USING ADAPTIVE SIGNAL LIMITER TOGETHER WITH WEIGHTING TECHNIQUES FOR NOISY SPEECH RECOGNITION

Wei-Wen Hung

Department of Electrical Engineering, Ming Chi Institute of Technology, Taishan, Taipei, Taiwan 243, R.O.C.
E-mail : wwhung@ccsun.mit.edu.tw

ABSTRACT

In an automatic speech recognition (ASR) system, environmental mismatch between speech models and testing speech utterances causes serious performance degradation. To alleviate this environmental mismatch problem, smoothing process and weighting technique are two of the most widely used methods. In this paper, an investigation is presented into the feasibility of combining an adaptive signal limiter with a weighting technique for seeking further performance improvement in noisy speech recognition.

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 signals 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 serious degradation in recognition performance. In order to reduce the spectral variation and alleviate its impact to the performance of a speech recognizer, several weighting techniques and smoothing techniques for speech signals or features have been proposed and proved to be effective.

Juang and Rabiner [1] 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 weighting process applied in the cepstral domain. In addition, Tetsuya and Hajime [2] found that invalid peaks of autocorrelation function easily cause a half or double pitch error. To alleviate this problem, a new approach using weighted autocorrelation function was proposed to improve the accuracy of pitch extraction.

Besides above weighting techniques, Rabiner et al. [3] used a combination of median and linear smoother to nonlinearily 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. Moreover, a signal limiter with fixed smoothing factor (i.e., a hard limiter) was used by Lee and Lin [4] to smooth the variation of speech features in noisy conditions. A signal limiting operation is equivalent to performing an arcsin transform on the autocorrelation domain of original speech signals. This hard limiter has the drawback of that the recognition accuracy 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. Based upon this reason, Hung and Wang [5] developed an adaptive signal limiter (ASL) 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, an investigation is presented into the feasibility of combining an adaptive signal limiter with a weighting technique for seeking further performance improvement. By applying both smoothing process and weighting technique on the autocorrelation domain of a speech signal, experimental results demonstrate that this new approach can achieve higher accuracy for noisy speech recognition.

2. SMOOTHING PROCESS BASED ON ASL

The smoothing process applied on the autocorrelation domain of a speech signal using an adaptive signal limiter is described as follows [5]. Consider a continuous density hidden Markov model (CDHMM), the output
likelihood measure of \( t-th \) frame in the testing utterance \( Y = \{y_t = [c_t, d_t], 1 \leq t \leq T_y \} \) based on the statistics of \( i-th \) state of word model \( \Lambda(w) = \{\Lambda_{w,i} = (\mu_{w,i}, \Sigma_{w,i}), 1 \leq i \leq S_w \} \) can be formulated by a Gaussian probability density function (pdf) and expressed as

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

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 \( \Lambda(w) \), and \( P \) is the order of speech feature vectors. 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 [6]. This context window is expressed as \( r_{w,i,j} = [r_{w,i,j}(1), \cdots, r_{w,i,j}(p)]^T \),

where \( j = 0 \) 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 \( \Lambda_{w,i,j} \), the cepstral vectors \( c_{w,i,j} \) of its context frames \( y_{i,j} \), for \( -2 \leq j \leq 2 \), are transformed to give the corresponding autocorrelation vectors \( r_{i,j} \). Then, these normalized autocorrelation vectors \( r_{i,j} = [r_{i,j}(1), \cdots, r_{i,j}(p)]^T \) are converted by the arcsin transformation

\[
\tilde{r}_{i,j}(\tau) = \sin^{-1} \left( \frac{r_{i,j}(\tau)}{1 + \delta(\text{SNR}_{t,j})} \right)
\]

for \(-2 \leq j \leq 2, 1 \leq \tau \leq p \) and \( \tilde{r}_{i,j}(\tau) \) is the smoothed version of \( r_{i,j}(\tau) \). In above equation, the smoothing factor \( \delta(\text{SNR}_{t,j}) \) is empirically formulated as

