NOISY SPEECH RECOGNITION USING VARIANCE ADAPTED LIKELIHOOD MEASURE

Jen-Tzung Chien, Lee-Min Lee and Hsiao-Chuan Wang

Department of Electrical Engineering,
National Tsing Hua University, Hsinchu, Taiwan

ABSTRACT

Because the norm of testing cepstral vector is shrunk in noisy environment, the model parameters, i.e., mean vector and covariance matrix, should be adapted simultaneously. In this study, we propose a method called variance adapted likelihood measure (VALM) which adapts the mean vector using a projection-based scale factor and adapts the covariance matrix using a variance reduction function estimated from the training database. The variance reduction function can be obtained according to various phonetic units. In the hidden Markov model based experiments, the speech recognition performance is greatly improved by applying VALM. The most significant improvement is achieved when the variance reduction function is separately estimated for different state parameters.

1. INTRODUCTION

In real application of speech recognition, the problem of environmental mismatch between training and testing utterances usually causes the unexpected recognition results. Among the mismatch sources, the noise effect created by background environment and low-quality microphone is an inevitable distortion source during the recording process. Before the noisy speech matches with the reference models, the speech can be enhanced by subtracting the corresponding noise spectrum [1]. Also, the reference models can be adapted to meet the testing environment. In order to reduce the mismatch problem, Mansour and Juang [2] proposed a family of distance measures based on projection operation for robust speech recognition. They found that the norm of cepstral vector was shrunk under noisy environment. Thus, they presented a first-order equalization method, which adjusted the reference cepstral vector by a shrinking factor, for minimizing the distance measure of each frame. The shrinkage of cepstral vector was a function of signal-to-noise ratio (SNR). The lower the SNR, the more the shrinkage. The remarkable improvement was reported in their dynamic time warping based experiments. In addition, Calson and Clements [3][4] proposed a weighted projection measure (WPM) for noisy speech recognition. By incorporating the algorithm of projection-based distance measure with the continuous-density hidden Markov model (HMM), the WPM was derived by adapting the mean vector using an equalization factor and keeping the covariance matrix unchanged. Because the magnitude of noisy cepstral vector is reduced, it makes intuitive sense to transform the covariance matrix as well as the mean vector. In the literature [5], the effect of additive noise was examined to induce the reduction of covariance matrix. In the assessment of the mismatch problem of mean and variance for noisy speech recognition, they reported that the mismatch yielded either by mean or variance would severely degrade the speaker identification results. However, the transformation of covariance matrix was not easily derived.

In this paper, we focus on the study of adaptation of covariance matrix in projection-based likelihood measure. The proposed measure for noisy speech recognition is called variance adapted likelihood measure (VALM). For a target speech with unknown noise level, the mean vector of reference model is adapted by a shrinking factor, and the covariance matrix is adapted by using a variance reduction function [6]. The variance reduction function is phone dependent and can be estimated from the training database. With the aid of prior knowledge of variance reduction function, the extra improvement of recognition rates for the adaptation of covariance matrix has been proved to be significant. Especially, when the variance reduction function is estimated according to various state parameters, the recognition rates under a SNR of 0 dB can be greatly increased from 10.1% to 40.2% compared with the baseline system.

2. VARIANCE ADAPTED LIKELIHOOD MEASURE

In the framework of conventional continuous-density HMM with Gaussian output probability, when an observed cepstral vector $c_i$ is time aligned with the $m^{th}$ component of $s^{th}$ state
parameter, the resulting observation probability \( P(\mathbf{c}_i | \Lambda_{x,m}) \) is represented by a likelihood measure of this form

\[
P(\mathbf{c}_i | \Lambda_{x,m}) = N(\mathbf{c}_i, \mu_{x,m}, \Sigma_{x,m}) =
(2\pi)^{-N/2} |\Sigma_{x,m}|^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{c}_i - \mu_{x,m})^T \Sigma_{x,m}^{-1}(\mathbf{c}_i - \mu_{x,m})\right)
\]

(1)

where \( N \) is the dimension of feature vector and \((\mu_{x,m}, \Sigma_{x,m})\) are the mean vector and covariance matrix, respectively. However, when the testing speech is contaminated by additive noise, the cepstral vector norm is shrunk and the reference mean vector can be adapted by a shrinking factor \( \lambda \) to equalize the testing speech. That is, the likelihood measure in Eq.(1) is modified as

\[
P(\mathbf{c}_i | \Lambda_{x,m}(\lambda)) = N(\mathbf{c}_i, \lambda \mu_{x,m}, \Sigma_{x,m}) =
(2\pi)^{-N/2} |\Sigma_{x,m}|^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{c}_i - \lambda \mu_{x,m})^T \Sigma_{x,m}^{-1}(\mathbf{c}_i - \lambda \mu_{x,m})\right)
\]

