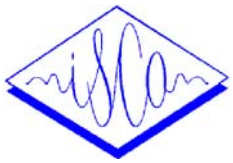
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

AN ADAPTIVE MEL-LPC ANALYSIS FOR SPEECH RECOGNITION

Yoshihisa NAKATOH, Makoto NISHIZAKI, Shinichi YOSHIZAWA, Maki YAMADA

Advanced Technology Research Laboratories, Matsushita Electric Industrial Co., Ltd.
3-4 Hikaridai, Seika-cho, Souraku-gun, Kyoto-fu, 619-0237, Japan
E-mail: nakatoh.yoshihisa@jp.panasonic.com

10.21437/Interspeech.2004-340

Abstract

This paper describes a new speech analysis method, an adaptive Mel-LPC (AMLPC) analysis method, using human auditory characteristics. The Mel-LPC analysis method that we have proposed is an efficient time domain technique to estimate the warped predictors from input speech directly. However, the frequency resolution of spectrum obtained by Mel-LPC analysis is constant regardless of the characteristics of input speech at each analysis frame. In the AMLPC analysis, it is probable to estimate the spectrum coefficients with optimal frequency resolution according to the characteristics of the phoneme at each analysis frame, because the spectral slope or the formant is different according to phoneme (vowels, fricatives and so on). The recognition performance of mel-cepstrum parameters obtained by the AMLPC analysis was compared with those of mel-cepstrum parameters obtained by the conventional LPC analysis and the Mel-LPC analysis through gender-dependent phoneme and word recognition. The results show that the proposed method leads to a significant improvement in recognition accuracy over conventional LPC analysis, and a slightly improvement of error rate about 10% over the Mel-LPC analysis.

1. Introduction

In the front end of speech recognition system, it is important to parameterize the perceptually relevant aspects of short-term speech spectrum. Usually, the LPC analysis has been widely used as a front end in speech recognition system or speech coding 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. The Mel-LPC analysis method that we have proposed is an efficient time domain technique to estimate the warped predictors from input speech directly. This analysis method was proved to be effective in speech recognition [1], and could potentially produce improved cepstral feature as the mel frequency cepstral coefficients (MFCC) [2], perceptual linear predictive (PLP) [3] and mel-linear predictive (LP) cepstral coefficients [4], because of their auditory like frequency resolution. However the frequency resolution of spectrum obtained by Mel-LPC analysis is constant regardless of the characteristics of input speech at each analysis frame.

This paper describes a new speech analysis method, an adaptive Mel-LPC (AMLPC) analysis method, using human auditory characteristics. In the AMLPC analysis, the frequency resolution of spectrum is different according to spectrum of input speech at each analysis frame. Because the important part on frequency axis is difference by each phoneme (vowels, fricatives and so on), it is reasonable to analysis by deference frequency resolution for each phoneme. The AMLPC analysis method is based on the Mel-LPC analysis, because the frequency resolution of spectrum can be changed easily by the only one factor to control the frequency resolution in the Mel-LPC analysis. In experiments, the recognition performance of mel-cepstrum parameters obtained by the AMLPC analysis is compared with those of mel-cepstrum parameters obtained by the conventional LPC analysis and those of the Mel-LPC analysis gender-dependent phoneme and word recognition tests.

2. Adaptive Mel-LPC Analysis

The feature extraction is very important factor to obtain a good recognition performance for ASR. In the AMLPC analysis, it is probable to estimate the spectrum coefficients with optimal frequency resolution according to the spectral slope or the formant. There is some potential to obtain a good recognition performance because of the adaptive feature extraction according to the phoneme such as vowels, fricatives and so on. The AMLPC analysis consists of five key parts, (1) calculation of normalized 1st order autocorrelation coefficient, (2) estimation of control factor, (3) Mel-LPC analysis, (4) compensation of pre-emphasis and (5) compensation for frequency warping. In the AMLPC analysis, first, the first order autocorrelation coefficient is calculated from the input speech at each analysis frame and is normalized by its own power (zero order autocorrelation coefficient). The control factor is estimated from the normalized autocorrelation coefficient. Although the warped predictors are estimated easily from input speech using control factor in the Mel-LPC analysis, their predictors have the characteristic of pre-emphasis caused by the Mel-LPC analysis. Therefore the warped predictors are completely compensated by filtering with the second order FIR filter. The warped predictors at each analysis frame are warped in the deference ratio of on frequency axis. Their predictors are compensated by Oppenheim's method [5]. The block diagram of the AMLPC analysis method is illustrated in Figure 1.

