ON IMPROVING THE PERFORMANCE OF CODED SPECTRAL PARAMETERS FOR SPEECH RECOGNITION

Seung Ho Choi[1,2], *Hong Kook Kim*[3], *and Hwang Soo Lee*[1,4]

[1] Dept of Electrical Engineering, KAIST
[2] Human & Computer Interaction Lab., SAIT, Samsung
[3] MMC Technology, Inc.
[4] Central Research Laboratory, SK Telecom
(E-mail: shchoi@green.sait.samsung.co.kr)

ABSTRACT

In digital communication networks, speech recognition systems conventionally reconstruct speech followed by extracting feature parameters. In this paper, we consider a useful approach by incorporating speech coding parameters into the speech recognizer. Most speech coders employed in the networks represent line spectral pairs (LSPs) as spectral parameters. In order to improve the recognition performance of the LSP-based speech recognizer, we introduce two different ways: one is to devise weighed distance measures of LSPs and the other is to transform LSPs into a new feature set, named a pseudo-cepstrum (PCEP). Experiments on speaker-independent connected-digit recognition showed that the weighted distance measures significantly improved the recognition accuracy than the unweighted one of LSPs. Especially we could obtain more improved performance by using PCEP. Compared to the conventional methods employing mel-frequency cepstral coefficients, the proposed methods achieved higher performance in recognition accuracies.

1. INTRODUCTION

Low-bit-rate speech coders are widely used in digital communication networks due to the recent advances in speech coding algorithms and DSP technologies. Many of these coders use linear predictive coding (LPC) parameters such as the line spectrum pair (LSP) frequencies to represent spectral information. If we use these spectral parameters, we can construct an efficient speech recognition system at the transmitters or receivers of digital communication networks.

Recent studies have investigated the degradation of recognition performances when the speech recognizer is used in digital communication networks [1] [2]. The conventional methods have to extract feature parameter for speech recognition using the reconstructed speech signals. Hence, the spectral distortion caused by a speech coder degrade the recognition performance severely.

In this paper, we propose a new speech recognition approach with low-complexity which is applicable to digital communication networks. The proposed method utilizes the quantized spectral parameters of a speech coder so that it does not have to reconstruct speech signals and to extract a baseline feature parameter. In order to increase the recognition accuracies of the proposed methods, two different methods are introduced. One is applying some weighting functions to LSPs. The other is converting LSPs to a new feature, pseudo-cepstrum (PCEP), by approximating the relationship between cepstrum and LSPs.

The paper is organized as follows. Section 2 describes weighting functions applied to LSP frequencies for the purpose of improving the recognition accuracy. Section 3 describes the extraction procedure of PCEP. Section 4 shows experimental results, and Section 5 summarizes the main contributions of the paper.

2. WEIGHTED LSP DISTANCE MEASURES FOR SPEECH RECOGNITION

In this section, weighting functions of LSP distance measure are introduced to increase the recognition accuracy. The weighting functions can be classified into three categories, which are (1) spectral weighting function, (2) frequency weighting function, and (3) hybrid weighting function.

These weighting functions are incorporated into a weighted Euclidean distance measure of the form

$$D^2(\omega,\hat{\omega}) = (\omega-\hat{\omega})^T W(\omega-\hat{\omega}), \tag{1}$$

where ω and $\hat{\omega}$ are the reference and test LSP vectors, respectively, and W is a diagonal weighting matrix which depends on ω and $\hat{\omega}$.

2.1. Spectral weighting functions

Paliwal and Atal [3] have proposed a weighted Euclidean distance measure using the localized spectral sensitivity property of the LSPs (referred to as LPCSW). In this measure, the weights assigned to an LSP vector of order 10 are given by

$$w_i^{LPCSW} = \{s_i[P(\omega_i)]^{0.15}\}^2, 1 \le i \le 10, \tag{2}$$

where $P(\omega_i)$ is the LPC power spectrum at the i^{th} LSP frequency ω_i. The scaling factors $\{s_i\}$ are used to decrease the sensitivity of the higher LSPs and have fixed values as

$$s_i = \begin{cases} 1, & 1 \le i \le 8 \\ 0.8, & \text{i=9} \\ 0.4, & \text{i=10.} \end{cases} \tag{3}$$

When adjacent LSPs come close, the speech spectrum has a peak near these frequencies. Therefore, a parameter that is close to one of its neighbors has a high spectral sensitivity and should be given a higher weight. From this point, Laroia *et al.*[4] have

defined the inverse harmonic mean weighting function (referred to as IHMW) as

$$w_i^{IHMW} = s_i^2\left(\frac{1}{\omega_i - \omega_{i-1}} + \frac{1}{\omega_{i+1} - \omega_i}\right), 1 \le i \le 10, \quad (4)$$

where $\omega_0 = 0$ and $\omega_{11} = \pi$.

