社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2005－ＳＬＰ－59 (31)
2005／12／22

メルLPC分析に基づく音声認識フロントエンドのCENSRECによる評価

イスラムMD. バブル† 松本 弘† 山本 一公†

† 信州大学工学部 〒380-8553 長野市若里 4-17-1
E-mail: †{babul,matsu,kyama}@sp.shinshu-u.ac.jp

あらまし 我々は，加法性雑音環境下における音声のメルLPCスペクトルを強調する手法として，メル周波数軸上でのWienerフィルタを提案してきた．このフィルタは直線周波数スケール上の波形の最小二乗誤差基準によって推定され，波形上ではなく，メル自己相関領域で効率的に適用される．本稿では，このフィルタに対して雑音の自己回帰(AR)モデルによる雑音検出を適用する．提案手法をAURORA-2J(CENSREC-1)により評価し，単語正解精度がベースラインに対して58%改善されるという結果を得た．

キーワード 音声認識，雑音環境，メルLPC分析，Wienerフィルタ，CENSREC

Evaluation of Mel-LPC based front-end for CENSREC database

Md. BABUL ISLAM†, Hiroshi MATSUMOTO†, and Kazumasa YAMAMOTO†

† Faculty of Engineering, Shinshu University
4-17-1 Wakasato, Nagano-shi 380-8553 Japan
E-mail: †{babul,matsu,kyama}@sp.shinshu-u.ac.jp

Abstract We previously proposed a Mel-Wiener filter to enhance Mel-LPC spectra in presence of additive noise. The proposed filter was estimated based on minimization of sum of the square error on the linear frequency scale and then efficiently implemented in the autocorrelation domain without denoising input speech. In this paper the autoregressive (AR) model of noise is used to detect noise. The performance of the filter is evaluated on AURORA-2J (CENSREC-I) database. With the proposed Wiener filter the word accuracy is improved by about 58% relative to the baseline.

Key words Speech recognition, noisy environment, Mel-Wiener filter, Mel-LPC analysis, CENSREC

1. Introduction

As an LP-based method, we previously proposed a simple and efficient time-domain technique to estimate an all-pole model on the mel-frequency scale [1], [2] which is referred to as Mel-LPC. This paper presents previously proposed Mel-Wiener filter [3] with some improvements incorporated with the Mel-LPC for noise robust speech recognition. The performance is evaluated on AURORA-2J.

2. Mel-Wiener Filter

In [3], we defined a Mel-Wiener filter on z domain by

$$\tilde{H}_w(\tilde{z}(z)) = \sum_{k=0}^{p-1} \tilde{h}_w[n]\tilde{z}^{-n} \qquad (1)$$

where $\tilde{z}(z)$ is the transfer function of the first order all-pass filter.

Filter estimation was based on the minimization of the sum of the square error which gives following normal equations

$$\sum_{k=0}^{p-1} \tilde{\phi}_{xx}(m,k)\tilde{h}_w(k) = \tilde{\phi}_{sx}(0,m) \quad (m=0,\ldots,p-1), \qquad (2)$$

where
$$\tilde{\phi}_{xx}(m,k) = \sum_{n=0}^{\infty} x_m[n]x_k[n], \quad \text{and} \qquad (3)$$

$$\tilde{\phi}_{sx}(m,k) = \sum_{n=0}^{\infty} s_m[n]x_k[n] \qquad (4)$$

$\tilde{\phi}_{xx}(m,k)$ and $\tilde{\phi}_{sx}(m,k)$ can be calculated from sum of finite terms and hence, reduce to generalized auto and crosscorrelation functions $\tilde{r}_{xx}[m]$ and $\tilde{r}_{sx}[m]$, respectively.

