FORMANT WEIGHTED CEPSTRAL FEATURE FOR LSP-BASED SPEECH RECOGNITION

Ho Young Hur and Hyung Soon Kim

Dept. of Electronics Engineering, Pusan National University, Pusan, Korea
E-mail: {hyher, kimhs}@hyowon.cc.pusan.ac.kr

ABSTRACT

In this paper, we propose a formant weighted cepstral feature for LSP-based speech recognition system. The proposed weighting scheme is based on the well-known property of LSPs that the speech spectrum has a peak when adjacent LSFs come close. By applying this scheme to pseudo-cepstrum (PCEP) conversion process [1], we can obtain formant weighted or peak enhanced cepstral feature. Results of speech recognition experiments using QCELP coder output show that the proposed feature set outperforms the conventional features such as LSP or PCEP. Moreover its performance also exceeds that of unquantized LPC cepstrum.

1. INTRODUCTION

Low bit rate speech coders are widely used in digital communication networks. Many of these coders use linear predictive coding (LPC) parameters in the form of the line spectrum pair (LSP) frequencies to represent vocal tract information. Although these coders are optimized for some perceptually related criterion, they introduce distortion into the speech signal. It is natural to expect that the recognition performance of these systems will deteriorate with the reduction in the bit rate [2].

Several researchers have addressed the problems of speech recognition systems in digital mobile communication environments. Salonidis showed that the robustness against tandeming could be achieved by using a feature compensation and model adaptation algorithm [3]. Laroria showed that the performance of a LSP-based speech recognition system can be improved by using an inverse harmonic mean weighted distance measure [4]. Kim improved recognition performance by introducing a simplified LSP-to-cepstrum conversion method [1]. Although these methods showed superior performance to the conventional method that uses reconstructed speech signals, recognition accuracy was still worse than that of un-encoded speech signals.

In this paper, we propose a formant weighted pseudo-cepstral feature set which is based on the property of LSPs that speech spectrum has a peak when adjacent LSPs come close. By emphasizing formant region, the proposed method outperforms even the unquantized LPC cepsrum (LPCC).

In order to evaluate the proposed method, we constructed a recognition system based on the Qualcomm code-excited linear predictive coder (QCELP) [5] and performed speech recognition experiments using LSP, PCEP, LPCC and proposed formant weighted pseudo-cepstrum.

2. FORMANT WEIGHTING METHOD FOR PSEUDO-CEPSTRUM

LSP frequencies are widely used in the speech processing areas including speech coding, synthesis, and recognition. In speech recognition research based on the LSP frequencies, it was found that the weighted LSP distance measures show better performance than unweighted distance. It was also found that pseudo-cepstrum (PCEP) directly converted from LSP frequencies is an effective feature for speech recognition [1].

In this section, we present the use of weighting function on cepstrum, based on the property that speech spectrum has a peak when adjacent LSPs come close.

2.1. Root power sum fomula of cepstrum

The all-pole model based on LP analysis is widely used in speech signal processing. An all-pole filter of order p is given by

$$S(z)=\frac{1}{A(z)}=\frac{1}{1+\sum_{k=1}^{p}a_k z^{-k}}=\frac{1}{\prod_{k=1}^{p}(1-z_k z^{-1})} \quad (1)$$

with roots z_k, k=1, 2, …, p of the model. Schroeder [7] expressed the cepstral coefficients $c(n)$ as a root power sum formula

$$c(n)=\frac{1}{n}\sum_{k=1}^{p}z_k^n=\frac{1}{n}\sum_{k=1}^{p}e^{-n(B_k+jw_k)} \quad (2)$$

where B_k and w_k is bandwidth and center frequency of root z_k, respectively. Thus the n^{th} cepstral coefficient can be interpreted as a nonlinear transformation of the components center frequencies and weighting according to their bandwidths. It was reported that the relationship between the cepstrum and the spectral components could be used to improve accuracy of speaker recognizer [8]. Inspired by the above facts, we propose the formant weighted pseudo-cepstrum feature in this paper.

2.2. Formant weighting function using LSP

There are several weighting functions that have been successfully applied to the weighted LSP distance measure to quantize LSPs for speech coding application [4][9][10]. Among them inverse harmonic mean weighting (IHMW) function is preferred in many cases because of its low computational burden [4]. The rationale behind the IHMW function is that a speech spectrum has a strong resonance at the formant frequency where adjacent LSPs are close together. Hence, a LSP that is close to one of its neighboring LSPs has high spectral sensitivity, and a large weighting value should be assigned to the LSP. From this viewpoint, the IHMW function is defined by

$$W_k=\left(\frac{1}{\omega_k-\omega_{k-1}}+\frac{1}{\omega_{k+1}-\omega_k}\right),\quad k=1,\cdots,p \quad (3)$$

where ω_k is k^{th} LSP frequency with $\omega_0=0$ and $\omega_{p+1}=\pi$.

