Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 43 (2004) 17–31
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

An efficient transcoding algorithm for G.723.1 and G.729A speech coders: interoperability between mobile and IP network ☆

Sung-Wan Yoon *, Hong-Goo Kang, Young-Cheol Park, Dae-Hee Youn

MCSP LAB, Department of Electrical & Electronic Engineering, Yonsei University, 134 Shinchon-dong, Sudaemun-gu, Seoul 120-749, South Korea

Received 16 December 2003

Abstract

In this paper, an efficient transcoding algorithm for G.723.1 and G.729A speech coders is proposed. Transcoding in this paper is completed through four processing steps: LSP conversion, pitch interval conversion, fast adaptive-codebook search, and fast fixed-codebook search. For maintaining minimum distortion, sensitive parameters to quality such as adaptive- and fixed-codebooks are re-estimated from synthesized target signals. To reduce overall complexity, other parameters are directly converted in parametric levels without running through the decoding process. Objective and subjective preference tests verify that the proposed transcoding algorithm has comparable quality to the classical encoder–decoder tandem approach. To compare the complexity of the algorithms, we implement them on the TI TMS320C6201 DSP chip. As a result, the proposed algorithm achieves 26–38% reduction of the overall complexity with a shorter processing delay.

Keywords: Speech coding; Transcoding; Tandem; G.723.1; G.729; G.729A; CELP; LSP conversion; Pitch conversion; Fast codebook search

1. Introduction

For the last two decades, many new speech coding standards have been established. Among them, ITU-T G.723.1 (ITU-T Rec. G.723.1, 1996) and G.729 (ITU-T Rec. G.729, 1996a) cover a wide range of applications requiring low bit rates. Both standards have different applications due to their distinctive features, but they obviously tend to share common applications such as voice messaging, and voice over internet protocol (VoIP). In such applications, the interoperability between the networks is crucial, so that endpoint devices are required to support both standard coders. However, supporting multiple coders in a single device costs system complexity and power, which will become a significant issue when the device needs to be portable and is powered by small dry cell batteries. Thus, to manifest interoperability, it is desirable for network systems to support communications between two endpoint devices employing different types of coders.

☆ This work was supported by BERC-KOSEF.
* Corresponding author. Tel.: +82-2-2123-4534; fax: +82-2-312-4584.
E-mail address: yocello@mcsp.yonsei.ac.kr (S.-W. Yoon).

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.01.004

A simple approach to overcome this interoperability problem is to place the decoder/encoder of one coder and the encoder/decoder of the other coder in tandem. However, the encoder–decoder tandem is often associated with several problems. First of all, quality degradation of the synthesized speech is inevitable because a speech signal is encoded and decoded twice by the two different speech coders. Quantization errors due to each encoding process are accumulated, which results in the degradation of speech quality. In the case of cross tandeming, the mean opinion score (MOS) drops more than that of single tandeming (Campos Neto and Crcoran, 1999, 2000). Second, computational demands for the decoder–encoder tandem may be too high to implement it into the service systems for a large number of users. When the tandem algorithms are placed in the gateways of cable networks or the base stations of wireless networks, the complexity eventually will be limit of the full communication systems. Finally, tandem coding is accompanied with an additional delay in the communication links because an additional look-ahead delay for the LPC analysis is inevitable to obtain target bitstreams.

Unlike the tandem coding, *transcoding* can be used to overcome these difficulties. Transcoding is to translate source bitstreams to target ones without running through complete decoding–encoding processes. Thus, it can minimize the quality degradation of the synthesized speech and computational complexities with no additional delay.

Past researches on transcoding, especially (Campos Neto and Crcoran, 1999) which proposed the transcoding between G.729 and IS-641 (TIA/EIA PN-3467, 1996), was focused on the conversion of LSP and gain information. In the process of the other encoding modules such as adaptive-codebook delay search and fixed-codebook index search, both coders share the information for excitation of each other. In other words, the pitch delay and fixed-codebook index are not changed, and directly mapped from one coder to the other coder. It is possible because of the many similarities in the structure of comprising the excitation. This algorithm showed the improved performance compared to cross tandeming connection. But, if the direct mapping is not possible due to dissimilarities in quantizing the excitation signal, the performance of the direct mapping method will not be guaranteed. In that case, the other algorithms such as partial encoding process that uses the source coder information to reduce the complexity of the encoding process in the target coder maintaining the quality should be applied.

In this paper, we propose an efficient *transcoding* algorithm working with 5.3/6.3 kbit/s G.723.1 and 8 kbit/s G.729A (ITU-T Rec. G.729, 1996b) speech coders. According to the survey in (Hersent et al., 2000), two coders are the most widely deployed standards in VoIP systems nowadays. Considering the frame length of two-speech coders, parameters corresponding to one frame of G.723.1 are converted to three sets of equivalent G.729A parameters. The proposed transcoding algorithm is composed of four processing steps: line spectral pair (LSP) conversion, pitch interval conversion, fast adaptive-codebook search, and fast fixed-codebook search. In the LSP conversion, since variation of LSP trajectory in a steady state speech segment of each coder is somehow negligible in quality aspects, we directly convert them in the LSP domain using a codebook matching procedure. For the pitch interval conversion, we cannot use a direct mapping method because the search ranges of pitch intervals in two coders are different. From the observation that pitch contour is monotonically varied with a small variation in voiced or voice-like speech signals, we estimate the open-loop pitch in constrained regions relating to the pitch values of the adjacent frames. This is an efficient way in the respect of complexity reduction. In the adaptive-codebook and fixed-codebook search process, the codebook structures of the two coders are different from each other. We apply a fast search algorithm for reducing the complexity of the transcoding system. Using the above four processing steps, the proposed transcoding algorithm translates bitstreams of the source coder to those of the target coder with minimum distortion and reduced complexity.

To verify the performance of the proposed algorithm, we perform objective and subjective tests such as LPC cepstral distance (LPC-CD)

(Kitawaki et al., 1998) and perceptual evaluation of speech quality (PESQ) (ITU-T Rec. P.862, 2000) with various speech sets. In addition we compare the complexity of them to that of the tandem algorithm by implementing the algorithms using TI TMS320C6201 DSP processor (Texas Instruments, 1996). Evaluation results confirm that the proposed transcoding achieves the reduction of overall complexity with a shorter processing delay. Also, the preference tests reveal the superiority of the proposed transcoding algorithm in comparison to the tandem.

This paper is organized as follows. G.723.1 and G.729 speech coding algorithms are briefly described in Section 2. In Section 3, transcoding algorithms from G.723.1 to G.729A and from G.729A to G.723.1 are proposed. Section 4 describes the performance evaluation of the proposed transcoding algorithm. Conclusions are presented in Section 5.

