AN EFFICIENT MEL-LPC ANALYSIS METHOD FOR SPEECH RECOGNITION

Hiroshi Matsumoto†, Yoshihisa Nakatoh†‡, and Yoshinori Furuhata †

†Dept. of Electrical & Electronic Eng., Faculty of Engineering, Shinshu University
500 Wakasato, Nagano-shi, Nagano 380, Japan,
‡Multimedia Development Center, Matsushita Electric Industrial Co., Ltd.
1006 Kadoma, Kadoma-shi, Osaka, 571-8501 Japan
E-mail: matsu@sp.shinshu-u.ac.jp and nakatoh@arl.drl.mei.co.jp

ABSTRACT

This paper proposes a simple and efficient time domain technique to estimate an all-poll model on a mel-frequency axis (Mel-LPC). This method requires only two-fold computational cost as compared to conventional linear prediction analysis. The recognition performance of mel-cepstral parameters obtained by the Mel LPC analysis is compared with those of conventional LP mel-cepstra and the mel-frequency cepstrum coefficients (MFCC) through gender-dependent phoneme and word recognition tests. The results show that the Mel-LPC cepstrum attains a significant improvement in recognition accuracy over conventional LP mel-cepstrum, and gives slightly higher accuracy for male speakers and slightly lower accuracy for female speakers than MFCC.

1. INTRODUCTION

In the front end of speech recognition system, it is important to parameterlize the perceptually relevant aspects of short-term speech spectrum. In filter-bank based systems, auditory-like frequency resolution has been incorporated into parameterlization such as mel frequency cepstral coefficients (MFCC) [1], perceptual linear predictive (PLP) [2] and mel-linear predictive (LP) cepstral coefficients [3]. These parameters have been shown to be superior to conventional LP cepstrum.

On the other hand, the LP analysis has been widely used as a front end in speech recognition system because of its computational simplicity and efficiency. However, the all-pole model approximates speech spectra equally well at all frequency band, and thus this property is inconsistent with human hearing. Although the LP spectrum is usually warped in cepstral or linear predictor domain after LP analysis [4], the frequency resolution is not improved yet by such a post processing. To alleviate this inconsistency between LP and auditory analysis, Strube [5] proposed a linear prediction on warped frequency scale based on a bilinear transformation, and investigated several computational procedures classified into "autocorrelation" and "covariance" methods. This analysis method was proved to be effective in speech coding [6], and could potentially produce improved cepstral feature as the MFCC or PLP analysis. However, this method has been rarely used in speech recognition due to relatively high computational load compared to conventional LP analysis. Another all-pole modeling on mel-frequency scale was proposed as a special case ($\gamma = -1$) in the mel-generalized cepstral analysis method [7]. This method needs an iterative procedure to minimize a non-linear criterion.

This paper proposes a simple and efficient time-domain technique to estimate an all-pole model on mel-frequency axis by Strube based on the error minimization on the linear frequency axis. The computational cost is only twice as much as conventional LP method without any approximation. This technique will be refer to as Mel-LPC analysis method (hereafter "warped" will be replaced by "mel"). The recognition performance of mel-cepstral parameters obtained by the Mel-LPC analysis is compared with those of conventional LP mel-cepstra and the mel-frequency cepstrum coefficients (MFCC) through gender-dependent phoneme and word recognition tests.

2. MEL-LPC ANALYSIS

2.1 Autocorrelation Method on Mel-Frequency Axis

In this study, we consider a speech segment of finite length, $x[0], .., x[N-1]$, which is usually windowed and preemphasized in advance. In the "autocorrelation" method by Strube [5], the standard autocorrelation method is applied to frequency-warped speech signal $\{\tilde{x}[n]\}$ which is defined by

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

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

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

Fig.1 illustrates this method. In spectral domain, an all-poll model $\tilde{\sigma}/\tilde{A}(\tilde{z})$ approximates the warped spectrum $\tilde{X}(e^{j\tilde{\lambda}})$ converted from the spectrum $X(e^{j\lambda})$ on the linear frequency axis by the following phase transfer function of the all-pass filter;

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

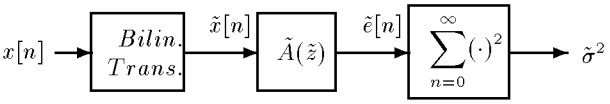

Figure 1: All-poll modeling on the mel-frequency axis.