2.1. Mel-LPC Analysis

The Mel-LPC analysis method that we have proposed is 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 LPC method. In this analysis, we use a same warped inverse filter on the linear frequency axis without any prefiltering unlike Strube's method [6],

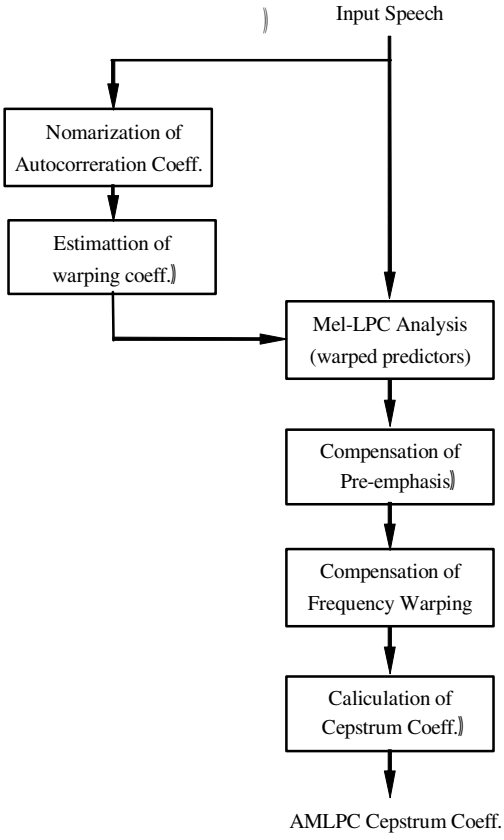


Figure 1: The block diagram of AMLPC analysis

$$A_w(z) = \tilde{A}_w(\tilde{z}) = \sum_{n=0}^{p} \tilde{a}_{w,n} \tilde{z}^{-n} \tag{1}$$

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha \cdot z^{-1}} \tag{2}$$

where $\tilde{z}^{-1}$ is the first order all-pass filter. For a windowed input signal segment $x[0],\ldots,x[N-1]$, the error power is given by equation (3), and the warped predictors $\{\tilde{a}_{w,k}\}$ are estimated so as to minimize the error power over infinite time interval unlike the "covariance method" in [6].

$$\tilde{\sigma}_w^{\,2} = \sum_{n=0}^{\infty} \left\{ \sum_{k=0}^{p} \tilde{a}_{w,k} \cdot y_k[n] \right\}^2 \tag{3}$$

It should be noted that the estimated predictors $\{\tilde{a}_{w,k}\}$ are different from the predictors $\{\tilde{a}_k\}$ defined in Strube's method [6]. The warped predictors are obtained by solving for the following normal equation:

$$\sum_{j=1}^{p} \phi(i,j)\tilde{a}_{w,j} = -\phi(i,0) \qquad (i = 1,\ldots,p) \tag{4}$$

