LSP WEIGHTING FUNCTIONS BASED ON SPECTRAL SENSITIVITY AND MEL-FREQUENCY WARPING FOR SPEECH RECOGNITION IN DIGITAL COMMUNICATION

Seung Ho Choi[*†], *Hong Kook Kim*[‡], *and Hwang Soo Lee*[*§]

[*] Dept. of Electrical Engineering, KAIST, Taejon 305-701, Korea
[†] Samsung Advanced Institute of Technology, Kyungki-Do 449-712, Korea
[‡] currently at AT&T Labs Research, 180 Park Avenue, Florham Park NJ 07932, USA
work performed while author was at MMC Technology, Inc., Seoul, Korea
[§] Central Research Laboratory, SK Telecom, Taejon 305-348, Korea
E-mail: [*] shchoi@spectra.kaist.ac.kr, [‡] hkkim@research.att.com, [§] hslee@sktelecom.re.kr

ABSTRACT

In digital communication networks, a speech recognition system extracts feature parameters after reconstructing speech signals. In this paper, we consider a useful approach of incorporating speech coding parameters into a speech recognizer. Most speech coders employ line spectrum pairs (LSPs) to represent spectral parameters. We introduce weighted distance measures to improve the recognition performance of an LSP-based speech recognizer. Experiments on speaker-independent connected-digit recognition showed that weighted distance measures provide better recognition accuracy than unweighted distance measures do. Compared with a conventional method employing mel-frequency cepstral coefficients, the proposed method achieved higher performance in terms of a recognition accuracy.

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 line spectrum pairs (LSPs) to represent spectral information. In order to construct an efficient speech recognition system at the transmitter or receiver of digital communication networks, directly using these spectral parameters is preferable to reconstructing speech signals and then extracting feature parameters.

In recent studies, the degradation of recognition performance in digital communication networks has been investigated [1] [2]. They extracted feature parameters by using reconstructed speech signals. However, the characteristic of feature parameters was far from that obtaining by original speech signals because a speech coder generates a large spectral distortion. Hence, the performance of a recognition system is degraded severely.

In this paper, we propose a new approach that utilizes the transmitted spectral parameters of a speech coder directly. In contrast with conventional methods, the low-complexity recognizer can be implemented with higher recognition performance. For this, we introduce weighted distance measures to an LSP-based speech recognizer.

This paper is organized as follows: In Section 2, several weighting functions are proposed for improving recognition accuracy. Section 3 shows the recognition performance of each weighted distance measure, and Section 4 summarizes the main contributions of this paper.

2. WEIGHTED DISTANCE MEASURES FOR SPEECH RECOGNITION

In this section, weighting functions to be used in the weighted distance measure are investigated. They can be classified into three categories; (1) spectral sensitivity-based weighting function, (2) mel-frequency warping-based weighting function, and (3) hybrid weighting function. These weighting functions are incorporated into the distance measure as follows:

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

where ω^r and ω^t are the reference and the test LSP vectors of order p, respectively, and W is a p-by-p weighting matrix that is decided by ω^r and ω^t. The W is represented as $W^r + W^t$, where W^r and W^t are weighting matrices obtained by ω^r and ω^t, respectively. To implement a speech recognition system with low-complexity, W should be a diagonal matrix and (1) be a weighted Euclidean distance measure. In other words, the i^{th} diagonal element of W is given by

$$w_i = w_i^r + w_i^t, 1 \le i \le 10, \tag{2}$$

where w_i^r and w_i^t are the i^{th} diagonal elements of W^r and W^t, respectively.

2.1. Spectral sensitivity-based weighting functions

Some weighting functions have been successfully applied to quantize LSPs in speech coding areas [3]. The (1) comes close to an accurate spectral distortion if we choose a proper W. Therefore, when they are applied to a speech recognition system, performance could be improved. We first investigate the usefulness of weighting functions used in a speech coding in view of speech recognition, and several weighting functions are proposed to further improve the accuracy of speech recognition system.