2. ITU-T G.723.1 and G.729A

Transcoding algorithm in this paper is associated with ITU-T G.723.1 (ITU-T Rec. G.723.1, 1996) and ITU-T G.729A (ITU-T Rec. G.729, 1996b). Each speech coder is widely used for multimedia communication with low bit-rate applications such as VoIP and digital cellular. Their operational features are briefly described in this chapter.

2.1. ITU-T G.723.1 speech coder

ITU-T G.723.1 (ITU-T Rec. G.723.1, 1996), the standard for the multimedia communication speech coder, has two modes whose bit rates are 5.3 and 6.3 kbit/s. G.723.1 takes 30 ms of speech or other audio signals for encoding, and signals with the same length are reproduced by the decoder. In addition, there is a look-ahead of 7.5 ms, resulting in a total algorithmic delay of 37.5 ms.

For every 60-sample sub-frame, a set of 10th-order linear prediction coefficients (LPC) is computed. The LPC set of the last sub-frame is quantized using a predictive split vector quantizer (PSVQ). The unquantized LPC coefficients are used to construct the short-term perceptual weighting filter which is used for filtering the entire frame to obtain the perceptual weighted speech signal. For every two sub-frames, the open-loop pitch period is computed using the weighted speech signal. Later, the speech signal runs through the adaptive and fixed-codebook search procedures on a sub-frame basis. The adaptive-codebook search is performed using a 5th-order pitch predictor, and the closed-loop pitch and pitch gain are computed. Finally, the non-periodic component of the excitation is approximated by two types of fixed codebooks: multi-pulse maximum likelihood quantization (MP-MLQ) for a higher bit rate, and an algebraic code excited linear prediction (ACELP) for a lower bit rate.

2.2. ITU-T G.729 speech coder

ITU-T 8 kbit/s G.729 (ITU-T Rec. G.729, 1996a) is developed based on conjugated-structure ACELP (CS-ACELP). G.729 operates on the frame length of 10 ms, and there is 5 ms look-ahead for linear prediction (LP) analysis, resulting in a total 15 ms algorithmic delay. For every 10 ms frame, 10th-order LPC coefficients are computed using the Levinson–Durbin recursion. The LPC coefficients for the 2nd sub-frame are quantized using multi-stage vector quantization (MSVQ). The unquantized LPC coefficients are used to construct the short-term perceptual weighting filter. After computing the weighted speech signal, the open-loop pitch period is computed. To avoid pitch multipling errors, the delay range is divided into three sections, and smaller pitch values are favored. Then, the adaptive- and fixed-codebooks are searched to find optimum codewords. The adaptive-codebook search is performed using a 1st-order pitch predictor, and a fractional pitch delay is searched in a 1/3 sample resolution. In the fixed-codebook search, non-periodic components of excitation are modeled using algebraic codebooks with four pulses. For the efficiency of quantization of pitch- and fixed-codebook gains, two codebooks with conjugate structures are used.

G.729 Annex A is a complexity-reduced version of G.729 (ITU-T Rec. G.729, 1996b). The complexity of G.729A is about 50% of G.729 (Salami

et al., 1997a). Algorithmic differences between G.729 and G.729A are found in (Salami et al., 1997b).

3. The proposed transcoding algorithm

A simple transcoding algorithm can be obtained by placing the decoder/encoder of one coder and the encoder/decoder of the other coder in tandem, as shown in Fig. 1(a). However, the decoder–encoder tandem often raises several problems such as degradation of speech quality, high computational load, and additional delay. All these problems are due to the fact that the speech signal should pass complete processes for the decoding and encoding of two associated coders.

Comparing with the tandem coding, *transcoding*, a direct translation of one bitstream to another, is more beneficial. Fig. 1(b) shows a block diagram of the transcoding. The transcoding can avoid the degradation of the synthesized speech quality because the several source parameters are directly translated in the parametric domain instead of being re-estimated from the decoded PCM data. In this respect, transcoding algorithm is expected to show less distortion than the tandem coding. Also, no additional delay is required, and a reduced computational load is expected.

3.1. Transcoding from G.723.1 to G.729A

The transcoding algorithm proposed in this paper has an asymmetric structure between Tx and Rx paths. For the transmission of speech data from G.723.1 to G.729A, the transcoding process is involved with the LSP conversion and the open-loop pitch conversion. Transcoding is executed on the basis of G.723.1 frame size; thus one G.723.1 frame is converted to three frames of G.729A. A block diagram of the proposed transcoding algorithm from G.723.1 to G.729A is shown in Fig. 2. The left dotted box in the figure corresponds to the decoder module of G.723.1 and the right one corresponds to the encoder module of G.729A.

For transcoding from the G.723.1 and G.729A, LSP sets of G.723.1 are converted to those of G.729A, and they are quantized by the G729A encoding scheme. Then, the quantized LSP sets are converted to LPC again, so the perceptually weighted synthesis filter is constructed for each sub-frame. Afterwards, the target signal for open-loop pitch estimation is computed by filtering the excitation through the perceptually weighted filter. For the open-loop pitch estimation of G.729A, a pitch smoothing technique with the closed-loop pitch intervals of G.723.1 and G.729A is used. Considering the similarity and the continuity of the pitch parameter for the consecutive frames, the

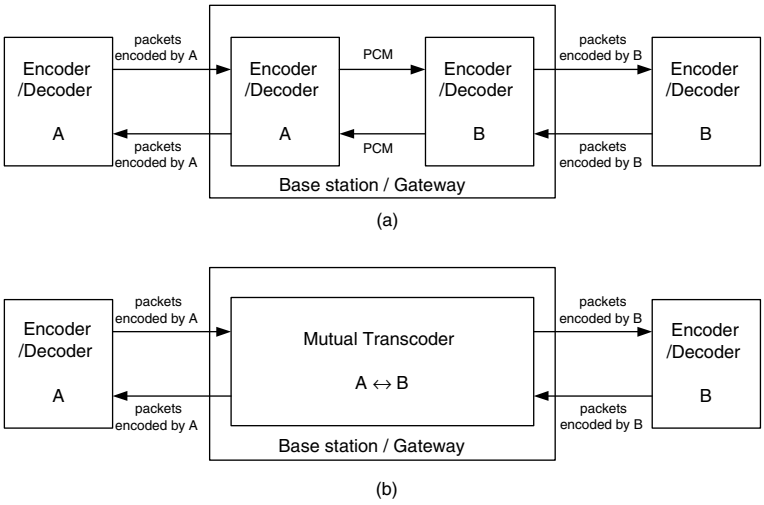

Fig. 1. (a) Tandem and (b) transcoding.