2.3. Formant Weighted Pseudo-Cepstrum

When the additive noise contaminate the speech signal, the bandwidths of speech peaks are broadened and LSFs are set apart. Thus, it is desirable to sharpen the bandwidth of spectral peaks of spectrum by some form of formant weighting.

We apply the above concept to the pseudo-cepstrum (PCEP) feature, which is known to be an effective LSP-based feature for speech recognition. The PCEP is obtained by approximating the relationship between the cepstrum and LSPs [1], given by

$$PCEP(n)=\frac{1}{n}\sum_{k=1}^{p}\cos\ n\omega_k \quad (4)$$

where ω_k is k^{th} LSP frequency.

By comparing Equation (4) and (2), we can interpret the n^{th} pseudo-cepstral coefficient as a nonlinear transformation of the LSFs and zero-bandwidth weighting. That is, it can be thought that poles of all pole filter are located on unit circle with center frequency of ω_k, even though LSP frequency ω_k is not the center frequency of the pole.

While previous research showed that the PCEP has a better recognition performance than LSPs, we could improve the recognition accuracy further by introducing formant weighted pseudo-cepstrum (FW-PCEP) as follows :

$$FW\ PCEP(n)=\frac{1}{n}\sum_{k=1}^{p}e^{n\beta_k}\cos\ n\omega_k \quad (5)$$

where $\beta_k=\alpha W_k$ is formant weighting function, α is a normalization factor which should be determined empirically, and W_k is the IHMW function described in Equation (3).

The FW-PCEP has an effect of narrowing the bandwidths of the narrow-band poles while their center frequencies are unchanged. The bandwidth narrowing can be carried out by selectively narrowing the pole bandwidths. In selective bandwidth narrowing, the degree of bandwidth narrowing is

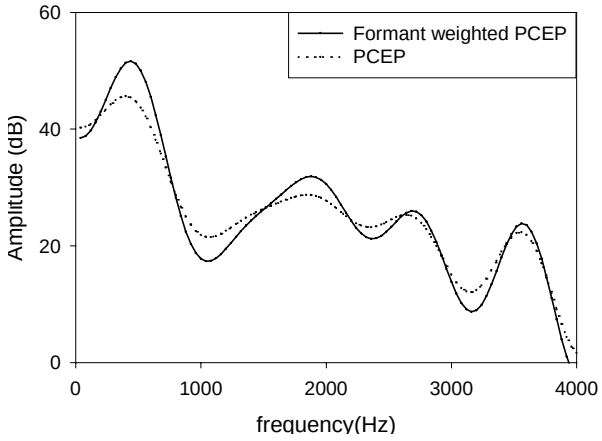

Fig. 1. Effect of formant weighting on a LPC spectrum.

controlled by the IHMW function of each LSP.

Fig. 1. shows the effect of formant weighting function. This figure is a spectrum of FW-PCEP and PCEP from the vowel sound /e/ uttered in clean environment. As it can be seen in the figure, the FW-PCEP has shrper peaks than PCEP spectrum.

3. RECOGNITION EXPERIMENTS

To evaluate the proposed method, we constructed a recognition system based on the Qualcomm code-excited linear predictive coder (QCELP) [5]. The spectral envelope in the QCELP is represented by ten LSPs that are updated every 20ms frame and each LSP is scalar-quantized. The database used for the recognition experiment consists of 28 isolated Korean words that were spoken 3 times by each of 9 speakers. The database was recorded in an ordinary office environment. To simulate the noisy environment, a zero mean white Gaussian noise was added to the test utterances. In all the experiments, DTW-based speech recognition system with the Sakoe-Chiba path constraint was used.

The various feature sets for the recognition experiments are illustrated in Fig. 2, where the solid lines show the proposed feature extraction method.

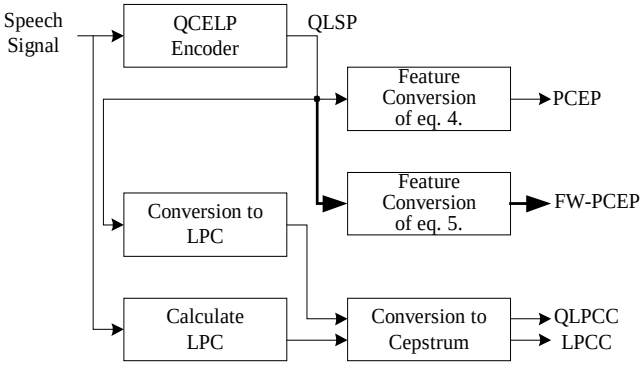

Fig. 2. Block diagram for several feature extraction methods

First, we performed a recognition experiment using a LPCC to investigate the effects of distortion caused by the speech coder on the speech recognition accracy. The recognition results with unquantized and quantized LPCC are shown in Table 1. Although the recognition accuracies are similar to each other when clean speech is used, some degradation of accuracy can be seen when SNR is below 30dB.