The localized spectral sensitivity property of LSPs can be used to obtain a weighted Euclidean distance measure [4]. In this measure (referred to as LPCW), the assigned weighting values to 10 LSPs are

$$w_i^{LPCW} = s_i^2[P(\omega_i)]^{0.3}, 1 \le i \le 10, \quad (3)$$

where $P(\omega_i)$ is a power spectrum of LPC at the i^{th} LSP, ω_i. The scaling factor, s_i, decreases the sensitivity of higher LSPs and has a fixed value as

$$s_i = \begin{cases} 1, & 1 \le i \le 8 \\ 0.8, & \text{i=9} \\ 0.4, & \text{i=10.} \end{cases} \quad (4)$$

In practice, LPCW requires heavy computations for calculating $P(\omega_i)$.

The LSPs has a following property: When adjacent LSPs come close, the speech spectrum has a peak near these frequencies. Therefore, an LSP that is close to one of its neighbors has a high spectral sensitivity, so a higher weighting value should be assigned to the LSP [5]. From this point, the inverse harmonic mean weighting function (referred to as IHMW) is defined with the scaling factor in (4) as

$$w_i^{IHMW} = s_i^2(\frac{1}{\omega_i - \omega_{i-1}} + \frac{1}{\omega_{i+1} - \omega_i}), 1 \le i \le 10, \quad (5)$$

where $\omega_0 = 0$ and $\omega_{11} = \pi$. The computational burden of IHMW is less than that of LPCW while IHMW gives approximately equal spectral sensitivity to LPCW. However, both the LPCW and the IHMW were designed on the basis of a heuristic sense that uses only the localized sensitivity properties of LSPs.

On the other hand, the authors in [6] derived the sensitivity matrix of LSPs from a quantization theory. They showed that a spectral distortion is equal to a weighted Euclidean distance if the distance is small. When $\mathbf{R}_A$ is the autocorrelation function of the impulse response of an LPC filter $1/A(z)$, a sensitivity matrix is given by

$$D_\omega(\omega) = 4\beta \mathbf{J}_\omega^T(\omega)\mathbf{R}_A\mathbf{J}_\omega(\omega), \quad (6)$$

where $\mathbf{J}_\omega(\omega)$ is an Jacobian matrix obtained by transforming LSPs to LPC coefficients, and β is a constant term. Based on (6), we define a weighting function, named Garner weighting function (GW) as

$$w_i^{GW} = s_i^2 d_i, 1 \le i \le 10, \quad (7)$$

where d_i is the i^{th} diagonal element of $D_\omega(\omega)$.

2.2. Mel-frequency warping-based weighting function

The human auditory system is more sensitive in low frequency ranges than high ranges. From this fact, spectral representations and matching measures were studied to improve the performance of recognizer [7]. As a way to incorporate these auditory characteristics into (1), a weighting function is proposed on the basis of mel-frequency warping, which gives more weighting values to the low order LSPs than to the high order ones.

The relationship between a mel-scale frequency, f^M, and a linear-scale frequency, f, is as follows:

$$f^M = f + 2\tan^{-1}\frac{a\sin f}{1 - a\cos f}, \quad (8)$$

where a controls the degree of frequency warping and is set to 0.45. Instead of the mel-frequency warping of LSPs (MLSPs) [8] by using (8), the proposed mel-frequency warping-based weighting function (referred to as MFW) is defined as

$$w_i^{MFW} = (1 + \frac{2}{\omega_i}\tan^{-1}\frac{a\sin\omega_i}{1 - a\cos\omega_i})^2, 1 \le i \le 10. \quad (9)$$

The weighted Euclidean distance obtained by this weighting function is approximately equal to the Euclidean distance of MLSPs.