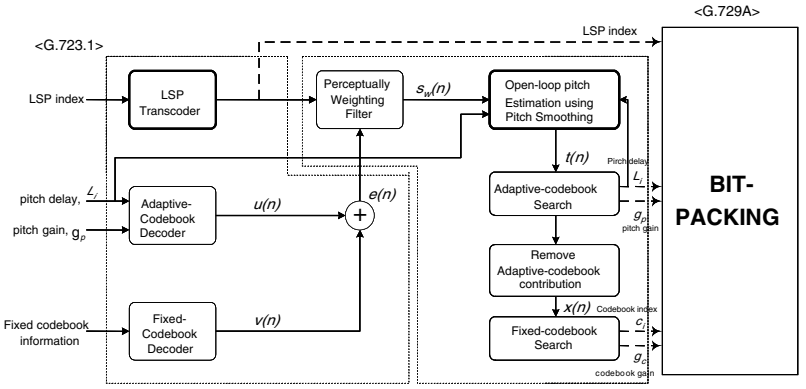

Fig. 2. Block diagram of transcoding from G.723.1 to G.729A.

pitch smoothing method estimates the open-loop pitch in the constrained narrow region, around the previous closed-loop pitch value of G.729A or the current closed-loop pitch of G.723.1. Using this approach, it is possible to generate equivalent speech quality to tandem coding while spending much less computational load.

After determining the open-loop pitch, the adaptive- and fixed-codebooks are searched for G.729. The closed-loop pitch of the second sub-frame at the current processing frame of G.729 is used for a candidate of the open-loop pitch estimation for the next frame. It means that the open-loop pitch of the next frame is estimated around the closed-loop pitch of the second sub-frame obtained in the current frame. After the translation of LSP parameters and the open-loop pitch from G.723.1 to G.729A, the adaptive-codebook and the fixed-codebook parameters of G.729A are determined by the standard encoding process. Finally, the parameters of G.729A are encoded to the bitstream and transmitted to the decoder of G.729A.

3.1.1. LSP conversion using linear interpolation

A linear interpolation technique is employed to translate the LSP parameters. Given a set of G.723.1 LSP parameters, three frame sets of G.729A LSP parameters are computed because the frame length of G.723.1 is three times longer than that of G.729A. Fig. 3 shows the LSP conversion procedure, which can be denoted by

$$p_i^B(j) = \begin{cases} p_1^A(j), & i = 1, \\ \frac{1}{2}(p_2^A(j) + p_3^A(j)), & i = 2, \\ p_4^A(j), & i = 3, \end{cases} \quad 1 \leqslant j \leqslant 10, \tag{1}$$

where p^A and p^B are the LSP parameters of G.723.1 and G.729A, respectively, and i denotes the frame index of G.729A.

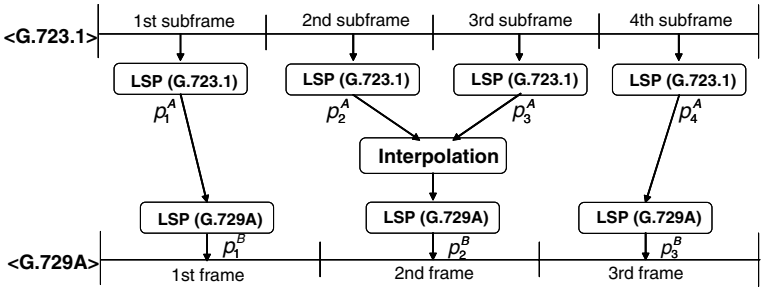

Fig. 3. LSP conversion from G.723.1 to G.729A using linear interpolation.

To compare the performance of the tandem and the LSP conversion using linear interpolation, we measured the spectral distortion to validate the efficiency of the proposed method. The decoded LPC coefficients of G.723.1 were used as a reference, to compare the distortions of the proposed and tandem method, because we intended to show the distortion added to the original spectrum. As shown in Table 1, the spectral distortion of the transcoding is much less than that of tandem. Moreover the both outliers of 2–4 dB and above the 4 dB are extremely small for transcoding cases. We also compared the LPC spectrum of the tandem and transcoding to that of the original one. Fig. 4 shows the LPC spectrum in the voiced region of the speech signal, in which the LPC spectrum of G.729A is also shown as a reference. As shown in Fig. 4, the LPC spectrum obtained after the LSP conversion given in Eq. (1) matches closely to the reference spectrum, especially in the low frequency region and around the formant. The LPC spectrum after the decoder–encoder tandem, however, indicates a larger spectral distortion than the proposed LSP conversion. Since speech quality is mainly determined from the accuracy of low and formant frequency region components (Rabiner and Schafer, 1978), it can be said that the proposed LSP conversion technique can provide better speech quality than the tandem.

Table 1
Spectral distortion—from G.723.1 to G.729A

Method	Average SD [dB]	Outliers (%)	
		2–4 dB	>4 dB
Tandem	2.81	61.71	14.03
Transcoding	0.81	0.29	0.00

Also, it should be mentioned that the proposed LSP conversion method is more beneficial than the simple tandem with regard to complexity and processing delay. As described in (Kang et al., 2000), the total delay of a speech communication system is decided by three factors: algorithmic delay, processing delay, and network delay. Considering the algorithmic and processing delays only, the total delays of the decoder–encoder tandem and direct LSP conversion, denoted by $D_{AB}^{\tan}$ and D_{AB}^{trans} are written by

$$D_{AB}^{\tan} = 42.5 + \alpha_A + \beta_A + \alpha_B + \beta_B, \tag{2}$$

$$D_{AB}^{\text{trans}} = 37.5 + \alpha_A + P_{AB} + \beta_B, \tag{3}$$

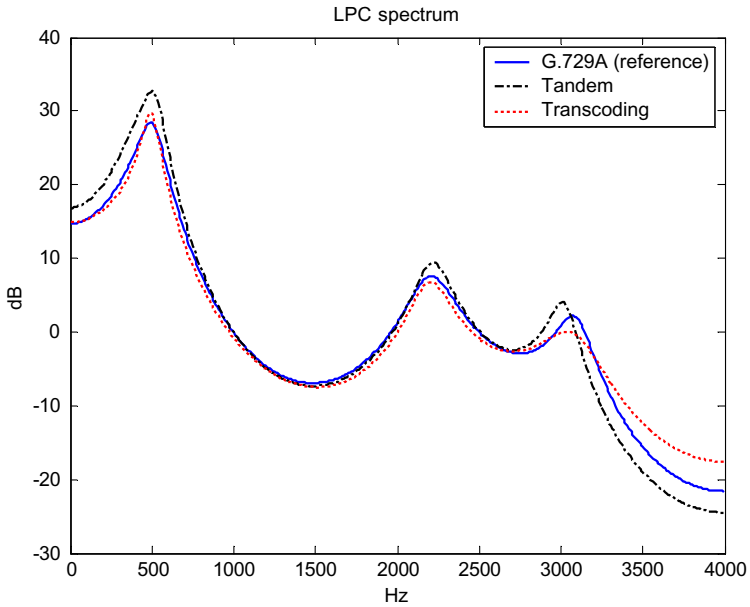

Fig. 4. Comparison of LPC spectrum (from G.723.1 to G.729A).