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

$$\phi(i,j) = \sum_{n=0}^{\infty} y_i[n] y_j[n] \tag{5}$$

using the output sequence $y_k[n]$ of the kth order all-pass filter excited by $y_0[n] = x[n]$. In terms of Parceval's theorem, $\phi(i,j)$ is proved to be equal to the autocorrelation function $\tilde{r}_w[i-j]$ of which Fourier transform is equal to the warped and frequency-weighted power spectrum, $\left|\tilde{X}\left(e^{j\tilde{\lambda}}\right)\cdot\tilde{W}\left(e^{j\tilde{\lambda}}\right)\right|^2$ as

$$\begin{aligned}\phi(i,j) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\tilde{X}\left(e^{j\tilde{\lambda}}\right)\cdot\tilde{W}\left(e^{j\tilde{\lambda}}\right)\right|^2 \cos(i-j)\tilde{\lambda}\,d\tilde{\lambda} \\ &= \tilde{r}_w[i-j]\end{aligned} \tag{6}$$

where the weighting function $\left|\tilde{W}\left(e^{j\tilde{\lambda}}\right)\right|^2$ is a frequency derivative of the phase transfer function of $\tilde{z}^{-1}$, and is given by

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

Therefore, equation (4) becomes an autocorrelation equation as in the standard LPC analysis, and the estimated spectrum $\tilde{\sigma}_w / \tilde{A}_w(\tilde{z})$ represents the envelope of $\tilde{X}\left(e^{j\tilde{\lambda}}\right)\cdot\tilde{W}\left(e^{j\tilde{\lambda}}\right)$. Furthermore, 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;

$$\phi(i,j) = \tilde{r}_w[i-j] = \sum_{n=0}^{N-1} x[n]\cdot y_{(i-j)}[n] \tag{8}$$

where the output sequence $y_k[n]$ is given by

$$\begin{aligned} y_k[n] &= \alpha\cdot\left(y_k[n-1] - y_{(k-1)}[n]\right) + y_{(k-1)}[n-1] \\ &\qquad (n = 0,\ldots,N-1, \quad k = 1,\ldots,p)\end{aligned} \tag{9}$$

Therefore, due to the cost for computing N points of $y_k[n]$ for each i, the Mel-LPC analysis is accomplished with about two-fold increase in computation over the standard LPC analysis.

2.2. Control of Frequency Resolution

In the Mel-LPC analysis, the frequency resolution of spectrum can be changed easily by only one control factor. This factor is the warping factor of equation (2), and its range is from -1.0 to +1.0. Therefore it is probable to estimate the spectrum coefficients with the frequency resolution according to the value of the control factor. For example, the frequency resolution is improved on low frequency band if the control factor is plus and is improved on high frequency band if the control factor is minus. In this paper, the frequency resolution is controlled adaptively at each analysis frame by

$$\tilde{\alpha}_k = \gamma\cdot\hat{\beta}_i + \alpha 0\,. \tag{10}$$

where $\hat{\beta}_i$ is the function to represent the characteristics of input spectrum at i frame, and γ is the sensitivity factor for adjustment, and $\alpha 0$ is the bias to decide the analysis condition of whole system. For $\hat{\beta}_i$ several function can be use such as the function related the spectral slope, the formant, the kind of phoneme and so on. In this paper, $\hat{\beta}_i$ is the normalized autocorrelation coefficient that is given by

$$\hat{\beta}_i = \frac{r_i[1]}{r_i[0]}. \qquad (11)$$

where $r_i[1]$ is the first order autocorrelation coefficient at i frame, and $r_i[0]$ is the zero order autocorrelation coefficient (power) at i frame.

2.3. Compensation of emphasis and Frequency Warping

Although the warped predictors are estimated easily from input speech using the warping factor α of equation (2) in the Mel-LPC analysis, their predictors have two types of distortion on the estimated spectrum. The first spectral distortion is the pre-emphasis of weighting function $\tilde{W}\left(e^{j\tilde{\lambda}}\right)$ given by equation (7). However, the effect of the weighting function is completely compensated by filtering $\tilde{r}_w[m]$ with the second order FIR filter $\left[\tilde{W}(\tilde{z})\cdot\tilde{W}(\tilde{z}^{-1})\right]^{-1}$. Inverse weighting is computed by

$$\tilde{r}[m] = C_0\tilde{r}_w[m] + C_1\{\tilde{r}_w[m-1] + \tilde{r}_w[m+1]\} \qquad (12)$$

$$\text{where } C_0 = \frac{1+\tilde{\alpha}_i^2}{\sqrt{1-\tilde{\alpha}_i^2}} \text{ and } C_1 = \frac{\tilde{\alpha}_i}{\sqrt{1-\tilde{\alpha}_i^2}}. \qquad (13)$$

The resultant warped autocorrelation coefficients $\{\tilde{r}[m]\}$ lead to the warped predictors $\{\tilde{a}_k\}$.