In [5], the authors have shown that the spectral distortion is equal to the weighted Euclidean distance for small distances and they have obtained a weighting function by spectral sensitivity (referred to as SSW). The i^{th} element of the diagonal spectral sensitivity matrix is computed as

$$d_i = \frac{1}{\pi \Delta \omega_i^2} \int_0^{\pi} [10\log_{10}|P(\omega)| - 10\log_{10}|P(\tilde{\omega})|]^2 d\omega, \quad (5)$$

where $\tilde{\omega} = \omega + \Delta\omega$. From (5), the SSW used in this paper is defined as

$$w_i^{SSW} = s_i^2 d_i, 1 \le i \le 10. \quad (6)$$

2.2. Frequency weighting function

Speech recognition systems increase recognition performance by employing the auditory spectral representation as a front-end [6]. As a way to incorporate an auditory model into the LSP distance measure, a frequency weighting function is proposed which gives more weight to the low order LSPs than to the high order ones.

The mel-scale frequency, f_m, is converted from the linear-scale frequency, f, as follows:

$$f_m = f + 2\tan^{-1}\frac{a\sin f}{1 - a\cos f}, \quad (7)$$

where a controls the degree of frequency warping and a is set to 0.47. Instead of the mel-scale conversion of LSPs by using (7), we propose a mel-scale frequency weighting function (referred to as MFW) which is defined as

$$w_i^{MFW} = (1 + \frac{2}{\omega_i}\tan^{-1}\frac{a\sin\omega_i}{1 - a\cos\omega_i})^2, 1 \le i \le 10. \quad (8)$$

2.3. Hybrid weighting functions

Spectral peaks in the low frequency are more important for speech recognition than those in the high frequency. This is not considered in the spectral weighting functions in Section 2.1. Hence, we propose a hybrid weighting function which is obtained from the spectral weighting function multiplied by the MFW.

3. PSEUDO-CEPSTUM

The PCEP is obtained by approximating the relationship between the cepstrum and LSPs. An inverse filter of order p whose roots are inside the unit circle is defined as $A_p(z) = \sum_{k=0}^{p} a_k z^{-k}$, where $a_0 = 1$ and $\{a_k\}$ are LPC coefficients of order p. Generally the LSPs of order p are defined as the complex roots of the following polynomials $P(z)$ and $Q(z)$ [7].

$$P(z) = A_p(z) + z^{-(p+1)}A_p(z^{-1}), \quad (9)$$

$$Q(z) = A_p(z) - z^{-(p+1)}A_p(z^{-1}). \quad (10)$$

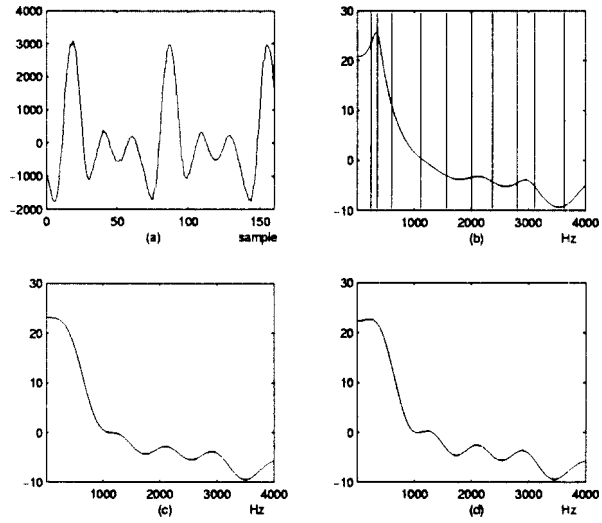

Figure 1: Comparison of the logarithmic spectra associated with LSPs corresponding to (a) a speech segment, obtained from (b) LPC, (c) cepstrum, and (d) PCEP.

To derive the relationship between the cepstrum and LSPs, we obtain the following equation by multiplying (9) and (10)

$$P(z)Q(z) = A_p^2(z)[1 - R^2(z)]$$

$$= (1 - z^{-2})\prod_{i=1}^{p}(1 - e^{j\omega_i}z^{-1})(1 - e^{-j\omega_i}z^{-1}), \quad (11)$$

where $\{\omega_i\}$ are LSPs of order p and $R(z) = z^{-(p+1)}\frac{A_p(z^{-1})}{A_p(z)}$. Taking the logarithm on both sides of (11) gives

$$2\log A_p(z) + \log(1 - R^2(z)) = \log(1 - z^{-2})$$

$$+ \sum_{i=1}^{p}[\log(1 - e^{j\omega_i}z^{-1}) + \log(1 - e^{-j\omega_i}z^{-1})]. \quad (12)$$