Table 2
Transmission delay—from G.723.1 to G.729A

	Operation	Tandem (ms)	Transcoding (ms)
A (G.723.1)	Buffering	37.5	37.5
	Encoding	p_A^E	p_A^E
Intermediate processing	Decoding	p_A^D	p_{AB}^{TR}
	Encoding	$3 \times p_B^E$	
	Delay	5	0
B (G.729A)	Decoding	$3 \times p_B^D$	$3 \times p_B^D$
Total delay (ms)		$42.5 + p_A^E + p_A^D + 3(p_B^E + p_B^D)$	$37.5 + p_A^E + p_{AB}^{\mathrm{TR}} + 3p_B^D$

where subscripts A and B denote G.723.1 and G.729A, respectively, and AB denotes the connection from A to B. Also, α_m and β_m ($m = A$ or B) denote encoder and decoder delays, respectively, and P_{AB} denotes the delay introduced by the proposed LSP conversion method. In the decoder–encoder tandem, an additional 5 ms look-ahead is needed for LPC analysis. But the proposed LSP conversion technique does not introduce this look-ahead delay because the processing of LPC analysis is not executed. As shown in Eqs. (2) and (3), the total delay of the proposed method is at least 5 ms less than that of the decoder–encoder tandem. Because the processing delay, P_{AB}, is less than the sum of β_A and α_B, the total delay will be further reduced. The delays induced by the decoder–encoder tandem and the LSP conversion technique are summarized in Table 2.

In this table, $p_A^E, p_A^D, p_B^E, p_B^D, p_{AB}^{\mathrm{TR}}$ is the processing time of encoding of G.723.1, decoding of G.723.1, encoding of G.729A, decoding of G.729A, and transcoding from G.723.1 and G.729A, respectively.

When the LSP conversion is completed, the perceptual weighting filter for the G.729A encoder is constructed using the converted LSP parameters. Then, the perceptually weighted speech signal for the open-loop pitch estimation is generated using the perceptual weighting filter.

3.1.2. Open-loop pitch estimation using a pitch smoothing

After the LSP conversion, the open-loop pitch for each frame of G.729A is estimated. In the proposed transcoding algorithm, the open-loop pitch is searched around the interval between the pitch values of the source and target coder corresponding to the adjacent sub-frame (Yoon et al., 2001).

The proposed open-loop pitch estimation is simplified using a pitch smoothing scheme shown in Fig. 5. At first, the closed-loop pitch of G.723.1 is compared with the pitch value obtained at the 2nd sub-frame of the previous G.729A frame. If the distance between the two pitch values is less than 10 samples, by considering the continuity of the pitch values, the closed-loop pitch of G.723.1 is determined as the open-loop pitch of G.729A. Otherwise, the pitch smoothing method is applied. To determine the threshold value of 10 samples for applying the pitch smoothing method in pitch interval conversion, the deviation of the adaptive-codebook pitch search range, depending on the corresponding open-loop pitch, is considered. [1]

If the pitch difference is larger than 10 samples, local delays maximizing $R(k_i)$ in Eq. (4) are searched in the range of ± 3 samples around the closed-loop pitch delays of G.723.1 and G.729A, respectively:

$$R(k_i) = \sum_{n=0}^{79} s_w(n) \cdot s_w(n - k_i), \qquad p_i - 3 \leqslant k_i \leqslant p_i + 3 (i = A, B), \tag{4}$$

where $s_w(n)$ is the weighted speech signal, subscripts A and B, respectively, denote G.723.1 and G.729A, $p_i s$ are the closed-loop pitch delays, and $k_i s$ are the open-loop pitch candidates.

[1] For example, the closed-loop pitch search range of G.729A is from −5 to +4 sample around the open-loop pitch or preceding sub-frame pitch.

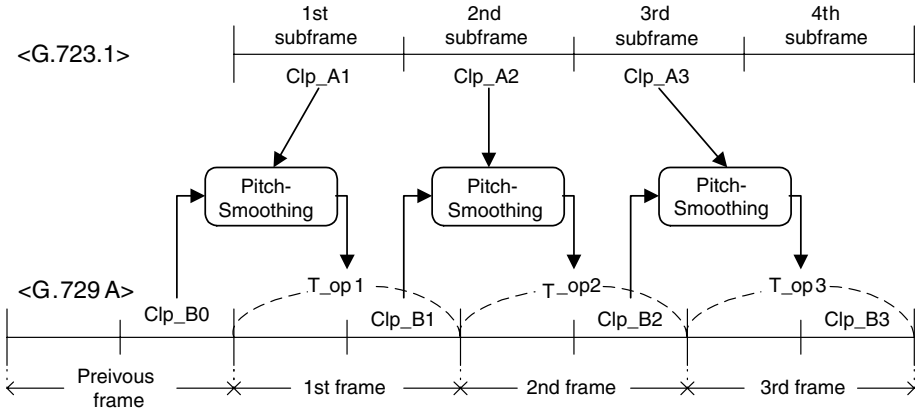

Fig. 5. Open-loop pitch estimation using pitch smoothing (G.723.1 → G.729A).

After determining the local delays, t_i, $i = A, B$ maximizing $R(k_i)$ in each range, $R(k_i)$ is normalized by the energy at the local maximum delays:

$$R'(t_i) = \frac{R(t_i)}{\sqrt{\sum_n s_w^2(n - t_i)}}, \quad i = A, B. \tag{5}$$

To evaluate the validity of this pitch estimation approach, we perform two experiments such as observation of the pitch contour of two coders and comparison of the estimated open-loop pitch contour of the target coder with the adaptive-codebook pitch contour of the source and target coders. The adaptive-codebook pitch contour of target coder matches well with the estimated open-loop pitch contour rather than the pitch value of source one. We infer from the observation that the variation of pitch value between adjacent subframes is relatively stable and the estimated open-loop pitch is close to the previous closed-loop pitch value. Finally, the normalized local maxima are compared with each other, and the local maximum from G.729A is favored. Thus, if the local maximum of G.729A is larger than 3/4 times of that of G.723.1, the open-loop pitch of G.729A is determined as the local maximum delay of G.729A. [2] Otherwise, the local maximum delay of G.723.1 is selected. Consequently, the smoothed open-loop pitch, T_{op}, is determined as

$$\begin{aligned}
&T_{\text{op}} = t_1 \\
&R'(T_{\text{op}}) = R'(t_1) \\
&\textit{if } R'(t_2) \geqslant 0.75 \cdot R'(T_{\text{op}}) \\
&\quad R'(T_{\text{op}}) = R'(t_2) \\
&\quad T_{\text{op}} = t_2 \\
&\text{end}
\end{aligned}$$

As shown in Fig. 6, the pitch contour of the proposed method, dashed line one, matches well with the original pitch value of G.729A without any severe fluctuation. But, in the case of the tandem approach, a drastic fluctuation appears such as pitch multiple errors even in the stable voiced speech segment.

