COMBINING MLLR ADAPTATION AND FEATURE EXTRACTION FOR ROBUST SPEECH RECOGNITION IN REVERBERANT ENVIRONMENTS

Aik Ming Toh, Roberto Togneri, Sven Nordholm

School of Electrical, Electronic and Computer Engineering
The University of Western Australia, Australia
aikming@ee.uwa.edu.au, roberto@ee.uwa.edu.au, sven@watri.org.au

Abstract

This paper presents an investigation on speech recognition performance in reverberant environments. Reverberant noise has been a major concern in speech recognition systems. Many speech recognition systems, even with state-of-art features, fail to respond to reverberant effects and the recognition rate deteriorates. This shows the limitations of robust feature extraction in reverberant environment. The maximum likelihood linear regression (MLLR) adaptation scheme is adopted for reverberant speech recognition on the TI-DIGIT database. The use of adaptation data improved the recognition performance significantly especially for strong reverberations. The performance of both MFCC₀ and MFCC₀D_A features improved by more than 10% for reverberations greater than 0.4s. This paper also demonstrates the optimal strength of both robust feature extraction and adaptation scheme for reverberant speech recognition. The recognition performance is maintained above 90% up to reverberation time 0.5s using both schemes.

1. Introduction

Current state-of-the-art speech recognition systems show impressive performance for clean acoustic environments recognition. However, the performance degrades when there is a mismatch between the training data and the testing data. Large amounts of training data are required to retrain speech recognition systems to a new environment. This alternative is not feasible as it is both computationally expensive and time consuming. Therefore, it is desirable to be able to improve the performance of the system while using a small amount of adaptation data. Several adaptation schemes have been introduced into speech recognition systems to enhance the performance and robustness in noisy environments [1]. However, most of these research focus on additive noise [2], [3].

There have also been some work done on model estimation [4] and microphone array processing [5] in reverberant speech recognition. While most of the research in reverberant environments incorporate the use of microphone array processing, this paper focuses on the use of adaptation data and speech features to improve the recognition accuracy and robustness.

From previous work [6] it has been shown that feature extraction alone is not enough to alleviate reverberant noise. No single feature set performed best for all different levels of reverberation. While one feature may be robust in additive noise, it may not be robust in reverberant noise. Even the state-of-art feature MFCC₀D_A suffers in reverberant environments.

There are papers which have reported that even with reverberant matched models, the recognition rate cannot be improved sufficiently, when the reverberation time is greater than 0.4s [7], [8]. The results presented in this paper demonstrate a different perspective. It is shown that by using a small portion of the training data as adaptation data, the recognition rate can be significantly improved.

In this paper, the performance of adaptation scheme in reverberant speech recognition is compared with the performance of acoustic matching scheme. In addition, the significance of using both state-of-art speech feature MFCC₀D_A and MLLR adaptation scheme in reverberant speech recognition is also demonstrated. Significant improvements are achieved for severe reverberation time.

The paper is organized as follows: The next section explores the effects of reverberation on speech signals. The MLLR adaptation scheme is presented in section 3. Section 4 specifies the experimental setup followed by the results in section 5. The last section comprises the conclusions.

2. Reverberant Effects on Speech Signals

Human auditory systems are accustomed to speech in moderate reverberant environment. However, the performance of speech recognition system degrades dramatically in reverberant condition. Reverberation is caused by the superposition of an acoustic signal...
and its reflected signals of different delays and amplitudes. It introduces a convolutional interference that comprises both spectral distortion and additive noise.

An obvious effect of reverberation is temporal smearing on the acoustic signal. This effect is also known as overlap masking in which segments of an acoustic signal are affected by the reverberant components of previous segments. Although this adds richness of sound to music, it makes the speech losing its intelligibility. Human listeners do not have much trouble understanding reverberant speech because most of the basic energy features are still in intact even under the influence of reverberations. However, it is difficult for speech recognition systems since they are trained in clean models. The smearing effects are evident in the temporal resolution of both spectral and cepstral features. Figure 1 shows the spectrogram of a speech utterance in clean and reverberant condition. The spectrogram of the reverberant utterance clearly depicts the temporal smearing effect.

Reverberation typically increases the loudness at given location because the energy generated over a range of time in the past is received in the present. The syllable onsets and identities of the speech can be masked by the decaying energy from previous syllables when the reverberation time, RT60 is long [9]. This can hurt the intelligibility of the speech severely. RT 60 is used to characterize the reverberation time. It is the time interval in which the reverberation level decays by 60dB.