2.3. Hybrid weighting functions

Spectral peaks in low frequency regions are more important than those in high frequency regions in view of speech recognition. This fact was not considered in the spectral sensitivity-based weighting functions in Section 2.1 except for the term s_i^2 in (3), (5) and (7). Greater improvement in recognition performance should be achieved by combining the spectral sensitivity-based weighting functions with MFW. Taking both the spectral sensitivity of LSPs and the human auditory characteristics into account, we introduce three hybrid weighting functions that are obtained by replacing the term s_i^2 of each spectral sensitivity-based weighting function with MFW. They are expressed as follows:

$$\begin{aligned} w_i^{LPCW\,with\,MFW} &= [P(\omega_i)]^{0.3} w_i^{MFW}, \\ w_i^{IHMW\,with\,MFW} &= (\frac{1}{\omega_i - \omega_{i-1}} + \frac{1}{\omega_{i+1} - \omega_i}) w_i^{MFW}, \\ w_i^{GW\,with\,MFW} &= d_i w_i^{MFW}. \end{aligned} \quad (10)$$

Fig. 1 shows 10 LSPs and the mean-normalized weighting values of the weighting functions. The weighting patterns of LPCW and IHMW are similar and GW gives large weights to the spectral peaks without regard to their frequencies, unlike other weight functions. The weighting values of low order LSPs obtained by the hybrid weighting functions are higher than those obtained by the spectral sensitivity-based weighting functions.

3. RECOGNITION EXPERIMENTS

We constructed a recognition system based on the Qualcomm code-excited linear predictive coder (QCELP) [9] for

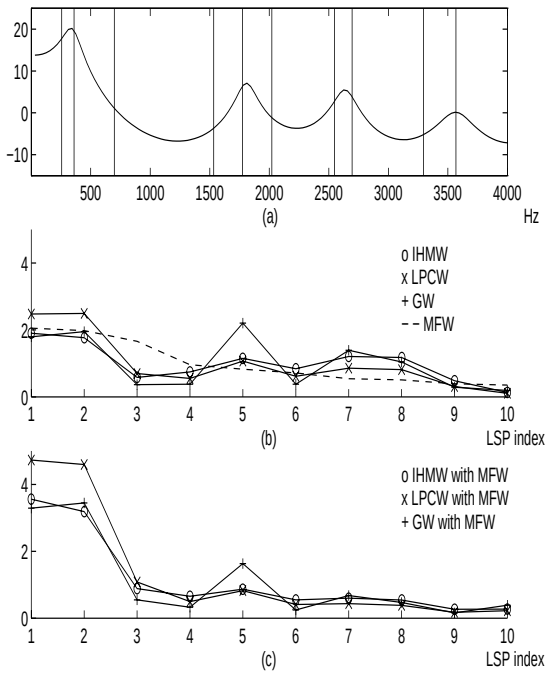

Figure 1. Example of (a) LSPs and corresponding logarithmic spectrum, and (b) the weighting values associated with the LSPs, and (c) the weighting values of hybrid weighting functions.

a feasibility test of the proposed methods. The spectral envelope in the QCELP is represented by 10 LSPs, and it is updated every 20 msec frame. A scalar quantizer is used for every coefficient.

To evaluate the recognition performance of the proposed methods, a connected Korean digit database was used, which was composed of 140 male and female speakers. 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 that are randomly generated with the number of digits varying from three to seven. All the feature vectors were concatenated with a 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 was used for modeling each digit with the codebook size of 256 for both the static and the delta feature vector. The HMM parameters were estimated by the segmental K-means algorithm.

First, we performed a recognition experiment by using a conventional feature vector to investigate the effects of distortion caused by a speech coder. A 19^{th} order mel-scale log filterbank energy vector was extracted with every 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 1. The results show that the speech coder gives an adverse effect to speech recognition. Especially, the accuracy degrades severely in case the condition of training and test is mismatched.

Table 1. Recognition rates of MFCC, where *Original* and *Reconstructed* mean that the feature vector is obtained by using the original and reconstructed speech signals, respectively.