The word "smoothed" means that the rapid and sudden pitch variation often produced by the estimation in distorted speech segments, an even voiced or voice-like region, is avoided, so the estimated open-loop pitch can represent the monotonically varying contour.

In the proposed scheme, the autocorrelation of the weighted speech is either computed around the constrained region or not at all. Also, as the local maximum value from the search process of original G.729A encoder in full search range is not involved, the open-loop pitch can be estimated with much less computational load in the constrained search range. Furthermore, the pitch

[2] The 3/4 was determined from the results of our experiments based on large number of test database.

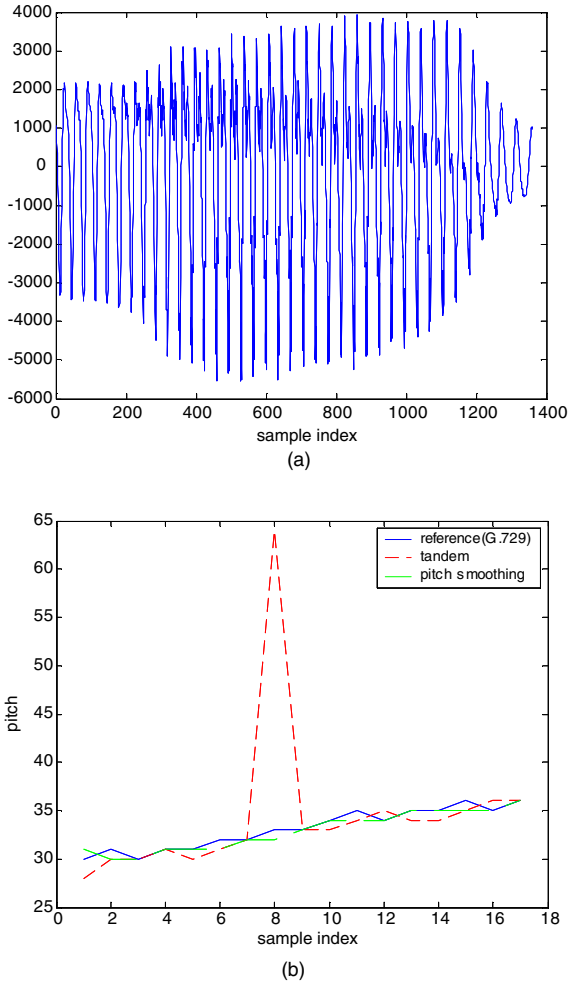

Fig. 6. Comparison of the open-loop pitch contour. (a) Voiced speech segment. (b) The estimated open-loop pitch contour.

smoothing scheme can reduce the drastic fluctuation of pitch contour in a voiced or voice-like segment, as shown in Fig. 6, like that of the tandem approach which results in sticky noise due to inaccurate pitch information.

3.2. From G.729A to G.723.1

For the case of speech transmission from G.729A to G.723.1, the proposed algorithm consists of LSP conversion using linear interpolation, pitch smoothing for the open-loop pitch conversion, fast adaptive-codebook search, and fast fixed-codebook search. Parameters corresponding to three frames of G.729A are collected and simultaneously converted to parameters corresponding to one frame of G.723.1. A block diagram of the proposed transcoding algorithm from G.729A to G.723.1 is shown in Fig. 7. The overall structure is similar to the connection from G.723.1 to G.729A, but fast adaptive-codebook and fixed-codebook search modules are different from G.723.1 to G.729A.

3.2.1. LSP conversion using a linear interpolation

LSP conversion from G.729A to G.723.1 is accomplished with parameters corresponding to three frames of G.729A. However, only the 4th frame parameters of G.723.1 are computed via a linear interpolation, as given by

$$p^A(j) = w_1 \cdot p_1^B(j) + w_2 \cdot p_2^B(j) + w_3 \cdot p_3^B(j), \quad 1 \leqslant j \leqslant 10, \tag{6}$$

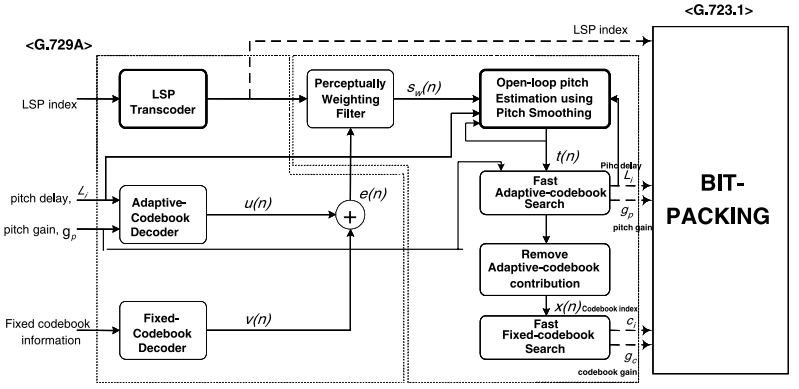

Fig. 7. Block diagram of transcoding from G.729A to G.723.1.