(2)

By applying the maximum likelihood (ML) criterion, the optimum shrinking factor \( \lambda_{opt} \) is then obtained by maximizing the logarithm of the above likelihood function. It can be expressed by

\[
\lambda_{opt} = \arg \max_{\lambda} \log P(\mathbf{c}_i | \Lambda_{x,m}(\lambda)) = \frac{\mathbf{c}_i^T \Sigma_{x,m}^{-1} \mu_{x,m}}{\mu_{x,m}^T \Sigma_{x,m}^{-1} \mu_{x,m} - \lambda \mu_{x,m}^T \Sigma_{x,m}^{-1} \mu_{x,m}}
\]

(3)

This equation is also equivalent to the orthogonality operation. Theoretically, the value \( \lambda_{opt} \) reflects the SNR level and its value locates in the interval \([0,1]\). The resulting distance measure is referred to as the weighted projection measure (WPM) and is given by

\[
d_{wpm}(\mathbf{c}_i, \Lambda_{x,m}) = (\mathbf{c}_i - \lambda_{opt} \mu_{x,m})^T \Sigma_{x,m}^{-1}(\mathbf{c}_i - \lambda_{opt} \mu_{x,m})
\]

(4)

In the experiments reported by [3], the recognition results outperformed that of weighted Euclidean distance (WED), which no adaptation is applied, for several performance factors. However, it is worth noticing that the WPM only scales the mean vector and remains the covariance matrix unchanged. In our analysis of training database, the reduction of noisy cepstral vector would result in the shrinkage of model mean vector and covariance matrix in the same time. Besides, we also find that the value \( \lambda_{opt} \) usually tends to zero or even negative at silence segment or when the speech is seriously degraded by additive noise. This leads to the ambiguity of pattern matching for different reference models. To moderately reflect the variation of model variances and enhance the model discriminability, the covariance matrix should be adapted for speech recognition in noise.

In the related work of WPM [4], a projection measure with noisy variance weightings was proposed for avoiding the tendency of disappearance of observed cepstral vector during the orthogonality process. They only weighted the observation vector with a grand noisy variance matrix which was a function of SNR. The derived projection measure with noisy variance weightings is shown as

\[
\hat{\mathbf{c}}_{wpm}(\hat{\mu}_{x,m}) = (\mathbf{c}_i - \frac{\mathbf{c}_i^T \hat{\mu}_{x,m}}{\hat{\mu}_{x,m}^T \Sigma_{x,m}^{-1} \hat{\mu}_{x,m}}) \left( \mathbf{c}_i - \frac{\mathbf{c}_i^T \hat{\mu}_{x,m}}{\hat{\mu}_{x,m}^T \Sigma_{x,m}^{-1} \hat{\mu}_{x,m}} \right)^T
\]

(5)

where the observed vector \( \mathbf{c}_i \) is weighted with a noisy variance matrix \( \hat{\Sigma}_{x,m} \) and yields the weighted cepstral vector \( \mathbf{c}_i = \hat{\Sigma}_{x,m}^{-1/2} \mathbf{c}_i \). And, the corresponding model mean vector is scaled by multiplying a grand variance matrix, i.e. \( \hat{\mu}_{x,m} = \Sigma_{x,m}^{-1/2} \mu_{x,m} \). The experiments showed that an additional gain of recognition rate was 10% at 0 dB. Instead of that, we propose a variance adapted likelihood measure (VALM) for noisy speech recognition. The VALM is constructed by merging the adaptation of covariance matrix into WPM. That is, when a target noisy frame is matched with a clean speech model parameter \((\mu_{x,m}, \Sigma_{x,m})\), the mean vector is shifted by using the corresponding optimum shrinking factor \( \lambda_{opt} \) and the covariance matrix is reduced according to the variance reduction function. The variance reduction function is a function of optimum shrinking factor \( \lambda_{opt} \) and can be estimated from the training database. For the mixture density HMM based recognizer, since the behavior of variance shrinkage differs from each other, we extract the variance reduction function \( K_{x,m}(\lambda_{opt}) \) for different state index and mixture index. The VALM is derived as

\[
P(\mathbf{c}_i | \Lambda_{x,m}(\lambda_{opt})) = N(\mathbf{c}_i, \lambda_{opt} \mu_{x,m}, K_{x,m}(\lambda_{opt}) \Sigma_{x,m})
\]

\[
= (2\pi)^{-N/2} |K_{x,m}(\lambda_{opt}) \Sigma_{x,m}|^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{c}_i - \lambda_{opt} \mu_{x,m})^T (K_{x,m}(\lambda_{opt}) \Sigma_{x,m})^{-1}(\mathbf{c}_i - \lambda_{opt} \mu_{x,m})\right)
\]

