ABSTRACT

The performance degradation of automatic speech recognition system due to acoustic mismatch in training and testing environment is a severe problem for practical use of speech recognizer [1]. In this paper, we explore the effects of noise on individual speech feature vector statistics, and several feature normalization methods are used to compensate environment influence on feature vectors. We try to find out what kind of normalization is most effective and which feature vectors should be normalized in order to achieve robust features under adverse noise.

Keywords : robust speech recognition, feature normalization, noise

1. INTRODUCTION

To achieve robust speech recognition, we need to understand the change of feature vectors under noise environment in order to compensate the influence of noise. Researches have shown that noise has the effects of shift the means and compress the variances of speech signal statistics [3,9]. Some of the feature normalization methods are intend to compensate such effects in order to be robust against the convolution and additive noise, for example CMN [5] (Cepstral Mean Normalization), RATZ [9] (Multivariate Gaussian Based Cepstral Normalization), E-CMN [6] (Exact CMN) and Recursive Feature Vector Normalization [2]. The improvement of recognition accuracy when applying such normalization methods is hard to explain especially in adverse noise condition without much calculation cost. In this paper, we present several normalization methods to overcome the degradation of automatic speech recognition due to acoustic mismatch in training and testing environment.

2. EFFECTS OF NOISE ON FEATURE VECTORS OF SPEECH

Various kinds of noise causes different effect of speech signals. We may classify the noise influences into three categories. First, additive noise, speech signal is mixed with ambient noise. The structure of additive noise is various. For example, car noise (coming from wind, engine, road, etc.), factory noise and human conversations (babble noise) are possible sources of additive noise. Sometimes non-acoustic noise like electronic noise also affects the recognition accuracy. The fact that additive noise is unpredictable makes robustness with respect to additive noise very difficult. Second, channel noise, speech signal being transmitted through a particular system, such as microphone, telephone handset, transmission channel, usually results in certain channel distortion. Channel distortion may change frequency structure and phase of speech signal in a usually non-linear way. Therefore, the use of different microphone in testing and training may lead to serious degradation of recognition system. The last noise category, the so-called Lombard effect, which occurs when a speaker talks under very noisy conditions, the speaking style may change in terms of pitch, sound duration, etc. This effect is highly dependent upon the speaker and the level of noise, therefore it is very hard to compensate or quantify such effect [10]. In this paper, we focus on how to compensate additive and channel noise in order to achieve robust speech recognition.

To model additive and channel noise, commonly, the noisy speech signal can be viewed as \( y = (x + n) \circ h \)
where $x$ is the input signal, $n$ is the additive noise and $h$ is the channel transfer function. Based on this model, the effects of the noisy environment can be viewed as a shift in mean and change in variance of speech signal statistics [3,9].

To investigate the noise effect of individual speech feature vectors further, we illustrate in figure 1 and 2 the effect of variance shifting of individual feature vectors. The noisy speech samples are obtained by adding white noise artificially to clean speech at signal-to-noise ratios (SNR) scales 15 db and 5 db, and then the statistics of the noisy speech samples is compared to clean speech samples.

First, the variances of individual speech feature vectors are calculated. In figure 1, the standard deviation $\sigma$ of individual cepstral coefficients are shown, standard deviation shrinks when the noise increases in all the feature vectors, such shrinking effect is obvious especially in the lower order of cepstral parameters. Figure 2 illustrates standard deviation shrinkage of delta-cepstral and delta-log-energy coefficients under noisy environment. The shrinking effect of standard deviation in delta-log-energy and delta-delta-log-energy are apparent as compare to delta cepstral coefficients. Both the figures show the need of variance adaptation, so as to compensate the influence of noise.

Second, the means of individual speech feature vectors are calculated. The means change its value when noise appears. In the lower order of cepstral parameters such shifting effect are especially evident. Nowadays, mean adaptation is widely applied to feature extraction system in order to compensate channel and additive noise.

3. FEATURE NORMALIZATION

As described in previous section, noise has the effects of shift in mean and change in variance of speech signal statistics. To compensate the effect of mean shift, we may normalize the noisy speech signal to have zero mean such as CMN method. To compensate the effect of change in variance, we may normalize the noisy speech signal to have unity variance.

As in neural network approach, input feature vectors are usually scaled to zero mean and unity variance as

$$\overline{x} = \left( x - \mu \right) / \sigma$$  \hspace{1cm} (1)

where $x$ is the original feature vector, and $\overline{x}$ is its normalized feature vector, $\mu$ and $\sigma$ are the mean and standard deviation of the input feature vector. The feature vectors can be normalized to have a zero mean and unity variance based on this method.

We tried two kinds of normalization in our experiments, utterance based normalization and speaker based normalization. When the cepstrum is normalized utterance by utterance, the normalization is called utterance based normalization. Speaker based normalization means the cepstrum is normalized from several utterances of a specific speaker’s speech to include characteristics of the speaker.
4. EXPERIMENTAL EVALUATION

Our experiments are based on MATDB-4 [11] Mandarin telephone speech database. The training database consists of 362 male speakers with 9563 utterances. The testing set consists of 80 male speakers with 2344 utterances. Noise (NOISEX-92) is artificially added to speech samples during testing at various signal-to-noise ratios (SNR). The recognition task is speaker independent Chinese isolated word recognition, the vocabulary size is 60 out of 1062 from the database. Each word consists of 2 to 4 syllables.