\[
\delta(\text{SNR}_{t,j}) = \begin{cases} 
\delta_{\min} & \text{if } \text{SNR}_{t,j} < \text{SNR}_{LB} \\
\frac{\delta_{\max} - \delta_{\min}}{\text{SNR}_{UB} - \text{SNR}_{LB}} (\text{SNR}_{t,j} - \text{SNR}_{LB}) + \delta_{\min} & \text{if } \text{SNR}_{LB} \leq \text{SNR}_{t,j} \leq \text{SNR}_{UB} \\
\delta_{\max} & \text{if } \text{SNR}_{t,j} > \text{SNR}_{UB}
\end{cases}
\]

and \( \text{SNR}_{t,j} \) is determined by

\[
\text{SNR}_{t,j} = 10 \cdot \log_{10} \left( \frac{(E_{i+j} - E_n)}{E_n} \right),
\]

where \( \delta_{\min} \), \( \delta_{\max} \), \( \text{SNR}_{LB} \), \( \text{SNR}_{UB} \) are tuning constants. \( E_i \) is the \( t-th \) frame energy in the testing utterance \( Y \), and \( E_n \) is the noise energy. For simplicity, the noise energy \( E_n \) is roughly obtained by \( E_n = \min\{E_1, E_2, \cdots, E_T\} \).

3. WEIGHTING PROCESS BASED ON AMDF

The autocorrelation function \( r_{i,j}(\tau) \) in Eq.(2) is generally calculated by

\[
r_{i,j}(\tau) = \frac{1}{N} \sum_{n=1}^{N} S_{i,j}(n) \cdot S_{i,j}(n + \tau)
\]

where \( S_{i,j}(n) \) represents the speech segment associated with the context frame \( y_{i,j} \), and \( N \) denotes the number of samples in the speech segment. The characteristic of \( r_{i,j}(\tau) \) is that it attains its maximum value at \( \tau = 0 \). If \( S_{i,j}(n) \) has a period of \( T \), then \( r_{i,j}(\tau) \) has peaks at \( \tau = n \cdot T \), where \( n \) is an integer. Essentially, \( r_{i,j}(\tau = 0) \) gives the largest value among \( r_{i,j}(\tau = n \cdot T) \), for \( n = 0, 1, 2, \cdots \). Other peaks of \( r_{i,j}(\tau) \) usually decrease as \( \tau \) increases. When the speech segment \( S_{i,j}(n) \) is contaminated by background noise, it is possible that some abnormal peaks may occur and result in a half-pitch or double-pitch error. For such reasons, it is necessary to emphasize the true peak of autocorrelation function by properly weighting the components in \( r_{i,j}(\tau) \). In this paper, the weights used for emphasizing the components
of the smoothed autocorrelation function \( \tilde{r}_{i,j}(\tau) \) are proportional to the reciprocal of the average magnitude difference function (AMDF) that is formulated as

\[
\phi_{i,j}(\tau) = \frac{1}{N} \sum_{n=1}^{N} |S_{i,j}(n) - S_{i,j}(n + \tau)|. \tag{6}
\]

The AMDF has the characteristic that when \( S_{i,j}(n) \) is similar with \( S_{i,j}(n + \tau) \), \( \phi_{i,j}(\tau) \) becomes small. This means that if \( S_{i,j}(n) \) has a period of \( T \), \( \phi_{i,j}(\tau) \) makes a peak at \( \tau = T \). Furthermore, the background noise included in \( \phi_{i,j}(\tau) \) behaves independently with that included in the \( r_{i,j}(\tau) \) [2]. Hence, using the weights proportional to the inversed AMDF, it can be expected that the true peak is emphasized. Once the arcsin transformation using Eq.(2) is performed, the smoothed autocorrelation vectors \( \tilde{r}_{i,j} \) is weighted by

\[
\hat{r}_{i,j}(\tau) = \frac{[\phi_{i,j}(\tau) + 1]^{-1}}{\sum_{\alpha=1}^{P} |[\phi_{i,j}(\alpha) + 1]^{-1}|} \cdot \tilde{r}_{i,j}(\tau) \tag{7}
\]

and then the smoothed and weighted testing cepstral vector \( \hat{c}_{t,j} \) of \( \hat{y}_{t,j} \) can be calculated. The corresponding smoothed testing delta cepstral vector \( \hat{d}_{i} \) can also be calculated as

\[
\hat{d}_{i} = \frac{2}{\sum_{j=1}^{2} j \cdot \hat{c}_{t,j}}. \tag{8}
\]

