diagram of the frame classification process.

First, an input frame is classified into either silence or speech. For this, ACB gain, FCB gain, noise-to-signal ratio (NSR) and the gradient of the pitch delay are used. An input frame declared as speech is further classified into either unvoiced or voiced speech using ACB gain. If the current frame is classified as voiced speech and the previous frame was unvoiced speech, the current frame is adjusted to the onset frame automatically. After that, voiced speech is further classified into either the stationary voiced or non-stationary voiced using ACB gain and the gradient of the pitch delay.

ACB gain can be used as an indicator of voice activity in speech. That is, ACB gain is large in the presence of speech and small in silence regions. Thus, it can be used to classify between silence and speech and between unvoiced and voiced speech. However, ACB gain cannot be used as it is since it varies too rapidly within a frame. To alleviate this problem, the minimum of ACB gains for each frame is smoothed. Then, the smoothed ACB gain is compared to the threshold value to determine whether the frame is speech or silence. In Figure 3,

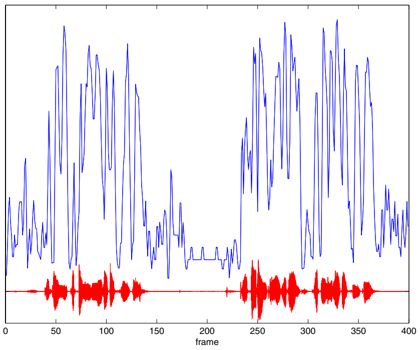

Figure 3: *ACB gain of G.729A and speech*

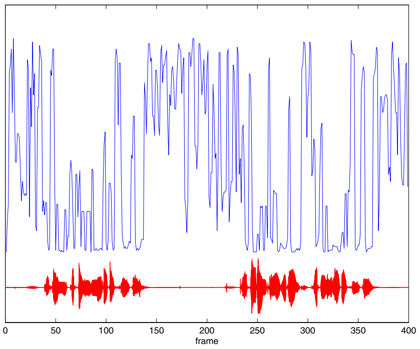

Figure 4: *Gradient of pitch delay and speech*

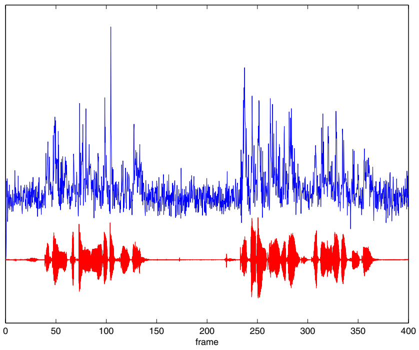

Figure 5: *FCB gain of G.729A for noisy speech*

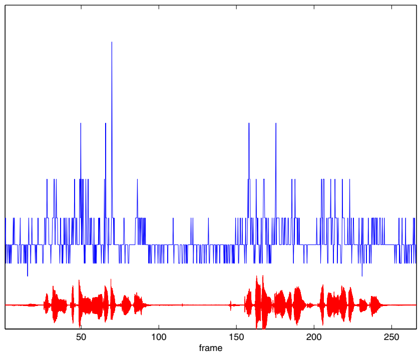

Figure 6: *FCB gain of G.723.1 for noisy speech*

smoothed ACB gains of G.729A speech codec is shown with corresponding speech waveform. For the classification of non-stationary voiced and stationary voiced, all ACB gains are used. If all ACB gains in a frame are bigger than a certain threshold value, the frame is classified into stationary voiced.

The gradient of the pitch delay is also used in the classification. The difference between the maximum and minimum pitch delays in the current frame and the previous frame are calculated and averaged. The value of the difference is large in silence regions and small where speech is present: pitch delay varies slowly in regions where speech is present. Figure 4 shows this tendency. Using this property, we can classify the input frame into either speech or silence. If the current frame is a stationary voiced frame, the pitch delays in the current frame and the previous frames are very similar to one another. So, the variance of pitch delays can be used to divide the stationary voiced and non-stationary voiced.

As the ACB gain, FCB gain also shows similar tendency in the presence of speech. Thus it can be used to classify signal into either silence or speech. However, if the input speech is noisy, FCB gain is not an useful parameter. Figure 5 and Figure 6 show that FCB gain cannot be used to determine whether the input frame is silence or speech when noise is present. The FCB gains of two figures are calculated by G.729A and G.723.1 in the presence of white Gaussian noise. Because the magnitude of FCB gains in speech is similar to that in only noise, it is difficult to classify using FCB gain. Extreme smoothing and adaptive threshold-setting can make FCB gain useful, but the result is not acceptable, thus we conclude FCB gain is useful only in low noise condition. To know noise level, NSR is used. FCB gain is used only in case of small NSR. In addition, NSR is also used by itself. If NSR is very large, the frame is classified to silence.

3.2. Rate Selection Based on Frame Classification

After the input frame is classified, the encoding rate is selected. The process is similar to that of SMV. Speech parameter that do not require LP analysis and open-loop pitch analysis are computed and utilized.

4. Simulation Results

The performance of the transcoder using the proposed rate selection algorithm is compared with that of the tandem transcoder. First, the rate selected by the proposed algorithm is compared to that selected by the rate selection algorithm of SMV. Second, the computational complexities of both cases are evaluated and compared based on weighted million operations per second (WMOPS). Finally the objective quality of the transcoded speech is evaluated using the perceptual evaluation of speech quality (PESQ) [8] scores.