In the following experiments, speech signals are sampled at 8 kHz and preemphasized with a digital filter \(1 - 0.95z^{-1}\). It was then analyzed for each Hamming-windowed frame of 20 ms with 10 ms frame shift. The recognition features consist of 12 mel-cepstral coefficients, 12 delta mel-cepstral coefficients, delta energy, and delta-delta energy. The HMM-based speech recognizer employed 60 sub-syllable models as basic recognition units, including 22 three-state context-independent INITIAL models and 38 five-state context-independent FINAL models. The observation distribution for each state of the HMM was modeled by a multivariate Gaussian mixture distribution. The number of mixture components in each state varies from five to ten, and each of the mixture components has a diagonal covariance matrix. Silence model uses a single-state model with ten mixtures.

5. EXPERIMENTAL RESULTS

Figure 3 shows the results obtained from speaker independent word recognition at different noise levels. Three kinds of noise were added to speech samples, white, factory and car noise. Figure 3 shows the average recognition error rate of speech samples corrupted by these three different kinds of noise. It is seen in figure 3, that normalization methods improved the recognition result remarkably over the no normalization case. The result also shows that variance-and-mean normalization reduces the recognition error further as compare to mean normalization only when the SNR decreases.

In order to investigate the normalization effect of diverse feature vectors, two types of normalization are used. First, all the 26 coefficients are normalized to obtain mean-and-variance normalization which called type 1 normalization, and second, only a portion of vectors are normalized (12 cepstral coefficients) which called type 2 normalization. There is no significant recognition rate difference between these two types of normalization as seen in table 1. Type 1 normalization gives 66.15% recognition improvement as compare to without normalization case, whereas type 2 normalization improves 67.89% recognition accuracy over no normalization case. The effect of normalization of delta-cepstral coefficients is not significant, because delta coefficients represent dynamic features.

Afterward, speaker based normalization is investigated. The feature coefficients were normalized over 20 words of the same speaker. It can be seen in table 1, speaker based normalization gives the best recognition results. Type 1 (all vectors) and type 2 (only 12 cepstrum) speaker based normalization have no significant recognition accuracy difference, which gives 73.11% and 73.02% improvement respectively. A drawback of speaker based normalization is that it requires a long-term speech data to carry out speaker normalization, which means the normalization should be done off-line or recursively.
### Table 1. Recognition error rates (%)

<table>
<thead>
<tr>
<th>Method</th>
<th>clean</th>
<th>20db</th>
<th>15db</th>
<th>10db</th>
<th>5db</th>
<th>0db</th>
<th>Average</th>
<th>Improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>without normalization</td>
<td>3.96</td>
<td>7.94</td>
<td>13.12</td>
<td>25.39</td>
<td>48.03</td>
<td>73.17</td>
<td>28.60</td>
<td>0.00</td>
</tr>
<tr>
<td>mean normalization</td>
<td>2.05</td>
<td>4.21</td>
<td>5.75</td>
<td>8.63</td>
<td>16.56</td>
<td>35.00</td>
<td>12.03</td>
<td>57.93</td>
</tr>
<tr>
<td>mean and variance normalization (1)</td>
<td>1.88</td>
<td>3.41</td>
<td>4.82</td>
<td>7.35</td>
<td>13.24</td>
<td>27.39</td>
<td>9.68</td>
<td>66.15</td>
</tr>
<tr>
<td>mean and variance normalization (2)</td>
<td>1.96</td>
<td>3.41</td>
<td>4.80</td>
<td>6.86</td>
<td>12.68</td>
<td>25.41</td>
<td>9.19</td>
<td>67.89</td>
</tr>
<tr>
<td>speaker based normalization (1)</td>
<td>1.66</td>
<td>2.97</td>
<td>4.11</td>
<td>6.29</td>
<td>10.67</td>
<td>20.45</td>
<td>7.69</td>
<td>73.11</td>
</tr>
<tr>
<td>speaker based normalization (2)</td>
<td>1.58</td>
<td>2.96</td>
<td>4.03</td>
<td>6.40</td>
<td>10.75</td>
<td>20.58</td>
<td>7.72</td>
<td>73.02</td>
</tr>
</tbody>
</table>

#### 6. CONCLUSION

In this paper, we explore the effects of noise on individual speech vector statistics. Noise results in variance shrinking and mean shifting in all the cepstral feature vectors, such effects are obvious especially in the lower order of cepstral parameters. In order to compensate such influence, various normalization methods are then investigated. Speaker based mean and variance normalization gives the best recognition accuracy in our experiments.

#### 7. ACKNOWLEDGEMENT

This paper is a partial result of the project No. 3P11100 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C.

The authors would like to thank the Association for Computational Linguistics and Chinese Language Processing in Taiwan for kindly supplying the database.

#### 8. REFERENCES


