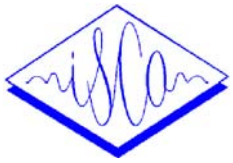

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

USING ADAPTIVE SIGNAL LIMITER TOGETHER WITH NOISE-ROBUST TECHNIQUES FOR NOISY SPEECH RECOGNITION

Wei-Wen Hung and Hsiao-Chuan Wang

Department of Electrical Engineering, National Tsing Hua University,
Hsinchu, Taiwan, 300, ROC
E-mail : hcwang@ee.nthu.edu.tw

10.21437/Eurospeech.1999-544

ABSTRACT

In a speech recognition system, environmental mismatch between speech models and test speech causes serious performance degradation. To solve this environmental mismatch problem, smoothing process is one of the most widely used techniques. In this paper, an adaptive signal limiter (ASL) is developed to smooth speech features so that the undesired spectral variations could be effectively reduced. In addition, we propose the method for combining ASL with other noise-robust techniques in the recognition of noisy speech.

1. INTRODUCTION

Spectral analysis is one of the most effective and frequently used methods for speech signal processing. In speech recognition, the basic technique is based on comparing the spectra of test signal and reference models. However, the spectrum variation may exist due to some factors, such as inherent constraints of analysis model, speaker's Lombard effect, changes in speech fundamental frequency, background noise and channel distortion, etc. This variation will result in unreliable spectral comparison and cause the degradation of recognition performance. In order to reduce the spectral variation and alleviate its impact to the performance of a speech recognizer, several smoothing techniques for the speech signal or speech features have been proposed and proved to be effective [1]-[4].

In general, smoothing techniques can be roughly classified into three categories according to their application domains. The first one is in waveform domain. Rabiner et al. [1] used a combination of median and linear smoother to nonlinearly smooth speech signals. This combined smoother is not only capable of preserving sharp discontinuities in speech signals, but also able to filter out additive noise superimposed on speech data. The second approach is in cepstral domain. Juang and Rabiner [2] showed that the variation of higher quefrency terms in cepstral domain is due to inherent artifacts of the analysis procedure. In contrast, the variation of lower quefrency terms is primarily due to variations in transmission channel, vocal tract and speaker's characteristics. Based upon these observations, a liftering function was used to incorporate into the traditional cepstral features so as to normalize the contributions from each cepstral term. This liftering procedure can be viewed as a kind of smoothing technique applied in the cepstral domain. In the third approach, a signal limiter with fixed smoothing factor (i.e., a hard limiter) is first used by Lee and Lin [3] to reduce the variation of speech features in noisy condition. A signal limiting operation is equivalent to performing an arcsin transformation on the autocorrelation domain of original speech signal. Experimental results for the recognition of 39-word alpha-digit vocabulary demonstrated that an equivalent gain of 5-7 dB in SNR could be achieved for a template-based DTW recognizer.

However, from those experiments, we also found that the recognition accuracy using a hard limiter for clean speech is relatively low. This is mainly due to that heavily smoothing can reduce feature variation of the speech segments with lower SNRs, but it also causes the loss of some important information embedded in the clean segments and the segments with higher SNRs. Therefore, a signal limiter with fixed smoothing factor might not work well for the all segments in a speech utterance. For this reason, we developed an adaptive signal limiter (ASL) [4] to further improve the noise-robustness of a signal limiter. In the proposed ASL method, the smoothing factor of a signal limiter is related to SNR value and adapted on a frame by frame basis.

In this paper, we try to extend the applications of signal limiters and propose a scheme to combine the adaptive signal limiter with a general noise-robust technique (e.g., weighted projection measure, WPM). This scheme is very straightforward and can provide a good compromise between the ASL and WPM methods.

2. ADAPTIVE SIGNAL LIMITER

The smoothing process of an adaptive signal limiter is

as follows [4]. Consider a continuous density hidden Markov model (CDHMM), the output likelihood measure of $t-th$ frame in the testing utterance $Y=\{y_t=[c_t,d_t],1\le t\le T_y\}$ based on the statistics of $i-th$ state of word model $\Lambda(w)=\{\Lambda_{w,i}=(\mu_{w,i},\Sigma_{w,i}),1\le i\le S_w\}$ can be formulated by a Gaussian probability density function (pdf) and expressed as

$$p(y_t|\Lambda_{w,i})=(2\cdot\pi)^{-p}\cdot|\Sigma_{w,i}|^{-1/2}\cdot\exp\left\{-\frac{1}{2}\cdot(y_t-\mu_{w,i})^T\cdot\Sigma_{w,i}^{-1}\cdot(y_t-\mu_{w,i})\right\}, \quad (1)$$