Next, we examined the performance of the recognizer with the IHMW function described in section 2. Table 2 shows the recognition accuracies resulting from the QLSP and weighted-QLSP using IHMW function. Weighted distance measures showed higher recognition accuracy than the unweighted one for all cases. From this result, we can guess that formant weight on the PCEP using IHMW function would also improve recognition accuracy.

Table 1. Recognition rate using LPCC and quantized LPCC (QLPCC)

SNR	Feature	
	LPCC	QLPCC
Clean	94.4 %	94.6 %
30 dB	88.9 %	88.3 %
20 dB	56.9 %	56.3 %
10 dB	17.7 %	16.3 %

Table 2. Recognition rates using unweighted and weighted LSP

SNR	Feature	
	QLSP	Weighted QLSP
Clean	93.5 %	95.4 %
30 dB	80.2 %	86.3 %
20 dB	33.7 %	40.3 %
10 dB	7.3 %	8.7 %

The normalization factor α in Equation (5) is determined experimently. Fig. 3 shows the recognition result with varying α and SNR. As it can be seen from the figure, the recognition rates using the normalization factor of 0.2 and 0.3 are superior to those with the normalization factor of 0.1 and 0.4.

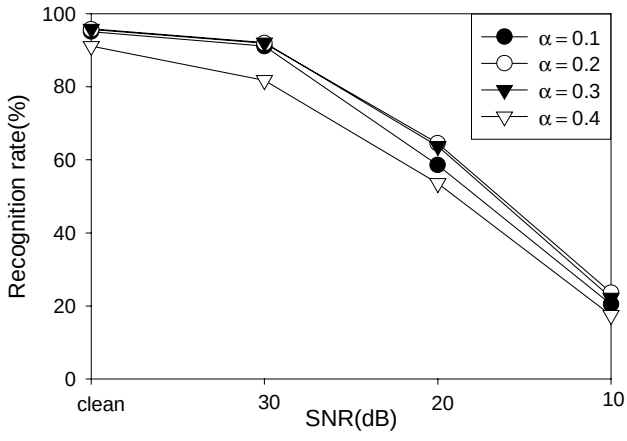

Fig. 3. Recognition rate with varying the normalization factor α and SNR.

To evaluate the recognition performance of the FW-PCEP

feature, we compared it with PCEP. Normalization factor α of 0.3 was used to extract FW-PCEP in this experiment. Table 3 shows the recognition results. From Tables 2 and 3, it can be seen that PCEP shows the superior performance to the weighted-QLSP when SNR is below 30 dB. FW-PCEP shows a consistent superiority to both the weighted QLSP and PCEP. By comparing Table 1 and 3, one can see that the proposed FW-PCEP feature outperforms even the unquantized LPCC.

Table 3. Comparison of recognition rates using PCEP and FW-PCEP

SNR	Feature	
	PCEP	FW-PCEP
Clean	94.2 %	95.8 %
30 dB	87.3 %	92.1 %
20 dB	53.2 %	63.7 %
10 dB	15.1 %	22.2 %

4. CONCLUSIONS

In this paper we presented a new LSP-based cepstral feature with formant weighting. The proposed weighting function, which is based on the fact that the speech spectrum has a peak when adjacent LSP comes close, has an effect of narrowing the bandwidths of the narrow-band poles while their frequencies are unchanged. In this paper we controlled the degree of bandwidth narrowing using the IHMW function of LSFs.

In order to evaluate the proposed feature set, we have constructed a recognition system based on the QCELP speech coder and performed speech recognition experiments using LPCC, LSP, pseudo-cepstrum and proposed formant weighted pseudo-cepstrum. The recognition results show that the proposed method yielded superior recognition accuracy to the other feature sets including unquantized LPCC in all the test environments.

5. REFERENCES

[1] H. K. Kim, K. C. Kim and H. S. Lee, “Enhanced distance measure for LSP-based speech recognition,” Electron. Lett., vol. 29, pp. 1463–1465, Aug. 1993.

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

[3] T. Salonidis and V. Digalakis, “Robust speech recognition for mutiple topological scenarios of the GSM mobile phone system,” in *Proc. ICASSP*, pp.101-104, 1998.

[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] Qualcomm, inc., *High Rate Speech Service Option For Wideband Spread Spectrum Communication Systems*, Feb. 1996.

[6] K. K. Paliwal, “A study of line spectrum pair frequencies for vowel recognition,” *Speech Commun.*, Vol. 8, Nol. 1, pp. 27-33, Mar. 1989.

[7] M. Schroeder, “Direct (nonrecursive) relations between cepstrum and predictor coefficients”, *IEEE ASSP*, vol. 29, pp. 297-301, April. 1981.

[8] D. Naik, K. Assaleh and R. Mammone, “Robust speaker identification using pole filtering,” Proc. ESCA Workshop on Speaker Recognition, pp. 225-230, April 1994.

[9] K. K. Paliwai 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, Feb. 1993.

[10] 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.