EVALUATION OF MEL-LPC CEPSTRUM IN A LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Hiroshi Matsumoto and Masanori Moroto

Shinshu University
Faculty of Engineering
4-17-1 Wakasato, Nagano-shi, Nagano 380-8553, Japan

ABSTRACT

This paper presents a simple and efficient time domain technique to estimate an all-pole model on the mel-frequency scale (Mel-LPC), and compares the recognition performance of Mel-LPC cepstrum with those of both the standard LPC mel-cepstrum and the MFCC through the Japanese dictation system (Julius) with 20,000 word vocabulary. First, the optimal value of frequency warping factor is examined in terms of monosyllable accuracy. When using the optimal warping factors, Mel-LPC cepstrum attains the word accuracies of 93.0% for male speakers and 93.1% for female speakers, which are 2.1% and 1.7 % higher than those of the LPC mel-cepstrum, respectively. Furthermore, this performance is slightly superior to that of MFCC.

1. INTRODUCTION

In automatic speech recognition (ASR), it is important to parameterlize the perceptually relevant aspects of short-term speech spectra. Among the psychoacoustic aspects, the auditory-like frequency scale has been incorporated into a number of spectral analysis methods.

In nonmetric spectral analysis, mel-frequency cepstral coefficients (MFCC) [1] are one of the most popular spectral features in ASR. In parametric spectral analysis, the LPC mel-cepstrum based on an all-pole model is widely used because of its computational simplicity and efficiency. However, while the LPC mel-cepstrum takes into account of auditory like frequency contribution, its frequency resolution is not improved by such a frequency warping of the LPC spectrum.

To alleviate this inconsistency between the LPC and the auditory analyses, several studies have simulated the auditory spectra before the all-pole modeling [3] [2] [4]. The perceptual linear predictive (PLP) analysis is a well-known method [3]. In contrast to these spectral modification, Strube [5] proposed an all-pole modeling to a frequency warped signal converted by the bilinear transformation [8], and presented several computational procedures to approximate the estimate. However, these methods have been rarely used in ASR due to relatively high computational costs compared to the conventional LPC analysis.

Therefore, we have previously proposed a simple and efficient time-domain technique to estimate the warped all-pole model, which is referred to as a "Mel-LPC" analysis, and showed the effectiveness through phoneme recognition tests [6]. This paper further compares the recognition performance of Mel-LPC cepstrum with those of conventional cepstral parameters: the LPC mel-cepstral, and the mel-frequency cepstral coefficients (MFCCs) through the Japanese dictation system (Julius) with 20,000 word vocabulary [7].

The remainder of this paper is organized as follows; Section 2 describes the Mel-LPC analysis. Section 3 demonstrates the superiority of Mel-LPC cepstrum in recognition performance over conventional analyses. Finally, Section 4 summarizes the results.

2. MEL-LPC ANALYSIS

2.1. LPC Analysis on Mel-frequency Scale

The linear prediction method on a warped frequency scale [5] is based on the standard "autocorrelation" method applied to to the bilinear transformed speech signal. Let $x[0], .., x[N-1]$ be a finite speech segment. The frequency warped signal $\{\tilde{x}[n]\}$ is defined by

$$X(z) = \sum_{n=0}^{N-1} x[n]z^{-n} = \tilde{X}(\tilde{z}) = \sum_{n=0}^{\infty} \tilde{x}[n]\tilde{z}^{-n} \qquad (1)$$

where $\tilde{z}^{-1}(z)$ is the first order all-pass filter,

$$\tilde{z}^{-1}(z) = \frac{z^{-1} - \alpha}{1 - \alpha \cdot z^{-1}}. \qquad (2)$$

In the frequency domain, the spectrum $X(e^{j\lambda})$ on the linear frequency axis λ is converted to the frequency-warped spectrum $\tilde{X}(e^{j\tilde{\lambda}})$ on the warped-frequency axis $\tilde{\lambda}$ by the frequency mapping function,