where $\mu_{w,i}=[c_{w,i},d_{w,i}]$ denotes the mean vector of $i-th$ state of word model $\Lambda(w)$ and consists of cepstral vector $c_{w,i}$ and delta cepstral vector $d_{w,i}$. $\Sigma_{w,i}$ is the covariance matrix of $i-th$ state of word model $\Lambda(w)$. The mean vector $\mu_{w,i}=[c_{w,i},d_{w,i}]$ of $i-th$ state of the word model $\Lambda(w)$ is indirectly represented by the normalized autocorrelation vectors of a five-frame context window [5]. This context window is expressed as $[r_{w,i,-2},r_{w,i,-1},r_{w,i,0},r_{w,i,1},r_{w,i,2}]$, and the component $r_{w,i,j}=[r_{w,i,j}(1),\cdots,r_{w,i,j}(p)]^T$, $j=0$ denotes the instantaneous frame. $j=-1,-2$ and $j=1,2$ denote the left context and the right context frames, respectively.

When the $t-th$ frame of a testing utterance Y is evaluated on the state $\Lambda_{w,i}$, the cepstral vectors $c_{t,j}$ of its context frames $y_{t,j}$, for $-2\le j\le 2$, are transformed to give the corresponding autocorrelation vectors $r_{t,j}$. Then, these normalized autocorrelation vectors $r_{t,j}=[r_{t,j}(1),\cdots,r_{t,j}(p)]^T$ are converted by the arcsin transformation

$$\tilde{r}_{t,j}(t)=\frac{\sin^{-1}\left\{\frac{r_{t,j}(t)}{\left[1+\delta(SNR_{t,j})\right]}\right\}}{\sin^{-1}\left\{\frac{1}{\left[1+\delta(SNR_{t,j})\right]}\right\}}, \quad (2)$$

for $-2\le j\le 2$ and $1\le t\le p$.

In this equation, the smoothing factor $\delta(SNR_{t,j})$ is empirically formulated as

$$\delta(SNR_{t,j})=\begin{cases}\delta_{\min} & if \quad SNR_{t,j}<SNR_{LB}\\ \left(\frac{\delta_{\max}-\delta_{\min}}{SNR_{UB}-SNR_{LB}}\right)\cdot\left(SNR_{t,j}-SNR_{LB}\right)+\delta_{\min} & \\ & if \quad SNR_{LB}\le SNR_{t,j}\le SNR_{UB}\\ \delta_{\max} & if \quad SNR_{t,j}>SNR_{UB}\end{cases}, \quad (3)$$

and $SNR_{t,j}$ is determined by

$$SNR_{t,j}=10\cdot\log_{10}\left\{\frac{(E_{t-j}-E_n)}{E_n}\right\}, \quad (4)$$

where E_t is the $t-th$ frame energy in the testing utterance Y, and E_n is the noise energy. For simplicity, E_n is obtained by $E_n=\min\{E_1,E_2,\cdots,E_{T_y}\}$. Once the autocorrelation vectors $\tilde{r}_{t,j}$, $for -2\le j\le 2$, are obtained, the smoothed testing cepstral vector $\tilde{c}_{t,j}$ of $\tilde{y}_{t,j}$ can be calculated. The corresponding smoothed testing delta cepstral vector $\tilde{d}_t$ can also be calculated,

$$\tilde{d}_t=\frac{\sum_{j=-2}^{j=2}j\cdot\tilde{c}_{t,j}}{\sum_{j=-2}^{j=2}j^2}. \quad (5)$$

Similarly, in order to avoid the mismatch between testing speech signal and reference model, the mean vector of state model $\Lambda_{w,i}$ is also smoothed by using Eq. (2) with the same smoothing factor so that a smoothed version $\tilde{\mu}_{w,i}=[\tilde{c}_{w,i},\tilde{d}_{w,i}]$ is obtained. In addition, the corresponding smoothed covariance matrix $\tilde{\Sigma}_{w,i}$ can also be obtained by means of maximum likelihood (ML) estimation. Finally, the smoothed output likelihood measure can be expressed by

$$\tilde{p}(\tilde{y}_t \left| \tilde{\Lambda}_{w,i} \right.) = (2 \cdot \pi)^{-p} \cdot \left| \tilde{\Sigma}_{w,i} \right|^{-1/2} \cdot \exp\left\{ -\frac{1}{2} \cdot (\tilde{y}_t - \tilde{\mu}_{w,i})^T \cdot \tilde{\Sigma}_{w,i}^{-1} \cdot (\tilde{y}_t - \tilde{\mu}_{w,i}) \right\} \quad (6)$$

3. COMBINATION OF ASL AND WPM

A signal limiter has the advantage that it can effectively reduce the variation of speech features when the SNR value is relatively low. However, smoothing in autocorrelation domain of the original speech will also inevitably suppress parts of formant peaks and reduce the discriminability of speech features. Especially, this suppression effect is harmful for speech recognition in clean condition and higher SNR.