(6)

When the stage of Viterbi scoring is applied for an observation vector with unknown noisy level, the optimum shrinking factor is first determined by Eq.(3). Then, the corresponding
covariance matrix is adapted by multiplying a variance reduction vector $\mathbf{K}_{s,m}(\lambda_{opt})$. The VALM is finally calculated by using Eq.(6). By employing the VALM in Viterbi decoding, the model discriminability can be enhanced. Thus, the recognition rates should be increased especially when the value $\lambda_{opt}$ is very small.

3. ESTIMATION OF VARIANCE REDUCTION FUNCTION

From the above description, we know that the effectiveness of VALM is determined by the accuracy of variance reduction function. In this study, the covariance matrix is simplified as a diagonal matrix and the variance reduction function is determined by analyzing the behavior of variance shrinkage from the training database. The estimation procedure of variance reduction function is shown in Figure 1. For each clean training utterance, the state sequence and mixture sequence are determined by Viterbi decoding using the clean HMM models. Then, the optimum shrinking factor $\lambda_{opt}$ and the ratio of corresponding noisy variance $\tilde{\sigma}_{s,m,i}^2$ to clean variance $\sigma_{s,m,i}^2$ for $i^{th}$ cepstral index are calculated for corresponding utterances with SNR at infinity, 20, 15, 10, 5 and 0 dB. Next, the relations of $\lambda_{opt}$ and $\tilde{\sigma}_{s,m,i}^2/\sigma_{s,m,i}^2$ are dotted in a scatter diagram. Finally, the variance reduction function for different cepstral component $\mathbf{K}_{s,m}(\lambda_{opt}) = \{k_{s,m,i}(\lambda_{opt}), i = 1, \ldots, N\}$ is piecewisely estimated by averaging the dotted values with step size of $\lambda_{opt}$ being 0.01.

![Figure 1: Estimation procedure for variance reduction function.](image)

In addition, in order to examine the property of phone dependence of variance reduction function, we regard various phonetic units as the basis of estimation procedure. Because the proposed method is applied for a recognizer using mixture density HMM, we evaluate the phone dependence of variance reduction function based on the following types of phonetic unit:

- **Type 1:** Global unit.
- **Type 2:** Speech/non-speech unit.
- **Type 3:** State unit.
- **Type 4:** Mixture unit.

That is, the variance reduction ratios are calculated according to the type of phonetic unit. The smaller the phonetic unit is, the more accurate the variance reduction ratio is for variance adaptation. Figure 2 shows the scatter diagram and the estimated curve of VALM of type 1 for cepstral index $i = 3$. Note that it is reasonable to limit the valuable $\lambda_{opt}$ of variance reduction function in the interval $[0, 1]$. Also, if the VALM of type 4 is applied, another example of the variance reduction functions of 4 mixture components of a chosen state is illustrated in Figure 3. We can see that the variance reduction curve is approximately proportional to $\lambda_{opt}$ and their shapes are changed for different cases.

![Figure 2: Scatter diagram showing relations of $\lambda_{opt}$ and $\tilde{\sigma}_{type=1,i}^2/\sigma_{type=1,i}^2$ , and function $k_{type=1,i}(\lambda_{opt})$ for $i = 3$.](image)

4. EXPERIMENTAL RESULTS

A multispeaker (50 males and 50 females) isolated Mandarin digit recognition was conducted to demonstrate the performance of proposed method. There were 2000 utterances for training and 1000 utterances for testing. The speech signal was sampled
at 8 kHz. The white Gaussian noise was added to clean speech with predetermined SNR to generate the noisy speech. The feature vector is consisted of 12-order cepstral coefficients (i.e. \( N = 12 \)). The model parameters are trained under clean speech. In our experiments, the continuous-density HMM without adaptation is referred to as a baseline result. The results using WPM are also included for comparative evaluation. For an unknown noisy frame, the optimum shrinking factor \( \lambda_{opt} \) is first calculated and model mean vector is adjusted by using \( \lambda_{opt} \). The adaptation of covariance matrix for VALM of type \( n \) is then found to be \( K_{type}(\lambda_{opt}) \). The resulting VALM is calculated.