In addition, spectral flattening effect can be observed on the envelope of speech features. Although reverberation affects the mean and variance of speech features, the shifts in these parameters can be effectively compensated with normalization strategies such as mean and variance normalization. However, the temporal smearing effect could not be effectively compensated in the feature context. Thus, adaptation is proposed to improve the performance of speech recognition in reverberant environment.

3. Maximum Likelihood Linear Regression

The adaptation scheme used in this experiment is the Maximum Likelihood Linear Regression (MLLR) method. This method is simple and known to be robust for unsupervised adaptation as well as effective for small amounts of adaptation data. It was initially developed for speaker adaptation [10]. The aim of MLLR is to obtain a set of transformation matrices that maximizes the likelihood of the adaptation data. The transformation sets are relatively small compared to the total number of Gaussian parameters computed from the training data.

A single global transformation is used for the small amount of data. In this paper, both the mean parameter estimation and Gaussian variance were updated. The means and variances are adapted in two separate stages. The variances are updated after the new mean is derived. The transformation matrices are tied across a number of Gaussians to ensure robust estimation of transformation parameters. The set of Gaussians which share a transform is referred to as a regression class. The HMM models are modified such that [11]

\[ \mathcal{L}(O_T | \mathcal{M}) \geq \mathcal{L}(O_T | \hat{\mathcal{M}}) \geq \mathcal{L}(O_T | \mathcal{M}) \]

where \( \mathcal{L} \) is the likelihood, \( \mathcal{M} \) is the original model set, \( \hat{\mathcal{M}} \) has the updated mean parameters, \( \mathcal{M} \) set has updated both means and variances and \( O_T = \{ o(1), o(2), ..., o(T) \} \) is the adaptation data.

The transformation matrix used to produce a new estimate of the adapted mean is given by [12],[10]

\[
\hat{\mu} = W \epsilon \quad \text{(2)}
\]

\[
\epsilon = [\omega \mu_1 \mu_2 ... \mu_n]^T \quad \text{(3)}
\]

where \( W \) is the \( n \times (n + 1) \) transformation matrix (for \( n \) dimensional data), \( \epsilon \) is the extended mean vector and \( \omega \) represents a bias offset of value 1.

The Gaussian variance vectors or the covariance matrices are updated using the transformation [12];

\[
\sum_m = B_m^T \hat{H}_m B_m \quad \text{(4)}
\]

where \( \hat{H}_m \) is the linear transformation to be estimated and \( B_m \) is the inverse of the Cholesky factor of \( \sum_m^{-1} \), such that \( \sum_m^{-1} = C_m C_m^T \) and \( B_m = C_m^{-1} \).

The variance transformation is shared over a number of Gaussians and the maximum likelihood estimate is given by

\[
\hat{H}_m = \frac{\sum_{r=1}^{R} C_r^T \left[ \sum_{\tau=1}^{T} L_r(\tau)(o(\tau) - \hat{\mu}_r)(o(\tau) - \hat{\mu}_r)^T \right] C_r}{\sum_{r=1}^{R} \sum_{\tau=1}^{T} L_r(\tau)} \quad \text{(5)}
\]

where \( \hat{\mu}_r \) is the previously calculated mean.

The formulations for the transformation matrix used to give a new estimate of the adapted mean and variance based on (1) have been well reported [11].
Table 2: Speech recognition with adaptation in reverberant environments

<table>
<thead>
<tr>
<th>Test_data</th>
<th>0.2s</th>
<th>0.4s</th>
<th>0.5s</th>
<th>0.6s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mismatch</td>
<td>92.60</td>
<td>77.35</td>
<td>66.09</td>
<td>53.27</td>
</tr>
<tr>
<td>Adapt 0.1s</td>
<td>93.91</td>
<td>82.83</td>
<td>72.55</td>
<td>62.38</td>
</tr>
<tr>
<td>Adapt 0.2s</td>
<td>95.02</td>
<td>84.16</td>
<td>74.60</td>
<td>63.12</td>
</tr>
<tr>
<td>Adapt 0.3s</td>
<td>94.93</td>
<td>86.61</td>
<td>78.17</td>
<td>68.09</td>
</tr>
<tr>
<td>Adapt 0.4s</td>
<td>94.08</td>
<td>86.09</td>
<td>79.26</td>
<td>69.85</td>
</tr>
<tr>
<td>Adapt 0.5s</td>
<td>92.28</td>
<td>86.39</td>
<td>80.37</td>
<td>71.19</td>
</tr>
<tr>
<td>Adapt 0.6s</td>
<td>92.72</td>
<td>86.21</td>
<td>81.09</td>
<td>73.49</td>
</tr>
<tr>
<td>Adapt 0.7s</td>
<td>92.05</td>
<td>85.42</td>
<td>81.29</td>
<td>74.85</td>
</tr>
</tbody>
</table>

4. EXPERIMENTAL SETUP

The database comprised both isolated and connected digit utterances from the TI-digit corpus. The training data contained utterances of 24 male and 24 female speakers. There were 8 male and 8 female speakers for the testing data. Each speaker subset composed of 77 digit utterances.