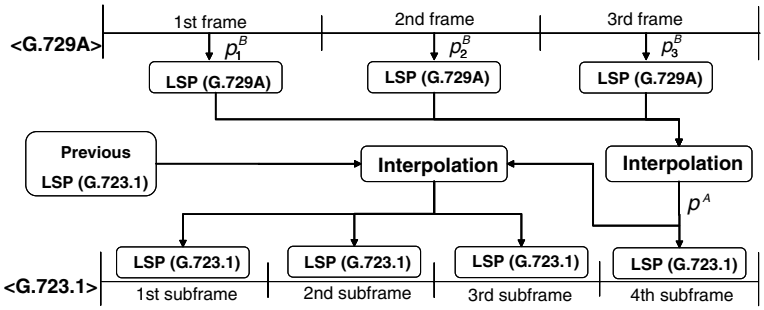

Fig. 8. LSP conversion from G.729A to G.723.1 using linear interpolation.

where superscripts A and B again denote parameters corresponding to G.723.1 and G.729, respectively, and subscript denotes frame indices of G.729. w_1, w_2, and w_3 are weighting factors. They are usually decided by considering geometric distance (Epperson, 2002). We set them at 0.05, 0.15 and 0.80 respectively. The developed LSP conversion process is illustrated in Fig. 8. In this transcoding algorithm, only the LSP set of the 4th sub-frame is computed by linear interpolation, because the LPC sets of the other sub-frames are computed by linear interpolation using the LSP information of 4th sub-frames between the previous and current frame. To compare the performance of the tandem and the LSP conversion using linear interpolation in this direction, we measure the spectral distortion. The decoded LPC coefficients of G.729A were used as a reference. As shown in Table 3, the spectral distortion of the LSP conversion is much less than that of tandem. Moreover the both outliers of 2–4 dB and above the 4 dB are extremely less, too. Also, as shown in Fig. 9, the LPC spectrum of the transcoding matches well with the reference spectrum obtained from the original G.723.1 encoder compared to that of the tandem coding.

Table 3
Spectral distortion—from G.729A to G.723.1

Method	Average SD [dB]	Outliers (%)	
		2–4 dB	>4 dB
Tandem	2.90	66.29	14.47
Transcoding	1.15	3.90	0.02

As described in Section 3.1.1, the proposed approach reduces overall complexity and overall processing delays. Processing delays are summarized and compared in Table 4, LPC spectra obtained in a voiced region shown in Fig. 9. As shown in the figure, the LPC spectrum of the proposed LSP conversion scheme matches well with the reference compared to the tandem coding.

3.2.2. Open-loop pitch estimation using pitch smoothing

Similar to the case from G.723.1 to G.729A, the perceptually weighted speech signal is computed as a target signal for the open-loop pitch estimation of G.723.1. The same pitch smoothing technique is applied. A block diagram of the developed open-loop pitch estimation is shown in Fig. 10. As a cost function, the normalized cross correlation at the pitch-lag candidate is computed, which have been used for the original open-loop pitch estimation module in G.723.1 (ITU-T Rec. G.723.1, 1996).

3.2.3. Fast adaptive-codebook search

The adaptive-codebook search in G.723.1 uses a 5th order pitch predictor, and estimates the pitch delay and gain simultaneously. Thus, the adaptive codebook search in G.723.1 is computationally demanding mainly because the pitch delay and pitch gain are searched simultaneously. Previously, we proposed a fast adaptive-codebook search algorithm (Jung et al., 2001). In this algorithm, the pitch delay and pitch gain are computed sequentially, rather than simultaneously. At first, the pitch delay is computed using a 1st-order pitch predictor, and later, the pitch gains of the 5th

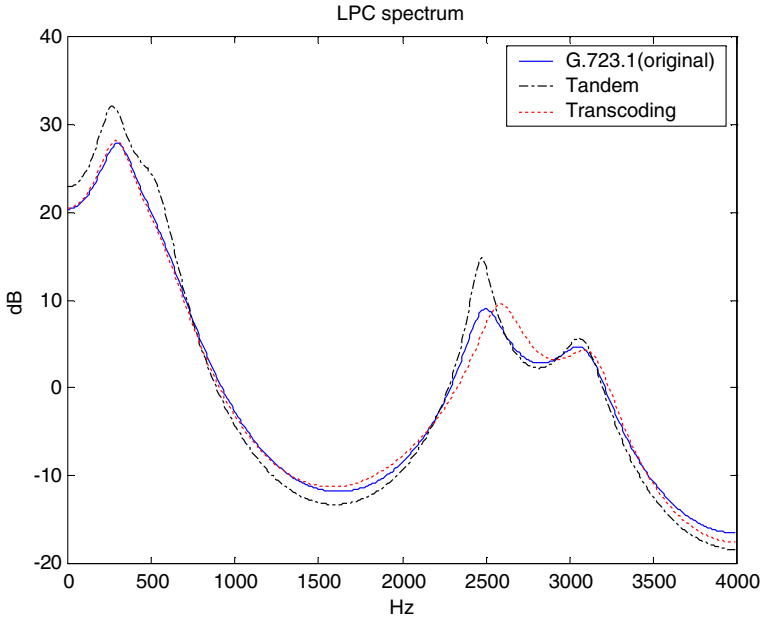

Fig. 9. Comparison of LPC spectrum (from G.729A to G.723.1).

Table 4
Transmission delay—from G.729A to G.723.1

	Operation	Tandem (ms)	Transcoding (ms)
B (G.729A)	Buffering	35	35
	Encoding	$3 \times p_B^E$	$3 \times p_B^E$
Intermediate processing	Decoding	$3 \times p_A^E$	p_{BA}^{TR}
	Encoding	p_A^E	
	Delay	5	0
A (G.723.1)	Decoding	p_A^D	p_A^D
Total delay (ms)		$40 + 3(p_B^E + p_B^D) + p_A^E + p_A^D$	$25 + p_A^E + p_{BA}^{\mathrm{TR}} + 2p_B^D$

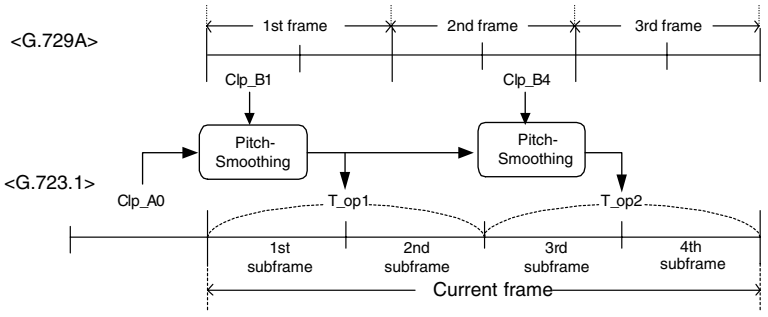

Fig. 10. Open-loop pitch estimation using pitch smoothing (G.729A → G.723.1).

order pitch predictor are estimated using the computed pitch delay. The pitch gain is computed using the function given by

$$MSE'_{ACB} = \sum_{n=0}^{N-1}\{t[n] - g\hat{t}[n-L]\}^2 = \sum_{n=0}^{N-1} t^2[n] - 2g\sum_{n=0}^{N-1} t[n]\hat{t}[n-L] + g^2\sum_{n=0}^{N-1}\hat{t}^2[n-L], \quad (7)$$

where $t[n]$ and $\hat{t}[n]$ are target signals of the adaptive-codebook search and weighted synthesis signal, respectively. g is the pitch gain of the 1st-order pitch predictor, and N is the sub-frame size. Optimum pitch delay minimizing Eq. (7) is obtained by differentiating Eq. (7) with respect to g and by setting the derivative to zero:

$$g = \frac{\sum_{n=0}^{N-1} t[n]\hat{t}[n-L]}{\sum_{n=0}^{N-1}\hat{t}^2[n-L]}. \quad (8)$$

Substituting Eq. (8) into Eq. (7) gives an expression for the weighted squared error in terms of the pitch delay, and we can conclude that minimizing Eq. (7) is equivalent to maximizing

$$C'_{\text{lag}} = \frac{\left(\sum_{n=0}^{N-1} t[n]\hat{t}[n-L]\right)^2}{\sum_{n=0}^{N-1}\hat{t}^2[n-L]}. \quad (9)$$

Finally, L maximizing C'_{lag} is determined from a closed-loop search procedure. This scheme is computationally very efficient, and there is no quality loss compared to the original coder (Jung et al., 2001).