The inverse filter on the mel-frequency axis,

$$\tilde{A}(\tilde{z}) = \sum_{k=0}^{p} \tilde{a}_k \tilde{z}^{-k}, \quad \tilde{a}_0 = 1 \tag{4}$$

is estimated by Durbin's algorithm using the following mel-autocorrelation coefficients:

$$\tilde{r}[m] \quad = \quad \sum_{n=0}^{\infty} \tilde{x}[n]\tilde{x}[n-m] \tag{5}$$

However, as shown in equation (1), since the bilinear transformation of a finite sequence results in an infinite sequence, the direct calculation of the mel-autocorrelation coefficients in equation (5) is not practical. Then, Strube proposed three methods to approximate $\tilde{r}[m]$ [5]. However, these require a FFT spectrum or a longer autocorrelation sequence of $x[n]$, and thus are computationally undesirable from a practical point of view.

2.2 Mel-Autocorrelation Method on Linear Frequency Axis

Strube also suggested two time domain methods based on direct error minimization for $x[n]$ [5]. The total error power $\tilde{\sigma}^2$ on the mel-frequency axis can be written by the integral on the linear frequency axis as follows:

$$\tilde{\sigma}^2 \quad = \quad \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \tilde{A}(e^{j\tilde{\lambda}}) X(e^{j\lambda}) W(e^{j\lambda}) \right|^2 d\lambda, \tag{6}$$

where

$$W(z) = \frac{\sqrt{1-\alpha^2}}{1-\alpha \cdot z^{-1}}. \tag{7}$$

Thus, on the linear frequency axis, the error minimization on the mel-frequency axis is equivalent to minimize the output power of $\tilde{A}(z)$ exited by the pre-filtered signal $x_w[n]$ with $W(z)$ as shown in Fig.2. However, since $x_w[n]$ is an infinite sequence, this minimization problem is not tractable.

This paper proposes another estimation method in which $W(z)$ in Fig.2 is removed as shown in Fig.3. This modification is equivalent to replacing x[n] in Fig.2 by the signal whose z-transform is $X[z]W^{-1}[z]$. Therefore, the inverse filter, which is denoted by $\tilde{A}_w(\tilde{z})$, is no longer the same as $\tilde{A}(\tilde{z})$, but instead $\tilde{A}_w(\tilde{z})$ includes the effect of $W^{-1}(z)$. However, as will be seen later, this effect can be exactly removed in the mel-autocorrelation domain.

As a result of minimizing the total error power $\tilde{\sigma}_w^2$ over infinite time interval, the mel-predictors $\tilde{a}_{w,k}$'s are obtained by solving for the following normal equation:

$$\sum_{j=1}^{p} \phi(i,j)\tilde{a}_{w,j} = -\phi(0,i), \qquad (i = 1,..,p), \tag{8}$$

where the coefficient $\phi(i,j)$ is given by

$$\phi(i,j) \quad = \quad \sum_{n=0}^{\infty} y_i[n] y_j[n], \tag{9}$$

using the output sequence $y_i[n]$ of the ith order all-pass filter excited by $y_0[n] = x[n]$. In terms of Parceval's theorem, $\phi(i,j)$ can be rewritten on the mel-frequency axis as

$$\phi(i,j) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \tilde{X}(e^{j\tilde{\lambda}}) \tilde{W}(e^{j\tilde{\lambda}}) \right|^2 \cdot \cos(i-j)\tilde{\lambda}\, d\tilde{\lambda}. \tag{10}$$

where $\tilde{W}(\tilde{z})$ is equal to $W^{-1}(z)$. Consequently, $\phi(i,j)$ is equal to the autocorrelation coefficient $\tilde{r}_w(|i-j|)$ whose Fourier transform is equal to the warped and frequency-weighted power spectrum $|\tilde{X}(e^{j\tilde{\lambda}})\tilde{W}(e^{j\tilde{\lambda}})|^2$. Therefore, the normal equation (8) becomes an autocorrelation equation as in conventional LP analysis.