Similarly, in order to avoid the mismatch between testing speech signal and reference model, the mean vector of state model \( \mu_{w,j} \) is also smoothed and weighted by using Eq. (2) and Eq. (7) with the same manner so that a smoothed and weighted version \( \hat{\mu}_{w,j} = [\hat{c}_{w,j}, \hat{d}_{w,j}] \) is obtained. In addition, the corresponding smoothed and weighted covariance matrix \( \hat{\Sigma}_{w,j} \) can also be obtained by means of maximum likelihood (ML) estimation. Finally, the smoothed and weighted output likelihood measure can be expressed by

\[
\hat{p}(\hat{y}_{t} | \hat{\lambda}_{w,j}) = (2 \cdot \pi)^{-p} \cdot \left| \hat{\Sigma}_{w,j} \right|^{-1/2} \cdot \exp \left\{ -\frac{1}{2} \cdot (\hat{y}_{t} - \hat{\mu}_{w,j})^{T} \cdot \hat{\Sigma}_{w,j}^{-1} \cdot (\hat{y}_{t} - \hat{\mu}_{w,j}) \right\}. \tag{9}
\]

4. EXPERIMENTS AND DISCUSSIONS

A task of multi-speaker 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. The white Gaussian noise is artificially generated by computer and added to clean speech with specific signal-to-noise ratio (SNR) values.

In Table 1, we demonstrate the digit recognition rates for the proposed method, the baseline and the adaptive signal limiter under different noisy conditions. From the experimental results, we can find that the smoothing process based on ASL provides a significant accuracy improvement at low SNRs relative to the baseline. Moreover, with the aids of incorporating weighting technique into the calculation of autocorrelation function, the performance of ASL can be further enhanced.

5. CONCLUSIONS

In this paper, we first review the basic formulation of the smoothing process using adaptive signal limiter (ASL). In order to achieve better performance for the ASL, a set of weights proportional to the reciprocal of the average magnitude difference function (AMDF) is incorporated into the weighting process of the smoothed autocorrelation function. Based on the experimental results, it has been shown that the proposed weighting technique is feasible to combine with the ASL and provides a moderate accuracy improvement relative to the ASL at different noisy conditions.

ACKNOWLEDGEMENT

This research has been partially sponsored by the National Science Council, Taiwan, R.O.C under contract NSC-90-2213-E-131-001.
REFERENCES


Table 1. Comparison of digit recognition rates (%) under different noisy conditions.

\( (\delta_{\text{min}} = 0.0, \delta_{\text{max}} = 1.0, SNR_{L_B} = 20\text{dB}, SNR_{U_B} = 30\text{dB}) \)

<table>
<thead>
<tr>
<th>Methods</th>
<th>SNRs</th>
<th>clean</th>
<th>20dB</th>
<th>15dB</th>
<th>10dB</th>
<th>5dB</th>
<th>0dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>98.9</td>
<td>80.2</td>
<td>65.7</td>
<td>48.8</td>
<td>25.6</td>
<td>10.6</td>
</tr>
<tr>
<td>ASL</td>
<td></td>
<td>95.2</td>
<td>85.1</td>
<td>76.4</td>
<td>68.1</td>
<td>58.5</td>
<td>49.7</td>
</tr>
<tr>
<td>Proposed method</td>
<td></td>
<td>97.8</td>
<td>89.0</td>
<td>81.7</td>
<td>72.6</td>
<td>61.8</td>
<td>52.1</td>
</tr>
</tbody>
</table>