Since the right-hand side of (12) has zeros on the unit circle, the zeros should be shifted radially by the factor $\alpha(0 < \alpha < 1)$ [8]. By taking the inverse Z-transform and using $\log A_p(z) = -\sum_{n=1}^{\infty} c_n e^{j\omega n}$, where c_n is the n^{th} LPC cepstral coefficient, we obtain the relationship between the LPC cepstrum and the LSPs as follows.

$$c_n = \frac{1}{2n}(1 + (-1)^n) + \frac{1}{n}\sum_{i=1}^{p}\cos n\omega_i + R_n, \quad n \ge 1, \quad (13)$$

where

$$R_n = \sum_{k=1}^{\infty}\frac{\alpha^{2k(p+1)-n}}{4\pi k}\int_{-\pi}^{\pi}\cos[(2k(p+1) + n)\omega$$

$$-4k\sum_{l=1}^{\infty}c_l\alpha^l\sin\omega l]d\omega.$$

The mid-term of the right-hand side of (13) gives the root-power-sum of the filter $\hat{A}(z) = 1/\prod_{k=1}^{p}(1 - e^{j\omega_k}z^{-1})(1 - e^{-j\omega_k}z^{-1})$

Table 1: Recognition rates of MFCC.

Condition		Recognition Rate(%)
Training	Test	
Original	*Original*	87.0
Original	*Reconstructed*	78.7
Reconstructed	*Reconstructed*	83.6

Table 2: Coding effects on recognition rates.

Training	Test	Recognition Rate (%)
LSP	LSP	83.6
LSP	QLSP	83.0
QLSP	QLSP	82.6
RLSP	RLSP	78.1

divided by the quefrency n. The PCEP $\{\hat{c}_n\}$ is defined as

$$\hat{c}_n = \frac{1}{n}\sum_{i=1}^{p} \cos n\omega_i, n \geq 1. \quad (14)$$

Fig. 1 shows the logarithmic LPC spectrum of a speech segment and the corresponding smoothed spectrum obtained from cepstrum and PCEP. As shown in the figure, the PCEP gives similar spectral envelope to the cepstrum.

4. RECOGNITION EXPERIMENTS

We have constructed a recognition system based on the Qualcomm code-excited linear predictive coder (QCELP) [9] for a feasibility test of the proposed methods. The spectral envelope in the QCELP is represented by LSPs of order 10, which are updated every 20 msec frame, and each frequency is scalar quantized.

To evaluate the recognition performances of the proposed methods, a connected digit database was used. Utterances from 93 speakers were used as training data and those from the other 47 speakers were used as test data. Each speaker pronounced 40 digit strings generated randomly with lengths varying from three to seven. All the feature vectors used in this work were concatenated with the corresponding time derivative vector, resulting in a 20-dimensional observation vector. In all the experiments, a left-to-right five state hidden Markov model (HMM) with discrete observation density is used for modeling each digit where the codebook size was 256 for both the static and delta feature vector. The HMM parameters were estimated by five iterations of the segmental K-means algorithm. We performed speech recognition experiments using various feature sets converted from the quantized LSPs. We compared the recognition results under various training and test conditions.

First, we performed a recognition experiment using a conventional feature vector to investigate the effects of distortion caused by the speech coder on the speech recognition accuracy. A 19-th order mel-scaled log filterbank energy vector was extracted for each 20 msec frame. By applying the discrete cosine transform, a mel-frequency cepstral coefficient (MFCC) vector of order 10 was derived. The recognition results of the MFCC are shown in Table 4, where *Original* and *Reconstructed* means that the feature vector is obtained by using the original and reconstructed speech signals, respectively. The results show that the mismatch between training and test conditions degrades the recognition accuracy severely.

Table 3: Average spectral distances (SD) and recognition rates.

Training	Test	SD (dB)	Recognition Rate (%)
LSP	QLSP	1.17	83.0
LSP	RLSP	2.79	75.0

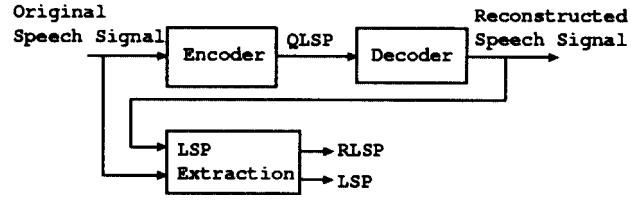

Figure 2: LSPs used in the experiments

Next, we examined the effects of LSP quantization and speech decoding on the speech recognition accuracy. In Table 2, QLSP means quantized LSP, and RLSP means LSP which is obtained from the reconstructed speech signals as depicted in Fig. 2. As shown in the table, the quantization of LSP parameters results in degraded recognition accuracy, but this degradation is much less than that of the RLSP.

