# A WIDEBAND CELP SPEECH CODER AT 16 KBIT/S BASED ON MEL-GENERALIZED CEPSTRAL ANALYSIS

*Kazuhito Koishida* †, *Gou Hirabayashi* †, *Keiichi Tokuda* ‡, *and Takao Kobayashi* †

†Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama, 226-8503 Japan
‡Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466 Japan
E-mail: {koishida,hirago,tkobayas}@pi.titech.ac.jp, tokuda@ics.nitech.ac.jp

**ABSTRACT**

This paper proposes a wideband CELP coder using frequency warping. Instead of linear prediction, the proposed coder adopts the mel-generalized cepstral analysis, and encodes fullband of the speech signal through a warped frequency scale. It is shown that the subjective quality of the proposed coder at 16 kbit/s is better than that of the ITU-T G.722 at 64 kbit/s. Furthermore, the proposed coder gives a much smaller difference in performance for male and female speakers than the conventional CELP coder. These results indicate that the frequency warping makes a large contribution to the improvement of the subjective quality for wideband speech coding.

## 1. INTRODUCTION

Wideband speech signals, which have a bandwidth of 50-7000 Hz, are more natural and intelligible than narrow-band speech signals with 300 to 3400 Hz bandwidth. Several applications such as teleconferencing, multimedia services and high-quality wideband telephony often require compression of wideband speech. In wideband speech, most of the important formants are typically located below 4 kHz, so that the energy at high frequencies is smaller than that at low frequencies. There are two well known techniques which efficiently utilize such characteristics; one is subband coding and another is adaptive transform coding. The basic principles of both schemes are to decompose the speech signal into subbands and encode each band separately. The ITU-T G.722 standard [1], subband ADPCM coder, is a representative of subband coding. Recently, various subband coders based on the CELP model [2] have been proposed around 16 kbit/s [3]-[5]. Other possible approach, which has not been caught so much attention in the area of wideband speech coding, is to incorporate frequency warping into spectral analysis and encode fullband of speech signal through the warped frequency scale. For narrow-band speech, we have proposed the speech coders which use frequency warping to enhance the speech quality [6]-[8].

In this paper, we investigate the effectiveness of frequency warping for wideband speech coding. The coder presented in this paper falls into fullband CELP coding. While CELP coders commonly use linear prediction (LP) as spectral estimation, the proposed CELP coder adopts the mel-generalized cepstral (MGC) analysis [9] which makes it possible to use frequency warping defined by an all-pass system. As a result, the proposed coder encodes fullband of the wideband speech signal through the warped frequency scale. The subjective performance of the proposed coder is evaluated from the following viewpoints: the effectiveness of the frequency warping, and comparison with the ITU-T G.722 standard and the conventional CELP coder.

## 2. STRUCTURE OF CELP BASED ON MEL-GENERALIZED CEPSTRAL ANALYSIS

In this section we describe the structure of the proposed CELP coder, called MGC based CELP (MGC-CELP). Since the basic algorithm is the same as the conventional CELP, we focus our discussion on the differences between the proposed and conventional coders.

### 2.1. Spectral analysis

In the MGC analysis [9], a speech spectrum $H(e^{j\omega})$ is assumed to be modeled by the MGC coefficients $c(m)$ as

$$H(z) = \begin{cases} \left(1 + \gamma \sum_{m=0}^{M} c(m)\tilde{z}^{-m}\right)^{1/\gamma}, & -1 \le \gamma < 0 \\ \exp \sum_{m=0}^{M} c(m)\tilde{z}^{-m}, & \gamma = 0 \end{cases} \tag{1}$$

where $\tilde{z}^{-1}$ is an all-pass transfer function defined by

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1. \tag{2}$$

For a sampling frequency of 16 kHz, the phase characteristics of the system give a good approximation to the mel scale when $\alpha = 0.42$. It is noted that $\gamma$ is a parameter to control the shape of poles and zeros, e.g., $H(z)$ becomes all-pole modeling for $\gamma = -1$ and cepstral modeling for $\gamma = 0$.

An optimum set of the MGC coefficients, which maximizes the expectation value of the prediction gain, can be obtained using efficient iterative algorithm based on FFT and recursive formulas [9]. In addition, such coefficients result in the stable system [9].

In the proposed coder, $\gamma$ is fixed to be $-1/2$, this leads to some advantages in the filter structure and the calculation of the quantization parameters. As concerns $\alpha$, we will investigate appropriate values through a listening test.