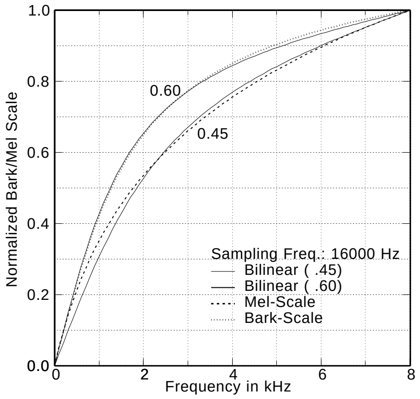

Fig. 1. The frequency mapping function in the bilinear transformation.

$$\tilde{\lambda} = \lambda + 2 \cdot \tan^{-1}\left\{\frac{\alpha \sin \lambda}{1 - \alpha \cos \lambda}\right\}. \tag{3}$$

Figure 1 shows the approximated frequency mapping functions of the bark and the mel scales (the solid lines) at the sampling frequency of 16kHz together with the "Mel" and the "Bark" scales (the dotted lines) based on the psychoacoustic works [9].

The prediction error minimization in the $\tilde{z}$ domain is equivalent to minimize the output energy $\tilde{E}$ of the inverse filter, $\tilde{A}(\tilde{z}(z)) = 1 + \sum_{n=1}^{p} \tilde{a}_n \tilde{z}^{-n}(z)$, in the z domain as shown in Figure 2. In this figure, $W(z)$ is defined by

$$W(z) = \frac{\sqrt{1-\alpha^2}}{1 - \alpha \cdot z^{-1}}, \tag{4}$$

and $\left|W(e^{j\lambda})\,\right|^2$ is equal to $d\tilde{\lambda}/d\lambda$. However, since either the frequency-warped signal $\tilde{x}[n]$ or the prefiltered signal $\{x_w[n]\}$ is an infinite sequence, the LPC analysis on the mel-frequency scale needs an approximation by truncating $\tilde{x}[n]$ or $x_w[n]$.

2.2. Mel-LPC Analysis

Unlike the Strube's formulation, the Mel-LPC analysis [6] removes $W(z)$ in Figure 2. That is, this method directly minimizes the output energy of a mel-inverse filter $\tilde{A}_w(\tilde{z}(z))$ in the z domain without pre-filtering $x[n]$ as shown in Figure 3.

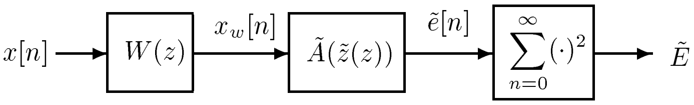

Fig. 2. Warped LPC analysis in the z domain.

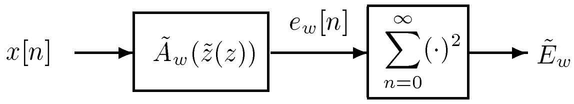

Fig. 3. Mel-LPC analysis in the z domain.

This modification is equivalent to replacing $x[n]$ in Figure 2 by the signal whose z-transform is $X[z]W^{-1}[z]$. Therefore, the estimated inverse filter $\tilde{A}_w(\tilde{z})$ is no longer the same as $\tilde{A}(\tilde{z})$, but instead includes the effect of $W^{-1}(z)$. Then, we write this estimated spectrum as

$$\tilde{H}_w(z) \quad = \quad \frac{\tilde{\sigma}_w}{1 + \sum_{n=1}^{p} \tilde{a}_{w,n} \tilde{z}(z)^{-n}}. \tag{5}$$

Given a finite speech segment, $x[0], .., x[N-1]$, the mel-prediction coefficients $\{\tilde{a}_{w,i}\}$ are estimated by minimizing the prediction error energy over an infinite time interval,

$$\tilde{E}_w = \sum_{n=0}^{\infty} \left(\sum_{i=0}^{p} \tilde{a}_{w,i} y_i[n] \right)^2 . \tag{6}$$