To examine how much the spectral distortions have an effect on the recognition accuracies, we computed the average spectral distance (SD) [3], which is defined as

$$SD = \{\frac{1}{N_f}\sum_{n=1}^{N_f}(\frac{100}{\pi}\int_0^{\pi}[\log_{10}|P_n(\omega)|^2 - \log_{10}|\hat{P}_n(\omega)|^2]^2 d\omega)\}^{\frac{1}{2}} \quad (15)$$

where P_n and $\hat{P}_n$ represent the LPC magnitude spectra of the reference and test speech segment, respectively, and N_f is the total number of frames. As shown in Table 3, the SD value between LSP and RLSP is more than that caused by the LSP quantization. This proves that the SD values are closely related to the recognition accuracies.

To improve the performance of the recognizer, the weighting functions described in Section 2 are assigned to the QLSPs of the speech coder. Table 4 shows the recognition accuracies resulted from the different weighting functions. It can be seen that all the weighted distance measures improve the recognition accuracy compared with the unweighted ones. Among them, LPCSW shows the best result. The table also shows that the hybrid weighting functions always give better performance than the spectral weighting functions do.

To evaluate the recognition performance of the PCEP introduced in Section 3, we compared it with the cepstrum obtained

Table 4: Recognition rates for weighted LSP distance measures.

Weighting function	Recognition Rate (%)
Unweighted	82.6
LPCSW	84.8
IHMW	84.5
SSW	83.9
MFW	84.5
LPCSW + MFW	85.5
IHMW + MFW	84.8
SSW + MFW	84.4

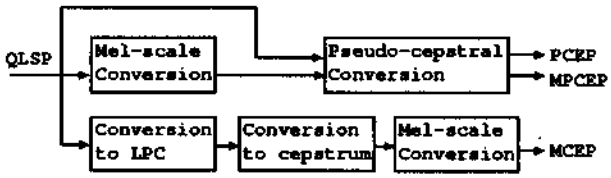

Figure 3: Block diagram for extracting cepstrum and PCEP from QLSP.

from QLSP as depicted in Fig. 3. As shown in Table 5, the mel-scale PCEP (MPCEP) gives the same performance as mel-scale LPC cepstrum (MCEP) does. However, the MPCEP saves the computational cost considerably compared to the MCEP since the MPCEP does not need the conversion of LSP to LPC.

5. CONCLUSIONS

We introduced a useful approach which incorporates speech coding parameters into the speech recognizer. We showed that large spectral distortion degrades speech recognition performance severely. To improve the recognizer using the quantized LSPs, first we proposed weighting functions: spectral weighting function, frequency weighting function, and hybrid weighting function. The weighted Euclidean distance measures increased the recognition performance compared with the unweighted one. The other is converting LSP to a new feature, pseudo-cepstrum, by approximating the relationship between LSP and cepstrum. The mel-scale pseudo-cepstrum (MPCEP) showed comparable performance to the mel-scale cepstrum (MCEP), while the MPCEP requires much less computations than MCEP does. We conclude that the proposed recognition methods can be efficiently used in a coder-based speech recognizer by incorporating the spectral parameters of the speech coder.

6. REFERENCES

[1] S. Euler and J. Zinke, "The influence of speech coding algorithms on automatic speech recognition," in *Proc. of ICASSP*, pp. 621-624, 1994.

[2] B. T. Lilly and K. K. Paliwal, "Effect of speech coders on speech recognition performance," in *Proc. of ICSLP*, pp. 2344-2347, 1996.

Table 5: Recognition rates of PCEP, MPCEP, and MCEP which are obtained from QLSP.

Feature	Recognition Rate (%)
PCEP	83.6
MPCEP	86.3
MCEP	86.3

[3] K. K. Paliwal and B. S. Atal "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. on Speech and Audio Process.*, vol. 1, no. 1, pp. 3-14, Jan. 1993.

[4] R. Laroia, N. Phamdo, and N. Farvardin, "Robust and efficient quantization of speech LSP parameters using structured vector quantizers," in *Proc. ICASSP*, pp. 641-644, 1991.

[5] W. R. Gardner and B. D. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 367-381, Sept. 1995.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp.1738-1752, 1990.

[7] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Am.*, (abstract) vol. 57, p.535, 1975.

[8] A. V. Oppenheim and R. W. Shafer, *Discrete-time signal processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[9] "Speech option standard for wideband spread spectrum digital cellular system," *Qualcomm Inc, TIA/EIA Interim Standard-96*, Apr. 1993.