Most importantly, since $\phi(i,j)$ is a function of the difference $|i-j|$, $\phi(i,j)$ becomes equal to the sum of the following finite terms without any approximation;

$$\phi(i,j) = \tilde{r}_w(|i-j|) = \sum_{n=0}^{N-1} x[n] y_{|i-j|}[n] \tag{11}$$

Therefore, due to the cost for computing N points of $y_i[n]$ for each i, the Mel-LPC analysis is accomplished with about two-fold increase in computation over conventional LP analysis. This computational load is much lower than those of both "autocorrelation" and "covariance" methods in [5].

Finally, the mel-inverse filter $\tilde{A}(\tilde{z})$ is easily obtained by deriving $\tilde{r}[m]$ from $\tilde{r}_w[m]$ as follows. Since the z-transforms of $\tilde{r}[m]$ and $\tilde{r}_w[m]$ are $|\tilde{X}(\tilde{z})|^2$ and $|\tilde{X}(\tilde{z})\tilde{W}(\tilde{z})|^2$, respectively, $\tilde{r}[m]$ is exactly calculated from $\tilde{r}_w[m]$ using the equation,

$$\tilde{r}[m] = \beta_0 \tilde{r}_w[m] + \beta_1 \left\{ \tilde{r}_w[m-1] + \tilde{r}_w[m+1] \right\}, \tag{12}$$

where $\beta_0 = (1+\alpha^2)(1-\alpha^2)^{-1/2}$, and $\beta_1 = \alpha(1-\alpha^2)^{-1/2}$. In speech recognition applications, however, $\tilde{r}_w[m]$ can be

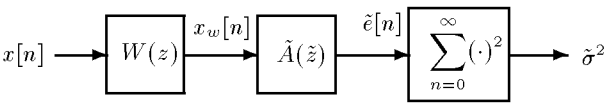

Figure 2: Mel all-poll modeling on the linear frequency axis.

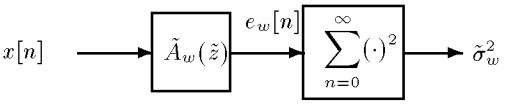

Figure 3: The proposed mel all-poll modeling on the linear frequency axis.

used, since the estimated spectrum $\tilde{\sigma}_w/\tilde{A}_w(\tilde{z})$ represents the envelope of $\tilde{X}(e^{j\tilde{\lambda}})\tilde{W}(e^{j\tilde{\lambda}})$ and $\tilde{W}(\tilde{\lambda})$ works like a pre-emphasis.

2.3 Smoothing on Mel-Frequency Axis

The harmonics in lower mel-frequency band become so much sparse that single harmonics appear as spectral poles. This is undesirable in all-pole modeling, especially for female voices. In order to alleviate this problem, the mel autocorrelation coefficient $\tilde{r}[m]$ is weighted by a lag window. This lag windowing is similar to a mel-filter bank in the mel LP or MFCC analysis. In this study, we choose the Blackman and Harris window as a lag window, and examine its appropriate length in later experiments.

3. EVALUATION

3.1 Database and Speech Analysis

In this study, we used 520 words uttered by each of 60 male and 60 female speakers from the ATR C-set database. The recognition performance of mel-cepstral parameters (MLPC) obtained by the Mel-LPC analysis was compared with those of conventional LP mel-cepstra (LPMC) and MFCC through gender-dependent phoneme and word recognition tasks. For three analysis methods, the speech signal from ATR database was down-sampled to 12kHz. A speech segment of 20ms with frame shift of 10ms was pre-emphasized with $(1 - 0.95z^{-1})$, and was weighed by Hamming window. A feature vector consists of both cepstral and delta-cepstral coefficients excluding the 0th terms (power terms). The number of cepstral coefficients was set equal to the predictor length for all-poll modeling and to the number of filter channels minus one for MFCC analysis.