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

Next, we examined the effect of recognition accuracy due to LSP quantization and speech decoding. In Table 2, QLSP means a quantized LSP, and RLSP means an LSP that is obtained by the reconstructed speech signals as depicted in Fig. 2. As shown in the table, recognition accuracy is dropped by 1.0 % when we used a quantized LSP, but the amount of degradation is much less than that of the RLSP. To examine the relationship between the degradation of recognition performance and the degree of spectral distortion, we computed the average spectral distance (SD) [4] using the training data. The SD is defined as $SD = \{\frac{1}{N}\sum_{n=1}^{N}(\frac{100}{\pi}\int_0^{\pi}[\log_{10}|P_n^r(\omega)| - \log_{10}|P_n^t(\omega)|]^2 d\omega)\}^{\frac{1}{2}}$, where $P_n^r(\omega)$ and $P_n^t(\omega)$ represent the LPC power spectra of the n^{th} reference and test speech segments, respectively, and N is the total number of frames. As shown in Table 3, the SD value between LSP and RLSP is much higher than that of the LSP quantization. This proves that the SD values are closely related to a recognition accuracy. In other words, recognition performance can be improved by lowering an SD. It was shown in [2] that the performance of speech recognizer was slightly degraded if it used a higher bit-rate speech coder.

To improve the performance of the recognizer for a given speech coder, the weighting functions described in Section 2 are assigned to the QLSPs of the speech coder. Table 4 shows recognition accuracies resulting from different weighting functions. All the weighted distance measures show greater recognition accuracy than the unweighted ones do. The recognition performance of the weighted distance obtained by MFW is approximately equal to that of the Euclidean distance of MLSPs. Among the spectral sensitivity-based weighting functions, LPCW shows the best result. The table also shows that the hybrid weighting functions always give better performance than the spectral sensitivity-based weighting functions do. As shown in Tables 1 and 4, LPCW with MFW reduces the errors by 11.2 % and 16.7 % compared with MFCC and LSP-Unweighted.

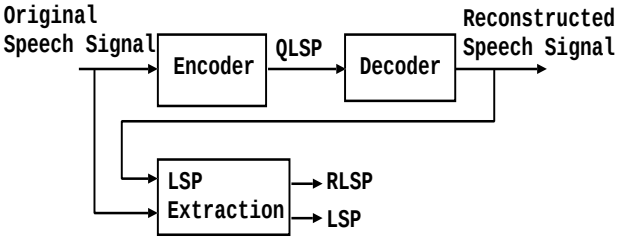

Figure 2. Block diagram for extracting LSP, QLSP, and RLSP.

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

4. CONCLUSIONS

We introduced a useful approach that incorporates coded spectral parameters into a speech recognizer in a digital communication system. To improve the performance of a speech recognizer which uses the quantized LSPs, we first introduced weighted distance measures based on spectral sensitivity. Also, we proposed a mel-frequency warping-based weighting function and hybrid weighting functions that adopt the characteristics of spectral sensitivity and mel-frequency warping function. The hybrid weighting functions showed the better performance than other weighting functions. We can conclude that the performance of a speech recognizer worked in digital communication channels can be greatly improved by directly using transmitted LSPs with a weighting function.

ACKNOWLEDGMENTS

The authors would like to thank Dr. William R. Gardner, Qualcomm Inc. and Dr. Hai Le Vu, Technical University of Budapest for their valuable discussions about Gardner's sensitivity matrix.

REFERENCES

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

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

[3] H. L. Vu and L. Lois, "A new general distance measure for quantization of LSF and their transformed coefficients," in *Proc. ICASSP*, pp. 45-48, 1998.

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

Table 4. Comparison of recognition rates of weighted distance measures.

Weighting Function	Recognition Rate (%)
LSP-Unweighted	82.6
LPCW	84.8
IHMW	84.5
GW	83.9
MLSP-Unweighted	84.4
MFW	84.5
LPCW with MFW	85.5
IHMW with MFW	85.3
GW with MFW	85.3

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

[5] 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.

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

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

[8] F. S. Gurgen, S. Sagayama, and S. Furui, "Line spectrum frequency-based distance measures for speech recognition," in *Proc. ICSLP*, pp. 521-524, 1990.

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