4.1. Rate Accuracy and Average Bit Rate

In order to determine the performance of the proposed rate selection algorithm, rate accuracy and ABR are employed. Rate accuracy is defined as the ratio of total number of frames to the number of frames exactly matched in terms of rate. Table 3 and Table 4 show that the accuracy is not very high, but the ABR of the proposed algorithm is similar to that of SMV rate selection algorithm.

Table 3: *Rate accuracy and ABR for G.729A→SMV transcoder*

Input	Mode	Rate accuracy(%)	ABR(bps)	
			Tandem	Proposed
Male	0	70.84	5830.50	5082.75
	1	68.19	4413.38	4086.75
	2	70.97	3410.63	3374.25
	3	77.06	3135.38	3000.75
Female	0	74.86	6231.88	5627.08
	1	68.74	4199.63	4390.89
	2	75.75	3470.03	3467.82
	3	77.35	3351.88	3293.54

Table 4: *Rate accuracy and ABR for G.723.1→SMV transcoder*

Input	Mode	Rate accuracy(%)	ABR(bps)	
			Tandem	Proposed
Male	0	84.17	6349.75	6447.74
	1	65.83	4478.90	5609.55
	2	82.16	3477.89	3517.09
	3	82.92	3334.68	3318.09
Female	0	80.91	6446.24	7028.14
	1	53.64	4843.72	5656.28
	2	75.46	3651.26	3720.61
	3	71.26	3999.50	3513.08

4.2. Complexity and Speech Quality

The performance of the transcoder based on direct conversion of LSF, without open-loop pitch analysis and the proposed rate selection algorithm are revealed in Table 5,6,7 and 8. Table 5 and Table 6 show that the computational complexity of the transcoder using the proposed rate selection algorithm is much lower than that of the tandem transcoder. Table 7 and Table 8 show that the speech quality is maintained.

5. Conclusion

A novel rate selection algorithm for transcoding between CELP-type codec and SMV is proposed. Although the rate accuracy is only 75%, the ABR and speech quality of transcoded speech using the proposed algorithm are comparable to those of the tandem transcoder while computational complexity of the proposed is considerably less than that of the tandem. The proposed algorithm can be used in the transcoding of not only G.729A and SMV or G.723.1 and SMV but of any CELP-type codec and SMV.

Table 5: *Comparison of computational complexity with WMOPS (G.729A→SMV)*

Mode	WMOPS(Male speech)		WMOPS(Female speech)	
	Tandem	Proposed	Tandem	Proposed
0	27.69	23.19	29.24	24.29
1	25.90	20.26	27.06	21.40
2	25.55	20.02	26.80	21.16
3	25.51	20.00	26.82	21.16

Table 6: *Comparison of computational complexity with WMOPS(G.723.1→SMV)*

Mode	WMOPS(Male speech)		WMOPS(Female speech)	
	Tandem	Proposed	Tandem	Proposed
0	29.46	23.30	30.22	24.97
1	27.69	21.90	29.47	22.81
2	27.44	21.53	28.47	22.65
3	27.52	21.52	27.52	22.67

Table 7: *Comparison of PESQ scores(G.729A→SMV)*

Mode	PESQ(Male speech)		PESQ(Female speech)	
	Tandem	Proposed	Tandem	Proposed
0	3.498	3.498	3.233	3.227
1	3.445	3.395	3.178	3.112
2	3.375	3.343	3.136	3.057
3	3.368	3.342	3.123	3.049

Table 8: *Comparison of PESQ scores(G.723.1→SMV)*

Mode	PESQ(Male speech)		PESQ(Female speech)	
	Tandem	Proposed	Tandem	Proposed
0	3.303	3.244	3.081	3.102
1	3.204	3.166	2.959	3.069
2	3.112	3.050	2.904	2.863
3	3.095	2.982	2.957	2.817

6. References

[1] Hong-Goo Kang, Hong-Kook Kim, R.V. Cox, "Improving transcoding capability of speech coders in clean and frame erasured channel environments," *Proc. IEEE Workshop on Speech Coding, 2000*, pp. 78–80, Jan., 2000.

[2] Sung Wan Yoon, Sung Kyo Jung, Young Cheol Park, and Dae Hee Youn, "An efficient transcoding algorithm for G.723.1 and G.729A speech coders", *Proc. Eurospeech 2001*, vol. 4, pp. 2499-2502, 2001.

[3] Sunil Lee, Seongho Seo, Dalwon Jang, Chang D. Yoo, "A novel transcoding algorithm for AMR and EVRC speech coders via direct parameter transformation," In. *Proc. ICASSP 2003*.

[4] 3GPP2 Spec. "Selectable Mode Vocoder Service Option for Wideband Spread Spectrum Communication Systems", 3GPP2-C.S0030-0 v2.0, Dec. 2001.

[5] ITU-T Rec. G.723.1, "Dual-rate Speech Coder For Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s", 1996.

[6] ITU-T Rec. G.729 Annex A, "Reduced Complexity 8 kbit/s CS-ACELP Speech Codec", 1996.

[7] Yang Gao, A. Benyassine, J. Thyssen, Huan-yu Su, E. Shlomot, "EX-CELP : A Speech Coding Paradig", *Proc. ICASSP 2001*, vol. 2, pp. 689-692, 2001.

[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 codecs," 2000.