The second spectral distortion is the characteristic of frequency warping. The spectrum obtained by he Mel-LPC analysis at each analysis frame are warped in the deference ratio on frequency axis. Therefore it is necessary to change spectrum into the warped spectrum with the same warping ratio on frequency axis regardless of analysis frame. In this paper, the spectrum warped in the deference ratio is compensated by Oppenheim's method [5]. In this method the predictors warped in the deference ratio are warped in the predictor's domain by equation (2), and the warping factor α of equation (2) is given by

$$\alpha_c = \frac{\alpha_T - \tilde{\alpha}_i}{1-\alpha_T\cdot\tilde{\alpha}_i} \qquad (14)$$

where $r_i[1]$ is the first order autocorrelation coefficient at i frame, and $r_i[0]$ is the zero order autocorrelation coefficient (power) at i frame.

3. Evaluation

3.1. Database and Speech Analysis

The recognition performance of mel-cepstral parameters obtained by the AMLPC analysis was compared with those of mel-cepstral parameters obtained by the conventional LPC analysis (LPC) and the Mel-LPC analysis (MLPC) through the phoneme and isolated word recognition tasks. In this study, we used 520 words uttered by each of 60 male speakers from the ATR C-set database. 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. The analysis order was set to 14 and the number of mel-cepstral coefficients was 14. A feature vector consists of both cepstral and delta-cepstral coefficients excluding the 0th terms (power terms). The warping factor of whole system is 0.41. In the phoneme and the isolated word recognition task, a set of gender-dependent phoneme HMMs were used a set of 35 phonemes including silence. The structure of HMMs is a left-to-right model with 3 emitting states, which consist of 4 gaussians for 35 phonemes including silence. Each phoneme model was trained using 520 words from each of 40 speakers for each gender. All the words from the other 20 speakers were 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 ($Acc = [(N-S-D-I)/N]\cdot 100\%$), where for N tokens, S, D, and I are substitution, deletion, and insertion errors, respectively.

3.2. Effect of Frequency Warping

In experiments, first, we evaluated the recognition performance in case of non-adaptive frequency resolution ($\gamma = 0.0$). In the AMLPC analysis method, the Mel-LPC analysis is a special case at $\gamma = 0.0$ and the conventional LPC analysis is a special case at $\gamma = 0.0$ and $\alpha 0 = 0.0$. Then, we examined the effect of the bias of warping factor $\alpha 0$ on phoneme and isolated word recognition accuracy. Four values of 0.0, 0.20, 0.41 and 0.60 were evaluated. Table 1 shows the phoneme and word recognition scores. This table shows that the best scores for phoneme and word recognition is obtained about 0.41. According to the mel-herz and Bark-herz transformations [7], [8], the mel frequency scale is approximated by 0.41. The result indicates the mel scale is suitable.

3.3. Effect of Adaptive Frequency Resolution

To examine the effect of adaptive frequency resolution, we found out the optimal bias $\alpha 0$ and the optimal sensitivity γ to control the frequency resolution. The phoneme recognition experiments were carried out when the bias is 0.0, 0.2 or 0.4

Table 1: Effect of frequency warping factor in AMLPC analysis ($\gamma = 0.0$) in phoneme and isolated word recognition.

Bias $\alpha 0$	0.0(LPC)	0.2	0.41	0.6
Phoneme Accuracy(%)	56.54	56.34	59.91	58.87
Word Accuracy(%)	89.49	90.01	91.35	91.18

Table 2: Effect of adaptation about frequency resolution in AMLPC analysis in isolated word recognition.