In the isolated word recognition task, two set of gender-dependent phoneme HMMs were used; a set of 35 phonemes and a set of 260 context-dependent phoneme including silence. In the phoneme recognition task, only the first set was used. The structure of HMMs is a left-to-right model with 3 emitting states, which consist of 4 gaussians for 35 phonemes and 2 gaussians for 260 phonemes. Each phoneme model was trained using 520 words from each of 40 speakers for each gender. All the words from the other 20 speakers was used for testing. A syntax consists of the preceding and following silences for a word. In phoneme recognition task, the results are evaluated in terms of percentage accuracy ($Acc = [(N - S - D - I)/N] \cdot 100\%$), and percentage correct ($Corr = [(N - S - D)/N] \cdot 100\%$), where for N tokens, S, D, and I are substitution, deletion, and insertion errors, respectively.

3.2 Phoneme Recognition

(1) Effect of Warping Factor

First, we examines the effect of the warping factor α on phoneme recognition accuracy. As the bilinear transformation is an approximation of herz-to-perceptual scale mapping, the optimum warping factor is not clear. According to the mel-herz and Bark-herz transformations [8], [9], the mel and bark frequency scales are approximated by $\alpha = 0.41$ and $\alpha = 0.50$, respectively. Then, three values of 0.37, 0.41 and 0.50 were evaluated. Table 1 shows the results of phoneme recognition for male and female speakers. This table shows that $\alpha = 0.41$ gives the best scores for both genders. Although the approximation to the Bark scale seems to be appropriate as to spectral resolution, the result indicates the mel scale is suitable. This value of α will be used throughout the following experiments.

Table 1: Effect of frequency warping parameter in Mel-LPC analysis in phoneme recognition.

α	Male		Female	
	Acc	Corr	Acc	Corr
0.37	63.9	74.9	54.9	74.1
0.41	64.9	75.0	55.7	74.3
0.50	64.4	74.2	55.5	74.0

(2) Effect of Analysis Order

Second, the performance of three analysis methods were compared in phoneme recognition task using the set of context independent HMMs. Table 2 shows the phoneme recognition scores as a function of the number of cepstral coefficients. MLPC attains 1.4 to 7.1 percent higher accuracy depending on the analysis order and genders than conventional LP mel-cepstral coefficients. These differences become greater in higher analysis order. The performance of conventional LP mel-cepstrum with the order of 18 corresponds to Mel-LPC cepstrum with the order of about 12. In comparison with MFCC, MLPC is slightly better than MFCC for male speakers, but slightly worse in higher orders than MFCC for female speakers. This is considered to be caused by high pitch harmonics of female voices. Insertion errors are larger for female speakers than for male speakers.

(3) Effect of Lag Window Length

To find out the optimal length of lag window for high pitch voices, phoneme recognition experiments were carried out for the analysis order of 14 with the lag window length of 100 to 160. Table 3 shows the scores as a function of the window length. The window length of about 140 seems to be

Table 2: Comparison of three type of mel-cepstral parameters in phoneme recognition.

(A) Male speakers

Analysis Order	MLPC		MFCC		LPMC	
	Acc	*Corr*	*Acc*	*Corr*	Acc	Corr
10	60.8	72.5	59.2	73.1	58.2	69.9
14	64.9	75.0	63.2	74.3	61.0	73.3
18	65.8	75.3	65.3	75.3	61.7	73.9

(B) Female speakers

Analysis Order	MLPC		MFCC		LPMC	
	Acc	*Corr*	*Acc*	*Corr*	Acc	Corr
10	50.1	71.4	48.7	73.4	48.7	71.4
14	55.7	74.3	57.6	75.3	48.6	72.9
18	57.5	75.0	57.2	75.0	51.3	73.9

Table 3: Effect of lag-window length for Mel-LPC analysis in phoneme recognition.