### 2.2. Synthesis filter

For $\gamma = -1/2$, the synthesis filter is realized by the rational transfer function of the form

$$S(z) = \frac{1}{\{C_1(\tilde{z})\}^2} \tag{3}$$

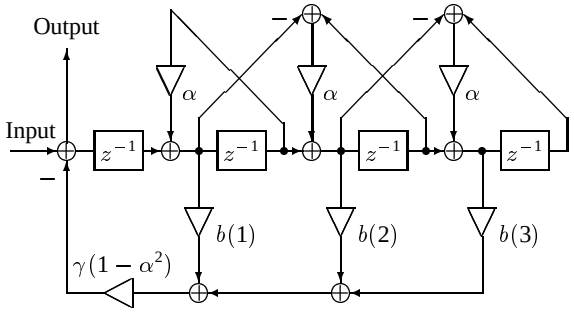



Figure 1: The structure of $1/C_1(\tilde{z})$ for $M = 3$.

where $S(z)$ is a gain-normalized version of $H(z)$ and

$$C_1(\tilde{z}) = 1 + \gamma \sum_{m=0}^{M} c_1(m)\tilde{z}^{-m}. \tag{4}$$

The coefficients $c_1(m)$ are calculated from $c(m)$:

$$c_1(m) = \begin{cases} (c(0) - \lambda)\big/(1 + \gamma\lambda), & m = 0 \\ c(m)\big/(1 + \gamma\lambda), & 1 \le m \le M \end{cases} \tag{5}$$

where

$$\lambda = \sum_{m=0}^{M} (-\alpha)^m c(m). \tag{6}$$

To remove a delay-free loop from $S(z)$, we modify (4) as follows:

$$C_1(\tilde{z}) = 1 + \gamma \sum_{m=1}^{M} b(m)\,\Phi_m(z) \tag{7}$$

where

$$\Phi_m(z) = \frac{(1-\alpha^2)z^{-1}}{1 - \alpha z^{-1}}\tilde{z}^{-(m-1)} \tag{8}$$

and the filter coefficients $b(m)$ are obtained using the recursive formula given by

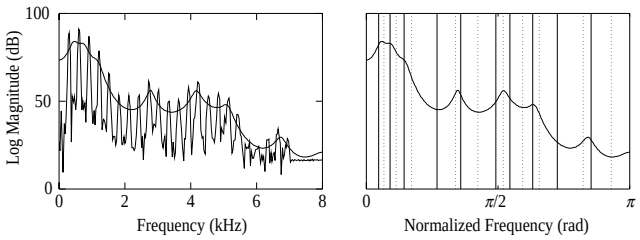
$$b(m) = \begin{cases} c_1(M), & m = M \\ c_1(m) - \alpha b(m+1), & 0 \le m \le M-1. \end{cases} \tag{9}$$

It is noted that $b(0)$ becomes zero, which means that the gain of $S(z)$ is unity. The structure of $1/C_1(\tilde{z})$ based on (7) is shown in Fig. 1.

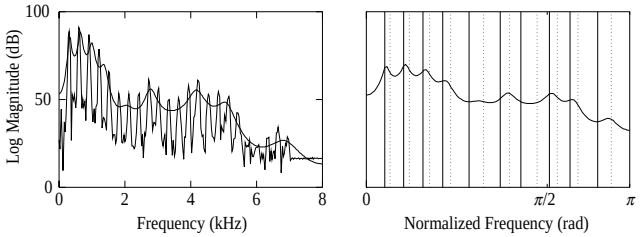
### 2.3. Quantization of MGC coefficients

Direct quantization of $c(m)$ or $c_1(m)$ may cause unstable synthesis filter. To avoid this problem, we have presented the spectral representation based on the MGC coefficients [10], referred to as MGC based LSP (MGC-LSP). The MGC-LSP is a frequency-domain representation of speech similar to LSP, and it is defined on the warped frequency scale. The procedure for obtaining the MGC-LSP parameters is as follows: First (4) is modified as

$$C_1(\tilde{z}) = (1 + \gamma c_1(0))C_2(\tilde{z}) \tag{10}$$




(a) LP spectrum and LSP parameters.




(b) MGC spectrum and MGC-LSP parameters.