As a result, $\{\tilde{a}_{w,i}\}$ and $\tilde{\sigma_w}$ are given by the Durbine's algorithm using the following "generalized" autocorrelation function in which a unit delay is replaced by the all-pass filter,

$$\tilde{r}_\alpha[m] = \sum_{n=0}^{N-1} x[n] y_m[n], \tag{7}$$

where $y_m[n]$ is the output signal of $\tilde{z}^{-m}(z)$ excited by $x[n]$. In terms of Parceval's theorem $\tilde{r}_\alpha[m]$ in equation (7) can be written in the mel-frequency domain $\tilde{\lambda}$ as

$$\tilde{r}_\alpha[m] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left|\tilde{X}(e^{j\tilde{\lambda}}) \tilde{W}(e^{j\tilde{\lambda}})\,\right|^2 \cdot \cos m\tilde{\lambda}\, d\tilde{\lambda}, \tag{8}$$

where $\tilde{W}(\tilde{z}) = W^{-1}(z) = \sqrt{1-\alpha^2}/(1 + \alpha\, \tilde{z}^{-1})$. Consequently, $|\tilde{H}_w(e^{j\tilde{\lambda}})|^2$ provides the spectral envelope of $|\tilde{X}(e^{j\tilde{\lambda}})\tilde{W}(e^{j\tilde{\lambda}})|^2$. The frequency-weighting function $\tilde{W}(e^{j\tilde{\lambda}})$ can be removed by filtering $\tilde{r}_\alpha[m]$ with the FIR filter $[\tilde{W}(\tilde{z})\tilde{W}(\tilde{z}^{-1})]^{-1}$.

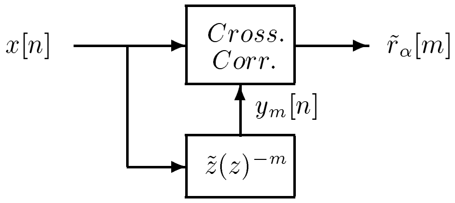

Fig. 4. The generalized autocorrelation function.

However, since this term works as a first order preemphasis $\beta(1-\alpha \cdot z^{-1})$ in the z domain, we use cepstral coefficients (Mel-LPC cepstral coefficients) derived from $\{\tilde{a}_{w,k}\}$ in the following experiments.

The computational cost for the Mel-LPC analysis is two times greater than that for the standard LPC analysis due to computation of $y_m[n]$ in equation (7). However, this computational load is much lower than Strube's method [5], and the prediction coefficients are estimated without any approximation.

3. EVALUATION

3.1. Experimental Conditions

The recognition performance of the Mel-LPC cepstrum was compared with those of conventional cepstral parameter: the LPC mel-cepstral, and the mel-frequency cepstral coefficients (MFCCs).

The speech data was sampled at 16kHz. A speech segment of 25ms with a frame shift of 10ms was preemphasized with $(1-0.90\tilde{z}^{-1})$, and was weighted by Hamming window. In both the LPC and the Mel-LPC analysis the number of poles was set to 16, and the number of channels in MFCC analysis was set to 24. Every feature vector consists of 12 cepstral and 13 delta-cepstral coefficients including the 0th delta-cepstral coefficient. As an example, Figure 5 compares a LPC spectrum (the dotted line) and a Mel-LPC spectrum (the bold line) for $\alpha = 0.6$ together with the FFT spectrum. It is seen that the Mel-LPC spectrum represents a more precise spectral envelope below 2kHz than the LPC spectrum with the same number of poles ($p = 16$), and especially separates the two adjacent formants around 500Hz.

In the recognition experiments, the IPA Japanese dictation systems with 20k word vocabulary [7] were used. The sub-word models are 124 gender-dependent monosyllable HMMs. The structure of HMMs is a left-to-right model with 3 emitting states for vowels, double consonant(/q/), syllabic nasal(/N/) and silences, and with 5 emitting states for other syllables. A state consists of 16 Gaussians.