In addition to above process, we apply an extra algorithm in the vector quantization of pitch gains. The pitch gain of the G.723.1 coder is vector-quantized using an 85 or 170-entry codebook, 170 for lower rate and 85 or 170 for higher rate, respectively. This process is another major computational burden for implementation. In the developed transcoding algorithm, the search range of the gain codebook is limited depending on the pitch gain of G.729A. In other words, the indices of the adaptive-codebook gain table are pre-selected depending on speech signal characteristics. The process of the gain index table generation or pre-selection is shown in Fig. 11.

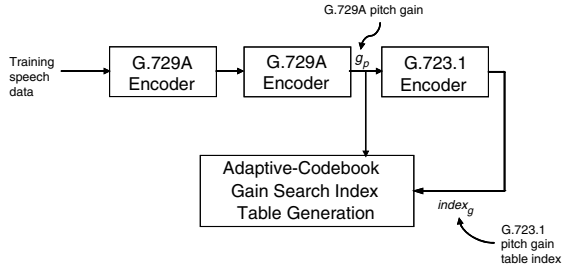

Fig. 11. Gain index table generation for fast adaptive-codebook search.

In this approach, the similarity or relationship of pitch gains of each coder is considered. As shown in Fig. 11, the decoded PCM signal from G.729A is encoded by G.723.1, like a tandem connection. In this process, we can find the statistical information between the pitch gains of two coders. The dynamic range of the pitch gain value of G.729A is from 0 to 1.2. We divide this pitch gain range into eight sub-sections, and the conjugate structure of G.729A gain codebook is considered for the boundary value of each sub-section. Then, in the following encoding process of G.723.1, the determined adaptive-codebook gain table indices are stored at each sub-frame. As a result, the distribution of the most probable top 40 or 85 gain indices for 85 or 170 entry gain table, respectively, can be listed up at each pitch gain sub-section of G.729A. For reliability of gain index distribution, we use speech signals recorded by a female and male speaker, with each sentence 8 s long, and a total of 96 sentences per speaker. Results of the subjective listening test confirm that no degradation of speech quality was introduced.

3.2.4. Fast fixed-codebook search

In the fixed-codebook search of G.723.1 at 5.3 kbit/s, four pulses are searched based on an ACELP structure for every sub-frame (ITU-T Rec. G.723.1, 1996). Each sub-frame is divided into four tracks, and the pulse and sign of each pulse are determined using a nested-loop search. In the theoretically worst case, the pulse locations are

searched with the combination of $8 \times 8 \times 8 \times 8 = 8^4$ with the analysis-by-synthesis technique. In practice, limiting the number of entering the loop for the last pulse search can significantly reduce the complexity of the fixed-codebook search. In order to simplify the fixed-codebook search of G.723.1, a depth-first tree search which is used in the G.729A fixed-codebook search is employed in this paper. Using this scheme, the combination of the pulse location can be reduced down to $2 \times \{(8 \times 8) + (8 \times 8)\}$ (Jung et al., 2001).

The fixed-codebook excitation of G.723.1 at 6.3 kbit/s is modeled by MP-MLQ based on multi pulse excitation. The fixed-codebook parameters such as quantized gain, the signs and locations of the pulses are sequentially optimized. This process is repeated for both the odd and even grids. To reduce the complexity of this module, we adopt the fast algorithm proposed in (Jung et al., 2001), which is based on ACELP search. In this fast search algorithm, the 60 positions in a sub-frame are divided into several tracks. Each pulse can have either amplitude +1 or −1. The result in (Jung et al., 2001) shows that the reduction ratio is about 80% when fully optimized for DSP implementation.

4. Performance evaluations

To evaluate the performance of the proposed transcoding algorithm, subjective preference tests are performed together with an objective quality evaluation and complexity check. Subjective quality is evaluated via an *A–B* preference test, and as an objective quality measure, LPC-CD and PESQ are used.

4.1. Objective quality evaluation

For objective evaluation measures, the NTT Korean speech database is used (NTT-AT, 1994). Each sentence is 8 s long with 8 kHz sampling frequency, four male and female speakers with 24 sentences per speaker are recorded under a quiet environment, so a total of 96 sentences were used for LPC-CD and PESQ evaluation. LPC-CD (Kitawaki et al., 1998) is defined as

$$\mathrm{LPC\_CD} = 10/\log_{10}\sqrt{2\sum_{i=1}^{p}(C_x(i) - C_y(i))^2}, \tag{10}$$

where $C_x(i)$ and $C_y(i)$ are LPC cepstral coefficients of the input and output of the codec, and p is the order of LPC filter. Measurement results are summarized in Tables 5 and 6, where lower LPC-CD and highly PESQ imply better speech quality. As shown, for both measures, the proposed transcoding scheme is judged as having better quality than tandem coding.

4.2. Subjective quality evaluation

For subjective evaluations, informal *A–B* preference tests are conducted. The tests are performed by 30 naive listeners. In the tests, the subjects are asked to choose a sound between pairs of samples presented through a headset. If the subjects can not distinguish the quality difference, they are

Table 5
Objective test results (G.723.1 at 5.3 kbit/s)

	LPC-CD (dB)		PESQ	
	Female	Male	Female	Male
Tandem (*A–B*)	3.98	3.90	3.023	3.362
Transcoding (*A–B*)	3.66	3.54	3.051	3.376
Tandem (*B–A*)	4.17	3.65	3.017	3.384
Transcoding (*B–A*)	3.86	3.22	3.077	3.426

Notice: A to *B* (from G.723.1 to G.729A), *B* to *A* (from G.729A to G.723.1).