Figure 2: Example of estimated spectrum and associated spectral parameters.

where

$$C_2(\tilde{z}) = 1 + \gamma \sum_{m=1}^{M} c_2(m)\tilde{z}^{-m} \tag{11}$$

and the coefficients $c_2(m)$ are calculated from the gain-normalized MGC coefficients using

$$c_2(m) = c_1(m)\big/(1 + \gamma c_1(0)), \quad 1 \le m \le M. \tag{12}$$

Next $C_2(\tilde{z})$ is decomposed into symmetric and antisymmetric polynomials:

$$C_P(\tilde{z}) = C_2(\tilde{z}) + \tilde{z}^{-(M+1)}C_2(\tilde{z}^{-1}) \tag{13}$$

$$C_Q(\tilde{z}) = C_2(\tilde{z}) - \tilde{z}^{-(M+1)}C_2(\tilde{z}^{-1}). \tag{14}$$

Finally, using the same method as LSP, the MGC-LSP parameters can be obtained as the angular positions of the roots of $C_P(\tilde{z})$ and $C_Q(\tilde{z})$.

Figure 2 shows an example of the spectra estimated by the 20th-order analysis and the associated spectral parameters. In the figure, (a) and (b) correspond to the LP and MGC analysis, respectively. In the MGC analysis, we let $\alpha = 0.3$ and $\gamma = -1/2$. It is seen that the LSP parameters concentrate in some frequency regions, while the MGC-LSP parameters are distributed almost equally. This leads to less computational complexity to transform the MGC coefficients into the MGC-LSP parameters, compared to LSP case.

### 2.4. Perceptual weighting filter and postfilter

The perceptual weighting filter is defined by the MGC coefficients as

$$S_{pw}(z) = \frac{C_1(\tilde{z}/\beta_1)}{C_1(\tilde{z}/\beta_2)} \tag{15}$$

where $\tilde{z}/\beta$ indeicates a bandwidth expansion in $\tilde{z}$-plane. The filter $C_1(\tilde{z}/\beta)$ can be realized using the same structure of $C_1(\tilde{z})$ [8]. We

Table 1: Bit allocations of 16 kbit/s wideband CELP coder.

| | Subframe | Frame |
|---|---|---|
| Spectral parameters | – | 21 |
| Power | – | 7 |
| Excitation codebook 1 | 9 | 9×4 |
| Excitation codebook 2 | 17 | 17×4 |
| Gain codebook | 7 | 7×4 |
| Total | – | 160 bits |

Table 2: Quantization performance of MGC-LSP and LSP (dB).

| | $\alpha$ (MGC-LSP) | | | | | | LSP |
|---|---|---|---|---|---|---|---|
| | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | |
| CD | 1.95 | 1.96 | 1.96 | 1.93 | 1.88 | 1.88 | 1.79 |

set the tunable parameters of the perceptual weighting to $\beta_1 = 1.0$ and $\beta_2 = 0.0$, i.e., $S_{pw}(z) = C_1(\tilde{z})$.

The short-term postfilter is defined by

$$S_{st}(z) = \frac{C_1(\tilde{z}/\beta_3)}{C_1(\tilde{z}/\beta_4)}. \tag{16}$$

The tilt compensation filter has a structure of the form

$$S_{tilt}(z) = (1 - \mu z^{-1})^p \tag{17}$$

where $\mu$ is a parameter to control the global spectral tilt. The parameter $\mu$ is determined in such a way that the first mel-cepstrum of $S_{st}(z)S_{tilt}(z)$ is set to be zero [8]. Under such constraint, $\mu$ is given by

$$\mu = \frac{-\gamma(\beta_4 - \beta_3)c_1(1)}{-\alpha\gamma(\beta_4 - \beta_3)c_1(1) + (1 - \alpha^2)p}. \tag{18}$$

By informal listening, we let $(\beta_3, \beta_4, p) = (0.8, 0.95, 2)$ for $\alpha = 0.0, 0.1$ and $0.2$, and $(0.8, 0.9, 2)$ for $\alpha = 0.3, 0.4$ and $0.5$.

## 3. FRAMEWORK OF CELP CODER AT 16 KBIT/S

This section describes each coding parameters. The input speech is sampled at 16 kHz and filtered by the sending filter P.341 with 50 to 7000 Hz bandwidth. The speech level is adjusted at $-26$ dB. The 10 msec frame is used and divided into four subframes of 2.5 msec. The 20th-order spectral parameters are computed using the 32 msec Hamming window centered by the middle of the last subframe. Table 1 shows the bit allocations of wideband CELP coding at 16 kbit/s. The postfilter consists of three filters: a pitch postfilter, a short-term postfilter and a tilt compensation filter.