In training HMMs, we used about 20k sentences uttered by 134 speakers for each gender, which are from database of phonetically balanced sentences (ASJ-PB) and newspaper article texts (ASJ-JNAS). The test set consists of 100 sentences uttered by 23 speakers for each gender. The results are evaluated in terms of percentage accuracy ($Acc[\%] = \frac{N-S-D-I}{N} \times 100$),where for N tokens, S, D, and I are substitution, deletion, and insertion errors, respectively.

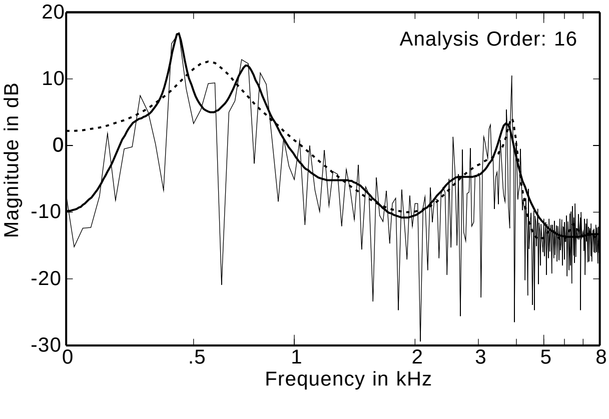

Fig. 5. Comparison of the Mel-LPC (bold) and the LPC (dotted) spectra.

3.2. Effect of Frequency Warping Factor

First, the optimal frequency warping factors in both the Mel-LPC cepstrum and the LPC mel-cepstrum were examined in terms of the syllable accuracy obtained without both the bigram and the trigram language models. Figure 6 compares the syllable accuracy as a function of the frequency warping factor α for both the LPC mel-cepstrum and the Mel-LPC cepstrum together with that for the MFCC.

In the Mel-LPC analysis, the optimal value of α for male speakers is around 0.5, which is between the mel and bark scales, whereas that for female speakers is 0.4, which is smaller than that corresponding to the mel-scale. This difference between both gender is also observed in the case of the LPC mel-cepstrum. This difference seems to be caused by the sparse harmonics in female voices. On the other hand, the optimal values for the LPC mel-cepstrum are much smaller than those corresponding to the bark and mel-frequency scales.

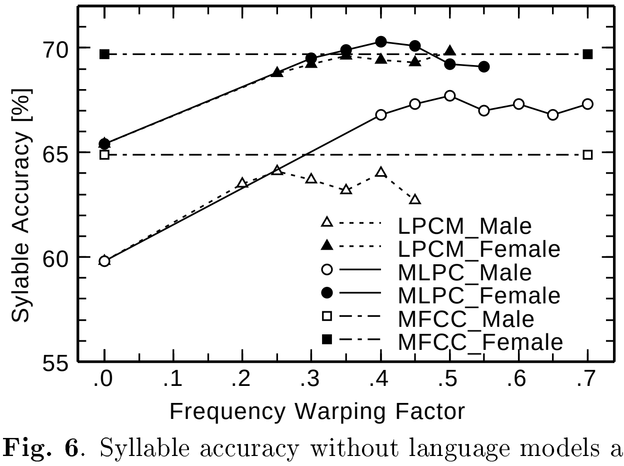

Fig. 6. Syllable accuracy without language models as a function of frequency warping factor α

As a result, it is clear that both Mel-LPC cepstrum and the LPC mel-cepstrum outperform the LPC cepstrum, that is, for $\alpha = 0$, due to auditory-like frequency contribution. Furthermore, the Mel-LPC cepstrum with the optimal frequency warping improves recognition accuracy over the LPC mel-cepstrum and the MFCC. The same tendency is also found for female speakers, but the improvement is small. This improvement is caused by higher frequency resolution in the Mel-LPC analysis.

3.3. Evaluation by Dictation

Since the optimal weights on both the language score and the insertion penalty depend on the kinds of feature parameters, the values of both weights were optimized. The optimal values for the first and the second pathes in the decoder are shown in Table 1. The optimal values for the Mel-LPC cepstrum tend to be larger than the LPC mel-cepstrum and the MFCC.
As a results, the Mel-LPC cepstrum improved the word accuracy from 90.9% for the LPC mel-cepstrum to 93.0% for male speakers and from 91.4% to 93.1% for female speakers as shown in Figure 7. Furthermore, Mel-LPC cepstrum attained slightly higher recognition accuracy than the MFCC. This improvement is caused by the reduction of substitution and deletion errors as shown in Table 2.