Table 6
Objective test results (G.723.1 at 6.3 kbit/s)

	LPC-CD (dB)		PESQ	
	Female	Male	Female	Male
Tandem (*A–B*)	3.95	3.86	3.106	3.465
Transcoding (*A–B*)	3.53	3.37	3.118	3.455
Tandem (*B–A*)	3.86	3.59	3.107	3.492
Transcoding (*B–A*)	3.70	3.11	3.133	3.507

Notice: A to *B* (from G.723.1 to G.729A), *B* to *A* (from G.729A to G.723.1).

Table 7
Preference test results (G.723.1 at 5.3 kbit/s)

Preference	G.723.1 → G.729A		G.729A → G.723.1	
	Female (%)	Male (%)	Female (%)	Male (%)
Tandem	26.9	30.6	26.3	35.8
Transcoding	42.5	36.9	25.4	45.4
No preference	30.6	32.5	48.3	18.8

Table 8
Preference test results (G.723.1 at 6.3 kbit/s)

Preference	G.723.1 → G.729A		G.729A → G.723.1	
	Female (%)	Male (%)	Female (%)	Male (%)
Tandem	32.4	35.8	39.8	38.6
Transcoding	47.7	45.4	41.4	41.5
No Preference	19.9	18.8	18.8	19.9

asked to choose "no preference". The test material includes 16 clean speech sentences obtained with four male and four female speakers. Tables 7 and 8 show the results. As shown, tandem and transcoding are preferred in a similar ratio or slightly preferred to transcoding. Results imply that the listeners cannot distinguish the quality of tandem coding from that of transcoding. Thus, it can be inferred that the proposed transcoding algorithm produces speech with a quality equivalent to that of tandem coding.

4.3. Complexity check

To check the complexity of the proposed algorithm, both tandem and transcoding algorithms are implemented on a TI TMS320C6201 DSP processor (Texas Instruments, 1996, 1998). Since our purpose is to compare the relative complexity both methods, we do not use a code optimization process. And the complexity for total encoding process can be varied because the constrained search range for open-loop pitch depends on the input signal. But, the variation of complexity in that module is extensively small. So the dependency of complexity reduction ratio on input signal is absolutely negligible.

Results in Table 9 indicate that, being compared with tandem coding, the processing time of each module is noticeably reduced by using the transcoding algorithm. Also, the total encoding time of transcoding is close to 62–74% of the encoding time needed for tandem coding. Thus, it can be said that the developed transcoding algorithm can synthesize equivalent quality to the tandem coding with complexity about 26–38% lower than tandem coding.

5. Conclusions

In this paper, we proposed an efficient transcoding algorithm that could convert 5.3/6.3 kbit/s G.723.1 bitstream into 8 kbit/s G.729A bitstream, and vice versa. This transcoding algorithm can reduce several problems caused by the tandem method, such as quality degradation, high complexity, and longer delay time. The proposed transcoding algorithm is composed of four steps: LSP conversion, open-loop pitch conversion, fast adaptive-codebook search, and fast fixed-codebook search. Subjective and objective evaluation results showed that the proposed transcoding

Table 9
Comparison of complexity

MIPS	G.723.1 → G.729A		G.729A → G.723.1(5.3k)		G.729A → G.723.1(6.3k)	
	Tandem	Transcoding	Tandem	Transcoding	Tandem	Transcoding
LPC & LSP	6.41	2.36	6.94	5.55	6.94	5.55
Open-loop	0.94	0.21	1.54	1.19	1.54	1.19
ACB	2.45	2.45	10.14	6.34	8.89	4.78
FCB	4.30	4.30	10.50	2.16	17.27	6.68
Others	4.04	4.04	8.05	8.05	8.05	8.05
Total (encoding)	18.15	13.37	37.17	23.28	42.69	26.25
Reduction ratio (%)	26.3		37.4		38.5	

algorithm could produce equivalent speech quality to the tandem coding with shorter delays and less computational complexity.

References

Campos Neto, A.F., Crcoran, F.L., 1999. Performance assessment of tandem connection of enhanced cellular coders. IEEE Proceedings of International Conference on Acoustics Speech Signal Processing, March 1999, pp. 177–180.

Epperson, J.F., 2002. Introduction to Numerical Methods and Analysis. Wiley.

Hersent, O., Gurle, D., Petit, J.-P., 2000. IP Telephony Packet-based Multimedia Communications Systems. Addison Wesley.

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

ITU-T Rec. G.729, 1996a. Coding of Speech at 8 kbit/s Using Conjugate Structure Algebraic-Code-Excited-Linear-Prediction (CS-ACELP).

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

ITU-T Rec. P.862, 2000. Perceptual evaluation of speech quality (PESQ), an objective method of end-to-end speech quality assessment of narrowband telephone networks and speech codecs, May 2000.

Jung, S.K., Park, Y.C., Yoon, S.W., Cha, I.H., Youn, D.H., 2001. A proposal of fast algorithms of ITU-T G.723.1 for efficient multi channel implementation. In: Proceedings of Eurospeech, September 2001, pp. 2017–2020.

Kang, H.G., Kim, H.K., Cox, R.V., 2000. Improving transcoding capability of speech coders in clean and frame erasured channel environments. In: Proceedings of IEEE Workshop on Speech Coding. pp. 78–80.

Kitawaki, N., Nagabuchi, H., Itoh, K., 1998. Objective quality evaluation for low-bit-rate speech coding system. IEEE J. Selected Areas Commun. 7 (2), 242–248.

NTT-AT multi-lingual speech database for telephonometry 1994. Available from <http://www.ntt-at.com/products_e/speech/index.html>.

Rabiner, L.R., Schafer, R.W., 1978. Digital Processing of Speech Signals. Prentice Hall.

Salami, R., Laflamme, C., Bessette, B., Adoul, J.P., 1997a. ITU-T G.729 Annex A: reduced complexity 8 kb/s CS-ACELP Codec for digital simultaneous voice and data. IEEE Commun. Mag., 53–63.

Salami, R., Laflamme, C., Bessette, B., Adoul, J-P., 1997b. Description of ITU-T recommendation G.729 Annex A: reduced complexity 8 kbit/s CS-ACELP Codec. In: IEEE Proceedings of International Conference Acoustic Speech Signal Processing, Vol. 2. pp. 775–778.

Texas Instruments, 1996. TMS320C62x/67x CPU and Instruction Set.

Texas Instruments, 1998. TMS320C6x C Source Debugger User's Guide.

TIA/EIA PN-3467, 1996. TDMA Radio Interface, Enhanced Full-Rate Speech Codec, February 1996.

Yoon, S.W., Jung, S.K., Park, Y.C., Youn, D.H., 2001. An efficient transcoding algorithm for G.723.1 and G.729A speech coders. In: Proceedings of Eurospeech, September 2001, pp. 2499–2502.