### 3.1. Spectral parameters

Two-stage vector quantizer with switched fifth-order moving average interframe prediction is used for spectral quantization. The selection of MA predictive coefficients and the first stage use 1 bit and 8 bits, respectively. In the second stage, the vector is split into a lower dimensional part and higher dimensional part, and 6 bits are assigned to each part. The MGC-LSP parameters are quantized with Euclidean distortion measure in the proposed coder, while LSP parameters with weighted Euclidean distortion measure in the conventional CELP coder.

Table 2 shows the quantization performance of the MGC-LSP for 8 Japanese speakers. In the table, the cepstral distortion (CD) measure is used for evaluation. It is shown that the quantization performance of the MGC-LSP is slightly worse than that of the LSP.

### 3.2. Power

The power parameter is calculated on a two-subframe basis, i.e., two-dimensional vector of the power parameter is obtained once a frame. The vector is quantized into 7 bits in the $\mu$-law domain.

### 3.3. Excitation codebooks and gain codebook

The excitation codebook 1 consists of an adaptive codebook and a fixed codebook [11]. The adaptive codebook represents the pitch periodicity, in which a fractional pitch delay is used with resolution: 1/4 in the range 33–96$\frac{3}{4}$, 1/2 in the range 97–160$\frac{1}{2}$ and integers only in the range 161-224. The fixed codebook represents the nonperiodic and nonstational speech, and it stores 64 random codevectors.

The excitation codebook 2 is based on an algebraic codebook structure. In this codebook, each codevector contains four non-zero pulses. Each pulse can have either the amplitude $+1$ or $-1$, and can assume the same positions as the ITU-T G.729 [12].

The gains of the excitation codebook 1 and 2 are vector-quantized using a 7-bit codebook. The gain codebook is trained by the generalized Lloyd algorithm.

Table 3: Speech quality versus $\alpha$ in terms of DMOS.

| $\alpha$ | Average | Female | Male |
|---|---|---|---|
| 0.0 | 3.41 | 3.13 | 3.70 |
| 0.1 | 3.56 | 3.30 | 3.83 |
| 0.2 | 3.72 | 3.58 | 3.86 |
| 0.3 | 4.20 | 4.23 | 4.17 |
| 0.4 | 4.18 | 4.22 | 4.14 |
| 0.5 | 4.01 | 4.09 | 3.92 |

## 4. SUBJECTIVE PERFORMANCE

Subjective tests were carried out in a sound-proof booth. Eight people took part in the tests. They listened to the speech sequences, which are sixteen Japanese sentences spoken by 4 female and 4 male speakers, and gave a rating using a 5-point scale.

### 4.1. Effect of frequency warping

Table 3 shows the speech quality for several values of $\alpha$ in terms of DMOS. It is seen from the table that the quality of $\alpha = 0.3$ and $0.4$ is significantly improved over $\alpha = 0$, especially for female speech. This is mainly due to decreasing of the perceived quantization noise. This result indicates that the frequency warping makes a large contribution to the improvement of the subjective quality.

### 4.2. Comparison with G.722 and conventional CELP

The MGC-CELP coders with $\alpha = 0.3$ and 0.4 are compared with the ITU-T G.722 standard at 48, 56 and 64 kbit/s, and the conventional CELP coder whose framework is the same as the proposed one. By informal listening tests, we chose $A(z/0.9)/A(z/0.6)$ and $A(z/0.65)/A(z/0.75)$ for the perceptual weighting filter and short-term postfilter of the conventional coder, respectively, where $A(z)$ is the LP inverse filter.

Figures 3 and 4 show the results of the ACR and DCR tests, respectively. It is shown that the MGC-CELP coders outperform the G.722 at 64 kbit/s in terms of MOS, and are comparable to the 64 kbit/s G.722 coder in terms of DMOS. Moreover, it is also found that the MGC-CELP coders give a much smaller difference between male and female speakers than the conventional CELP coder.

## 5. CONCLUSIONS

We proposed a wideband CELP coder based on mel-generalized cepstral analysis. The coder encodes fullband of the speech signal through the warped frequency scale. It was shown from the subjective tests that the performance of the proposed coder at 16 kbit/s is better than that of the ITU-T G.722 standard at 64 kbit/s. It was also found that the proposed coder gives a much smaller difference in the performance for male and female speakers than the conventional CELP coder.