The crosscorrelation function between clean and noisy speech is approximated as

$$\tilde{r}_{sx}[m] \approx \tilde{r}_{xx}[m] - s \cdot \tilde{r}_{nn}[m] \qquad (5)$$

where s is a scaling factor, given by

$$s = \begin{cases} s'(0.9\tilde{r}_{xx}[0]/\tilde{r}_{nn}[0]) & \text{if } \tilde{r}_{xx}[0] - s' \cdot \tilde{r}_{nn}[0] \leqq 0 \\ s' & \text{otherwise} \end{cases} \qquad (6)$$

and the value of s' is 1.75 for noise frame, otherwise 1.

3. Application to Mel-LPC Analysis

To estimate the Mel-LPC cepstrum, the generalized autocorrelation function of the filtered speech $\hat{s}_w[n]$ is required, which is given by

$$\hat{\tilde{r}}_{ss}[m] = \sum_{n=0}^{\infty} \hat{s}_w[n]\hat{s}_{w,m}[n] = \sum_{k=-p+1}^{p-1} r_{\tilde{h}\tilde{h}}[k]\tilde{r}_{xx}[m-k] \quad (7)$$

where $r_{\tilde{h}\tilde{h}}[k]$ is the autocorrelation function of $\tilde{h}_w[m]$.

4. Noise Estimation

Initial 40 frames are used to create a noise model based on autoregressive (AR) model [4], i.e., the model is assumed to be Mth order autoregressive with coefficients $\mathbf{a}^t = [a_0 a_1 \ldots a_M]$, where $a_0 = 1$. Now, for frame t, $\tilde{r}_{xx}[m,t]$ is calculated to estimate the likelihood ratio (LR) as follows:

$$LR[t] = r_a[0]\tilde{r}_{xx}[0] + 2\sum_{i=1}^{M} r_a[i]\tilde{r}_{xx}[i] - 1 \quad (8)$$

where $r_a[i]$ is the autocorrelation function of AR coefficients.

Finally, this ratio is compared with a threshold value, η. For $LR[t] < \eta$, the frame t is detected as noise, otherwise, speech frame. When frame t is detected as noise, noise model is updated by accumulating $\tilde{r}_{xx}[m,t]$ to $\hat{\tilde{r}}_{nn}[m]$ as follows:

$$\hat{\tilde{r}}_{nn}[m,t] = \begin{cases} \beta\hat{\tilde{r}}_{nn}[m,t_p] + (1-\beta)\tilde{r}_{xx}[m,t]; \\ \qquad\qquad \text{if frame } t \text{ is noise} \\ \hat{\tilde{r}}_{nn}[m,t_p]; \qquad \text{if frame } t \text{ is speech} \end{cases} \quad (9)$$

where t_p is the previous noise frame and β is the forgetting factor with value of 0.96. Before accumulating, $\tilde{r}_{xx}[m,t]$ is smoothed using a lag window with a length of 50.

5. Experiments

5.1 Experimental Setup

Experiment was performed using AURORA-2J (CENSREC-1) database. The order of Mel-LPC, window length, frame shift, preemphasis and warping factors are 12, 25ms, 10ms, 0.0 and 0.4, respectively. The HMM was trained on clean condition with 16 states per word and mixture of 20 Gaussians. A feature vector consists of 14 mel-cepstral coefficients and their delta and acceleration cepstrums including 0^{th} terms. Thus the category was 5.

5.2 Recognition Results

From Fig.1, it has been shown that the highest word accuracy is attained at the order of 5. Therefore, the order of filter is set to 5 in this experiment.

The recognition accuracies for baseline and for Mel-LPC based front-end, without and with Wiener filter are given in Table 1, 2 and 3, respectively. The average recognition accuracies with proposed model are 72.1%, 67.3% and 70.2% for sets A, B and C, respectively.

6. Conclusion

It has been shown that the Mel-Wiener filter incorporating with the Mel-LPC can be used as front-end to improve the recognition accuracy. As a result of recognition experiments,

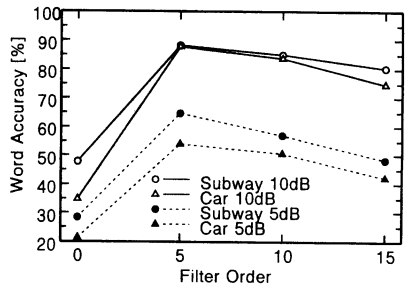

Fig. 1 *Recognition accuracy as a function of filter order.*

Table 1 Recognition accuracy for baseline.