On the other hand, most of the noise-robust techniques, e.g., the weighted projection measure (WPM), are more adequate and effective for the cases of higher and medium SNRs. It is obvious that the merits of signal limitation processing and WPM method are complementary. Based upon this observation, an interpolation scheme is used to combine these two adaptation schemes, adaptive signal limiter and WPM. Let $p(Y|\Lambda(w))_{WPM}$ and $p(Y|\Lambda(w))_{ASL}$ denote the accumulated likelihood scores calculated by using the WPM and ASL methods, respectively. Then, the resulting likelihood score can be expressed as

$$p(Y|\Lambda(w)) = \alpha \cdot p(Y|\Lambda(w))_{ASL} + (1.0 - \alpha) \cdot p(Y|\Lambda(w))_{WPM}, \quad (7)$$

where α is the interpolation factor.

4. EXPERIMENTS AND DISCUSSIONS

A task of multispeaker isolated Mandarin digit recognition was conducted to demonstrate the effectiveness of the proposed scheme. The speech data were collected from 50 male and 50 female speakers. There were three sessions of data collection. For each session, a speaker uttered 10 Mandarin digits. The first two sessions were used for training the word models and the other for testing. Each digit is modeled as a left-to-right HMM without jumps. The output of each state in HMM is a mixture of two Gaussian densities of feature vectors where each feature vector consists of 12 LPC-derived cepstral coefficients and 12 delta cepstral coefficients. Also, a conventional hidden Markov model (HMM) without incorporating signal limiters is referred as a baseline for comparison.

In order to show the effectiveness of adaptive signal limiter, a sample utterance of Mandarin digit '1' uttered by a male speaker is used for illustrating the Log LPC-derived spectrum. The 12-order LPC spectrum analysis is performed on a 32-msec window with 16-msec frame shift. The Log LPC spectrum of the clean speech '1' is depicted in Figure 1. From this figure, we can see that the formants of utterance '1' occur at the positions of 200, 1950, 3100 and 3350 Hz. When the speech is artificially contaminated by 10 dB white noise, as shown in Figure 2, the second, third and fourth formants are severely suppressed. If an adaptive signal limiter is applied, we can retain the formants in some degree even the higher formants are still suppressed. This is shown in Figure 3.

For speech recognition, the effectiveness of the proposed scheme that combines ASL and WPM methods is demonstrated in Table 1. The best result comes from the combination factor equal to 0.3. In Table 2, we can find that both ASL and WPM improve the performance in noisy environment. The improvement due to ASL is more in lower SNR while WPM is in higher SNR. When this two methods are combined, we can obtain much better result.

5. CONCLUSIONS

In this paper, an SNR-dependent signal limiter is proposed to adaptively smooth speech features. This smoothing technique is proved to be more feasible for noisy speech recognition. In addition, using WPM as an example, we also show that the ASL and general noise-robust techniques can be properly combined to achieve higher recognition rates

ACKNOWLEDGEMENT

This research has been partially sponsored by the National Science Council, Taiwan, ROC under contract NSC-88-2614-E-007-002.

REFERENCES

[1] L. R. Rabiner, M. R. Sambur and C. E. Schmidt, "Applications of a nonlinear smoothing algorithm to speech processing," *IEEE Trans. ASSP*, vol. 23, No. 6, pp. 552-557, December, 1975.

[2] B. H. Juang and L. R. Rabiner," On the use of bandpass liftering in speech recognition," *IEEE Trans. ASSP*, vol. 35, No. 7, pp. 947-954, July, 1987.

[3] C. H. Lee and C. H. Lin, " On the use of a family of signal limiters for recognition of noisy speech," *Speech Communication*, Vol. 12, pp. 383-392, 1993.

[4] W. W. Hung and H. C. Wang, "Smoothing hidden Markov models by using an Adaptive signal limiter for noisy speech recognition," to appear in *Speech Communication*, 1999.