Bias $\alpha 0$	Sensitivity γ			
	0.05	0.1	0.2	0.4
0.4	91.57	91.89	91.61	91.30
0.3	91.18	91.20	91.61	91.27
0.2	91.08	91.18	91.30	91.22

and the sensitivity is 0.05, 0.1, 0.2 or 0.4. Table 2 shows the effect of adaptation about frequency resolution in the AMLPC analysis in isolated word recognition. The best score (91.89%) was obtained, when the bias was about 0.4 and the sensitivity was about 0.1. However, the sensitivity over 0.2 have the opposite effect to what was intended and lead to go down the recognition performance. Therefore, the AMLPC analysis was proved to be effective in speech recognition because of the spectrum coefficients with optimal frequency resolution according to the spectral slope or the formant and so on. The results showed that the proposed method (91.89%) led to a significant improvement in recognition accuracy over the conventional LPC analysis (89.49%), and a slightly improvement of error rate about 10% over the Mel-LPC analysis (91.35%).

4. Discussion

The AMLPC analysis led to an improvement in recognition accuracy over conventional LPC analysis and the Mel-LPC analysis. Then, to examine the ground of its performance, we compare each confusion matrix when the bias $\alpha 0$ is 0.0, 0.41 or 0.6 in the AMLPC analysis ($\gamma = 0.0$). Table 3 shows the number of error in case of typical error pattern in phoneme recognition. In this table, For example, in the case of the error from /o/ to /a/ or from /u/ to /e/, the number of error at $\alpha 0 = 0.6$ is smaller than the number of error at $\alpha 0 = 0.0$. While, in the case of the error from /u/ to /i/ or from /e/ to /i/, the number of error at $\alpha 0 = 0.0$ is smaller than the number of error at $\alpha 0 = 0.6$. These results indicated that the optimal frequency resolution was different according to the kind of phoneme. In the case of the error from /a/ to /o/, it might be required to improve the frequency resolution in lower frequency band. While, in the case of the error from /e/ to /i/, it might be required to improve the frequency resolution in higher frequency band. Therefore, it is reasonable to analysis by deference frequency resolution for each phoneme.

Table 3: Number of error in case of typical error pattern in the AMLPC analysis ($\gamma = 0.0$) in phoneme recognition.

Error Pattern	Number of Phoneme Recognition Error		
	$\alpha 0$ =0.0(LPC)	$\alpha 0$ =0.41	$\alpha 0$ =0.6
/o/ to /a/	91	72	65
/u/ to /e/	228	159	132
/u/ to /i/	92	142	196
/e/ to /i/	116	132	137

5. Conclusions

This paper has presented a new speech analysis method, an adaptive Mel-LPC (AMLPC) analysis method, using human auditory characteristics. In our proposed method, it is probable to estimate the spectrum coefficients with optimal frequency resolution according to the characteristics of the phoneme at each analysis frame. The proposed method has achieved a significant improvement in recognition accuracy over conventional LPC analysis, and a slightly improvement recognition accuracy than the Mel-LPC analysis.

In future work, it is necessary to develop an efficient method to control a spectral resolution and to evaluate the performance in noisy speech recognition, and in continuous speech recognition.

6. References

[1] H.Matsumoto, Y.Nakatoh, Y.Furuhata, "An efficient Mel-LPC analysis method for speech recognition," Proc. of ICSLP98, pp. 2591-2594 (1998)

[2] 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

[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. , 1996.

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

[6] H.W.Strube, "Linear prediction on a warped frequency scale," J.of Acoust.Soc.America, vol.68, no.4, pp.1071-1076, 1980.

[7] S.Senefi, "Pitch and spectral estimation of speech based on an auditory synchrony model," ICASSP84,(1984).

[8] J.Makhoul and L.Cosell, "LPCW: An LPC vocoder with linear predictive warping," Proc. of ICASSP76, pp.446-469 (1976).