	SNR (dB)							Avg
	cln	20	15	10	5	0	-5	
Set A	99.7	90.8	66.7	41.6	22.8	10.7	6.7	46.5
Set B	99.7	84.5	59.8	40.8	23.3	11.5	5.2	44.0
Set C	99.8	91.9	73.1	45.4	25.5	13.7	8.8	49.9
Overall Avg	99.7	88.5	65.2	42.0	23.5	11.6	6.5	46.2

Table 2 Recognition accuracy without Wiener filter.

	SNR (dB)							Avg
	cln	20	15	10	5	0	-5	
Set A	99.6	82.9	59.9	39.7	23.4	10.9	6.2	43.4
Set B	99.6	73.7	51.6	35.2	19.8	8.9	2.9	37.78
Set C	99.6	93.3	77.3	49.2	28.1	16.4	8.8	52.9
Overall Avg	99.6	81.3	60.1	39.8	22.9	11.1	5.4	43.0

Table 3 Recognition accuracy with Wiener filter.

Noise	SNR (dB)							Avg
	cln	20	15	10	5	0	-5	
Subway	99.6	99.1	97.2	88.3	64.7	30.4	10.8	75.9
Babble	99.6	98.7	94.7	79.7	47.5	10.3	-2.7	66.2
Car	99.9	99.5	97.9	87.8	54.1	17.6	3.3	71.4
Exhibition	99.8	98.7	95.8	85.3	62.9	31.9	10.9	74.9
Average	99.7	99.0	96.4	85.3	57.3	22.5	5.6	72.1
Restaurant	99.6	97.9	93.1	77.0	39.7	6.1	-5.7	62.8
Street	99.6	98.4	92.8	76.3	45.2	16.4	4.6	65.8
Airport	99.9	98.6	94.5	80.9	52.5	15.9	0.7	68.5
T-Station	99.8	98.9	96.1	84.4	58.1	22.2	3.9	71.9
Average	99.7	98.5	94.1	79.6	48.9	15.1	0.9	67.3
Subway	99.7	98.8	96.4	85.9	61.7	31.7	12.3	74.9
Street	99.5	98.4	94.3	80.7	52.5	23.4	7.2	69.9
Average	99.6	98.6	95.4	83.3	57.1	27.5	9.7	72.4
Overall Avg	**99.7**	**98.7**	**95.3**	**82.6**	**53.9**	**20.6**	**4.5**	**70.2**

Table 4 Relative improvements with respect to baseline.

	SNR (dB)							Avg
	cln	20	15	10	5	0	-5	
Set A	15.0	81.6	88.5	75	44.8	13.5	-1.2	60.7
Set B	15.0	89.6	85.0	65.6	33.4	4.1	-4.6	55.5
Set C	-59.7	82.5	82.4	69.3	42.5	16.2	1.1	58.6
Overall	**0.1**	**85.0**	**85.9**	**70.1**	**39.8**	**10.3**	**-2.1**	**58.2**

it is found that the optimum filter order is 5, and accuracy is improved by about 58% relative to the baseline.

7. Acknowledgement

The present study was conducted using AURORA-2J database developed by IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

References

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

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

[3] M. B. Islam, H. Matsumoto, K. Yamamoto, "An improved Mel-Wiener filtering for Mel-LPC based speech recognition," *IEICE Technical Report, Japan*, vol.105, no.199, pp.1–6, 2005.

[4] B. Juang, "On the hidden Markov model and dynamic time warping for speech recognition - a unified view," *AT&T Bell Laboratories Technical Journal*, vol.63, no.7, pp.1213–1243, 1984.