[5] L. M. Lee and H. C. Wang, "Representation of hidden Markov model for noise adaptive speech

recognition," *IEE Electronics Letters*, Vol. 31, No. 8, pp. 616-617, 1995.

Table 1. Recognition rates (%) for various α values.

($\delta_{min} = 0.0, \delta_{max} = 1.0, SNR_{LB} = 20dB, SNR_{UB} = 30dB.$)

α \ SNRs	clean	20dB	15dB	10dB	5dB	0dB
0.0 (WPM)	97.2	92.2	84.8	69.5	52.5	30.5
0.1	97.3	93.3	86.8	74.6	59.7	38.7
0.2	97.7	93.6	88.3	76.6	62.9	46.2
0.3	98.0	93.3	88.9	77.0	64.4	49.5
0.4	97.9	93.0	87.8	76.6	64.8	50.9
0.5	97.6	92.7	87.0	76.3	65.4	52.8
0.6	97.5	91.7	85.5	75.0	66.4	54.0
0.7	97.0	91.0	84.6	74.1	65.5	54.5
0.8	96.8	89.3	82.2	73.0	64.1	54.0
0.9	96.1	87.6	79.5	71.5	62.4	52.5
1.0 (ASL)	95.2	85.1	76.4	68.1	58.5	49.7

Table 2. Comparison of digit recognition rates (%)

Methods \ SNRs	clean	20dB	15dB	10dB	5dB	0dB
Baseline	98.9	80.2	65.7	48.8	25.6	10.6
WPM	97.2	92.2	84.8	69.5	52.5	30.5
ASL	95.2	85.1	76.4	68.1	58.5	49.7
ASL+WPM (α = 0.3)	98.0	93.3	88.9	77.0	64.4	49.5

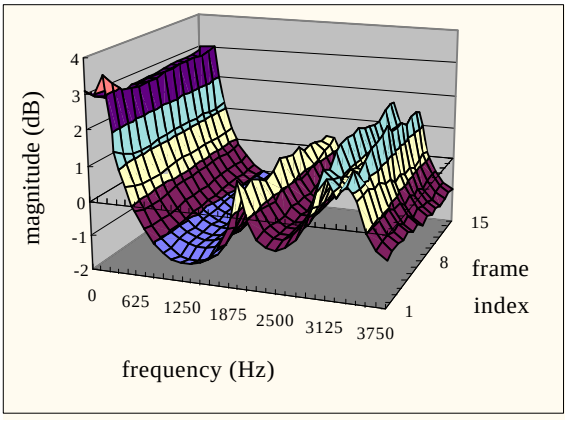

Fig. 1. Log spectra of clean speech '1'

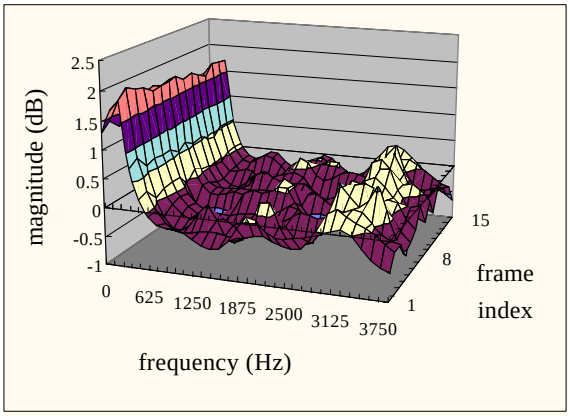

Fig. 2. Log spectra of noisy speech

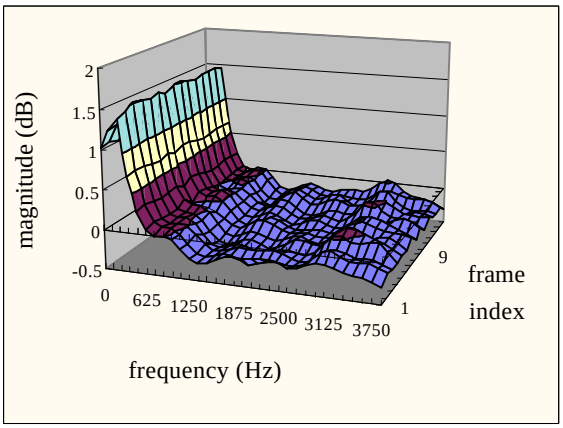

Fig. 3. Log spectra of noisy speech with ASL