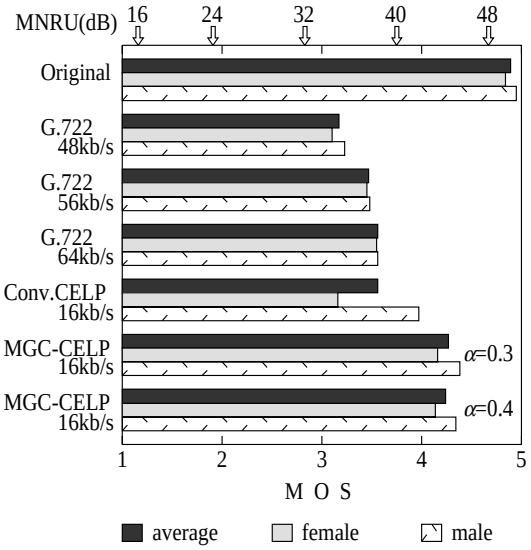



Figure 3: Result of ACR test.

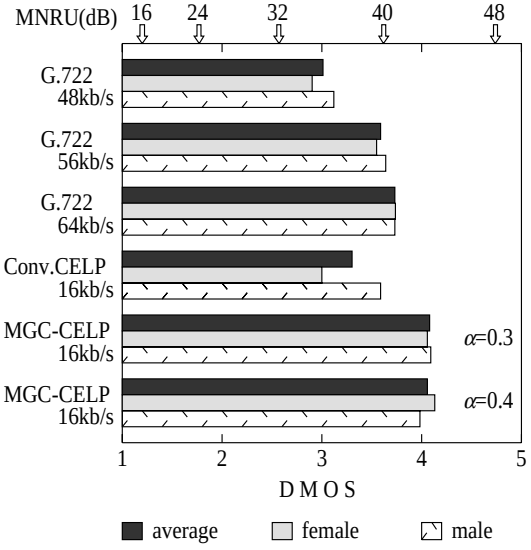



Figure 4: Result of DCR test.


## ACKNOWLEDGMENT

This work is supported in part by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists, and in part by Support for International Research of International Communication Foundation.


## REFERENCES


[1] P. Mermelstein, "A new CCITT coding standard for digital transmission of wideband audio signals," *IEEE Communications Magazine*, vol. 26, no. 1, pp.8–15, Jan. 1988.

[2] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP) : high quality speech at very low bit rates," *Proc. ICASSP'85*, pp.937–940, 1985.

[3] R. Drago, R. Montagna, F. Perosino and D. Sereno, "Some experiments of 7 kHz audio coding at 16kbits/s," *Proc. ICASSP'89*, pp.192–195, 1989.

[4] J. W. Paulus and J. Schnitzler, "16 kbit/s wideband speech coding based on unequal subbands," *Proc. ICASSP'96*, pp.255–258, 1996.

[5] A. Kataoka, S. Kurihara, S. Sasaki and S. Hayashi, "A 16-kbit/s wideband speech codec scalable with G.729," *Proc. EUROSPEECH'97*, pp.1491–1494, 1997.

[6] K. Tokuda, H. Matsumura, T. Kobayashi and S. Imai, "Speech coding based on adaptive mel-cepstral analysis," *Proc. ICASSP'94*, pp.197–200, Apr. 1994.

[7] K. Koishida, K. Tokuda, T. Kobayashi and S. Imai, "CELP coding based on mel-cepstral analysis," *Proc. ICASSP'95*, pp.33–36, 1995.

[8] K. Koishida, K. Tokuda, T. Kobayashi and S. Imai, "CELP coding system based on mel-generalized cepstral analysis," *Proc. ICSLP'96*, pp.318–321, 1996.

[9] K. Tokuda, T. Kobayashi, T. Masuko and S. Imai, "Mel-generalized cepstral analysis — a unified approach to speech spectral estimation," *Proc. ICSLP-94*, pp.1043–1046, 1994.

[10] K. Koishida, K. Tokuda, T. Kobayashi and S. Imai, "Spectral representation of speech using mel-generalized cepstral coefficients," *Proc. 3rd Joint Meeting of ASA and ASJ*, pp.963–968, 1996.

[11] S. Miki, K. Mano, H. Ohmuro and T. Moriya, "Pitch synchronous Innovation CELP (PSI-CELP)," *Proc. EUROSPEECH'93*, pp.261–264, 1993.

[12] R. Salami, C. Laflamme, J-P. Adoul, A. Kataoka, S. Hayashi, C. Lamblin, D. Massaloux, S. Proust, P. Kroon and Y. Shoham, "Description of the proposed ITU-T 8kb/s speech coding standard," *IEEE Speech Coding Workshop*, Annapolis, 1995.