Lag	Male		Female	
[point]	Acc	Corr	Acc	Corr
∞	64.9	75.0	55.7	74.3
160	65.1	74.9	55.9	75.4
140	65.0	75.0	56.7	75.5
120	64.6	74.8	56.7	75.4
100	62.4	74.8	56.7	74.8

Table 4: Comparison of word recognition rates obtained by three mel-cepstral parameters.

(A) Male Speakers

Context	MLPC	MFCC	LPMC
Context Free	92.1	91.5	90.7
Context Depend.	96.3	96.2	95.4

(B) Female Speakers

Context	MLPC	MFCC	LPMC
Context Free	87.2	89.1	84.8
Context Depend.	93.4	94.4	90.9

best, giving 1.0 percent improvement in accuracy for female speakers.

3.3 Word Recognition

Using two sets of phoneme HMMs, the performances of three analysis methods were compared through isolated word recognition of 520 word vocabulary. The analysis order was set to 14, and the lag window with a lenght of 140 was applied only to female speech. As shown in Table 4, the relative recognition scores among three methods are similar to those in the phoneme recognition. MLPC attains the highest scores for male speaker, but slightly lower scores than MFCC. The improvements in recognition rate by MLPC and MFCC over LPMC are larger for female speakers than for male speakers.

4. DISCUSSION

The Mel-LPC analysis has been shown to be superior to conventional LPC analysis in speech recognition. While the performance of the mel-LPC is slightly better than that of MFCC for male speakers, it is slightly worse than MFCC for female speakers due to too much frequency resolution in low frequency band. Although this disadvantage was improved by lag windowing on the mel-frequency axis, it was unsatisfactory. Therefore, it might be required to reduce frequency resolution in lower frequency band while preserving spectral resolution of close formants. Further improvements are expected by choosing appropriate time window as well as lag window.

The performance of the Mel-LPC cepstrum is comparable to that of MFCC, but the Mel-LPC analysis has still an advantage on computational load over the MFCC analysis. This method does not need FFT calculation and log conversion. The major computations required are all-pass filtering, correlation calculation, Durbin's recursion, and predictor-to-cepstral conversion. Therefore, the Mel-LPC analysis is desirable for practical implementation.

5. CONCLUSION

This paper has presented a simple and efficient time domain method for mel-scaled all-pole modeling on the linear frequency axis. The computational cost is only twice as much as conventional LP method, The proposed method has achieved a significant improvement in recognition accuracy over conventional LP analysis, and a slightly higher recognition accuracy for male speakers than the MFCC analysis.

In future work, it is necessary to develop a spectral smoothing method for high pitch voices, and to evaluate the performance in noisy speech recognition, and in continuous speech recognition.

Acknowledgment

This work is partially supported by Grants from Ministry of Education, Science and Culture of Japan, #10680376.

References

1. S.Davis and P.Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol.ASSP-28, No.4, pp.357-366, 1980.
2. H.Hermansky, "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am., Vol.87, No.4, pp.1738-1752, 1990.
3. 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. , 1996.
4. A.V.Oppenheim and D.H.Johnson, "Discrete Representation of Signal," Proc. IEEE, Vol.60, No.6, pp.681-691, 1972.
5. H.W. Strube, "Linear prediction on a warped frequency scale," J.Acoust.Soc. America, vol.68, no.4, pp.1071-1076.
6. E.Kruger and H.W.Strube, "Linear prediction on a warped frequency scale," IEEE Trans. Acoust., Speech, and Signal Processing, Vol.ASSP-36, pp.1529-1531, 1988.
7. K.Tokuda, et. al., "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation," Proc. of ICSLP94, pp.1043-1046, 1994.
8. S.Seneff, "Pitch and spectral estimation of speech based on an auditory synchrony model," ICASSP84,(1984).
9. J.Makhoul and L.Cosell, "LPCW: An LPC vocoder with linear predictive warping," Proc. of ICASSP76, pp.446-469 (1976).