Table 1. The optimal values of the language weight and the insertion penalty in three kinds of parameters.

	1st Pass		2nd Pass	
	W_L	P_I	W_L	P_I
MLPC	8.6	2.0	8.8	1.5
LPCM	7.2	1.0	7.0	1.0
MFCC	6.5	-0.8	7.0	-0.8

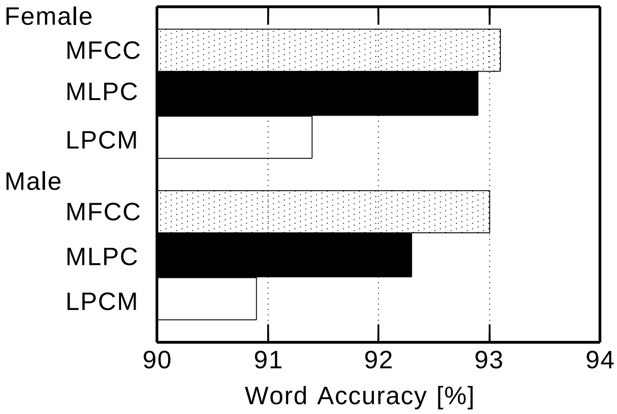

Fig. 7. Word accuracy for the Mel-LPC(MLPC) cepstrum, the LPC mel-cepstrum and the MFCC.

Table 2. Three types of recognition errors for MLPC, LPCM, and MFCC.

	male			female		
	Sub	Del	Ins	Sub	Del	Ins
MLPC	5.0	0.9	1.0	5.2	0.5	1.2
MFCC	5.6	1.2	1.2	5.2	0.8	1.1
LPCM	6.6	1.1	1.4	6.3	0.8	1.6

4. CONCLUSIONS

This paper has presented a simple and efficient time domain method in all-pole modeling on the mel-frequency scale, and has evaluated the performance through large vocabulary continuous speech recognition. The Mel-LPC cepstrum has achieved a significant improvement in recognition accuracy over the LPC mel-cepstrum, and has attained slightly higher recognition accuracy than the MFCC.

Acknowledgment

This work is partially supported by Grants from Ministry of Education, Science and Culture of Japan, No.12680376.

5. REFERENCES

[1] S.Davis and P.Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans., Vol.ASSP-28, No.4, pp.357-366, 1980.

[2] S.Itahashi and S.Yokoyama, "A formant extraction method utilizing mel scale and equal loudness contour," Speech Transmission Lab.-QPSR (Stockhlm) (4), pp.17-29, 1987.

[3] H.Hermansky, "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am., Vol.87, No.4, pp.1738-1752, 1990.

[4] M.G.Rahim and B.H.Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," IEEE Trans. on Speech and Audio Processing, Vol.4, No.1, pp.19-2.50, 1996.

[5] H.W. Strube, "Linear prediction on a warped frequency scale," J.Acoust.Soc. Am., Vol.68, No.4, pp.1071-1076.

[6] H.Matsumoto, Y.Nakatoh and Y.Furuhata, "An efficient Mel-LPC analysis method for speech recognition," Proc. of ICSLP98, pp.1051-1054, 1998.

[7] T.Kawahara, et al., "Japanese dictation toolkit - free software repository for speech recognition -," ESCA Workshop on Interactive Dialogue in Multi-Modal Systems, p.161, 1999.

[8] A.V.Oppenheim and D.H.Johnson, "Discrete Representation of Signal," Proc. IEEE, Vol.60, No.6, pp.681-691, 1972.

[9] E.Zwicker and E.Terhardt, "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," J. Acoust. Soc. Am., 68, pp.1523-1525, 1980.