Four types of phonetic unit are investigated in the estimation procedure of variance reduction function. Two sets of experimental results are reported. One is for 1-mixture HMM and the other is for 4-mixture HMM. The experimental results are listed in Table 1 and Table 2 respectively. From the results, we can see that the proposed VALM is superior to WPM under every noisy level. The improvement is particularly significant when the SNR is low. It is probable because that the adaptation of variance enhances the model discriminability. Moreover, we find that the best results of VALM are achieved if the phonetic unit of variance reduction function is as precise as state unit (i.e. type 3). In addition, the comparison of different likelihood measures is illustrated in Figure 4. In this figure, 1-mixture HMM is applied and the VALM of type 3 is used. We find that the additional gain of recognition rate is 30% at 0 dB compared with the baseline results. This demonstrates the importance of adaptation of model variance in projection-based likelihood measure. Therefore, we conclude that the proposed VALM is a remarkable and robust likelihood measure for noisy speech recognition.

### Table 1: Experimental results using 1-mixture HMM

<table>
<thead>
<tr>
<th>Method/SNR</th>
<th>clean</th>
<th>20 dB</th>
<th>15 dB</th>
<th>10 dB</th>
<th>5 dB</th>
<th>0 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>95.8</td>
<td>69.9</td>
<td>52.1</td>
<td>35.8</td>
<td>24.4</td>
<td>10.1</td>
</tr>
<tr>
<td>WPM</td>
<td>92.1</td>
<td>78.1</td>
<td>66.1</td>
<td>48.4</td>
<td>32.9</td>
<td>23.3</td>
</tr>
<tr>
<td>VALM (Type 1)</td>
<td>94.2</td>
<td>79.9</td>
<td>67</td>
<td>54</td>
<td>42.9</td>
<td>33.5</td>
</tr>
<tr>
<td>VALM (Type 2)</td>
<td>94.4</td>
<td>79.8</td>
<td>66.6</td>
<td>55.9</td>
<td>44.8</td>
<td>35.1</td>
</tr>
<tr>
<td>VALM (Type 3)</td>
<td>95.3</td>
<td>81.3</td>
<td>70.8</td>
<td>58.4</td>
<td>48.7</td>
<td>40.2</td>
</tr>
</tbody>
</table>

### Table 2: Experimental results using 4-mixture HMM

<table>
<thead>
<tr>
<th>Method/SNR</th>
<th>clean</th>
<th>20 dB</th>
<th>15 dB</th>
<th>10 dB</th>
<th>5 dB</th>
<th>0 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>95.7</td>
<td>69.8</td>
<td>51.4</td>
<td>37.2</td>
<td>28.3</td>
<td>12.9</td>
</tr>
<tr>
<td>WPM</td>
<td>94.4</td>
<td>82.5</td>
<td>69.3</td>
<td>52.6</td>
<td>38</td>
<td>26.1</td>
</tr>
<tr>
<td>VALM (Type 1)</td>
<td>95.3</td>
<td>83.8</td>
<td>70.6</td>
<td>57.4</td>
<td>44.6</td>
<td>32.5</td>
</tr>
<tr>
<td>VALM (Type 2)</td>
<td>95.3</td>
<td>83.8</td>
<td>72</td>
<td>59.2</td>
<td>45.3</td>
<td>35.3</td>
</tr>
<tr>
<td>VALM (Type 3)</td>
<td>95.5</td>
<td>85.5</td>
<td>74</td>
<td>62.5</td>
<td>50.9</td>
<td>42.5</td>
</tr>
<tr>
<td>VALM (Type 4)</td>
<td>95.9</td>
<td>85</td>
<td>74.3</td>
<td>62</td>
<td>51</td>
<td>42.1</td>
</tr>
</tbody>
</table>

![Figure 4: Recognition rates for different likelihood measure.](image)

### 5. CONCLUSION

For noisy speech recognition, the projection-based distance measure is useful and computationally simple. But, in the HMM based recognition, if only the model mean vector is adapted, the improvement of recognition rate is limited. In this paper, we modify the conventional projection-based likelihood measure by adapting the mean and variance simultaneously. The adjustment of variance is based on the variance reduction function which is estimated from the training database. Experiments show that the extra gain of recognition rates is significant by adapting model variance as well as model mean.

### 6. REFERENCES