Eleven isolated digit utterances from each training data subset were extracted as adaptation data. The adaptation data accounted for one-seventh of the total number of training utterances. The amount or proportion of adaptation data used was less than that reported in [5] and [13]. These data were corrupted to match the reverberant condition of the testing data described below.

The MLLR scheme was used to adapt clean speech models to the reverberant adaptation data. The aim was to keep the adaptation data to a minimum while optimizing the improvements in the recognition accuracy.

Reverberant effects were captured by estimating the impulse response of the room environments from long segments of speech. The experiment used the room impulse response designed to match the characteristic of a 2.2m high, 3.1m wide and 3.5m long room. The microphone and the speakers were localized 0.5m from the wall at opposite end. The speech was convolved with the RT60 room impulse response. The number of filter coefficients was adjusted according to the reverberation time.

All the speech files were pre-emphasized and windowed with a Hamming window. The speech signal was analyzed every 10ms with a frame width of 25ms. A Mel-scale triangular filterbank with 26 filterbank channels was used to generate the Mel-frequency cepstral coefficients (MFCC) features. The MFCC coefficients consisted of 12 static MFCCs and the zeroth cepstral coefficients. The HMM model used 15 states and 5 mixtures for the connected digit recognition.

5. EXPERIMENT RESULTS

In the author’s previous work on robust features for speech recognition in hostile environments [6], it had been shown that different features showed robustness in different levels of reverberation. Figure 2 shows the recognition accuracy of speech recognition in a mismatched environment where clean data train and reverberant test data were used. It is observed that the recognition rate degraded dramatically as the reverberation time increased. The system failed to sustain recognition accuracy of 90% after RT60 0.2s.

5.1. Matched Acoustic Conditions

The training data was filtered with the same room impulse response used by the test data respectively. Optimal performance in the recognition accuracy was achieved. Figure 2 illustrates the performance of matched condition across the incremental reverberation time. Recognition rates of more than 90% are achieved up to a reverberation time of 0.4s.

A number of slightly mismatched speech recognition tests were also performed to show the effectiveness of acoustic matching scheme in reverberant environments. The recognition accuracy for mismatch data is 77.35% for RT60 of 0.4s. Table 1 shows that training in even a slightly mismatched environment such as 0.1s will improve the recognition accuracy by 10.53%. The use of better matched conditions would further improve the recognition accuracy.

5.2. MLLR Adaptation

The use of acoustic matching scheme shows optimal results but this requires the training data be corrupted to match the acoustic condition of the reverberant testing data. Such an option is computationally expensive as well as time consuming, considering the vast amount of training data required. Thus, the use of adaptation scheme such as MLLR is recommended to adapt the clean speech models to the corrupted speech models.

Table 2 depicts the results of speech recognition with adaptation scheme in reverberant environments. The results show that significant improvements were achieved in the recognition accuracy by adapting the HMM models to noisy data. With adaptation, even slightly mismatched adaptation data contributes to improvement in the recognition performance.

It can be observed that the MLLR adaptation scheme offers more contribution to the recognition in more severe reverberations. The adaptation scheme improved the recognition accuracy by more than 10% for speech recognition in RT60 of 0.4s to 0.6s. The MLLR adaptation method is an effective and feasible solution to reverberant speech recognition since it only requires small amounts of adaptation data.

It is also observed that the optimal recognition accuracy in each reverberant environment was not achieved by adapting to the same reverberation level respectively. This could be due to the small amount of adaptation data used in this experiment. However, the emphasis of this work is to demonstrate the robustness and strength of adaptation scheme in reverberant environment rather than the optimality.

5.3. MFCC_0_D_A and Adaptation

We extended our baseline speech features to MFCC_0_D_A to show that adaptation are more effective than feature extraction in reverberant conditions. Table 3 shows the recognition accuracy of MFCC_0_D_A in mismatch environments. This feature is recog-
Table 4: Speech recognition with MFCC_0, D, A in reverberant environments

<table>
<thead>
<tr>
<th>RT60 (s)</th>
<th>clean</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matched</td>
<td>99.50</td>
<td>99.28</td>
<td>99.28</td>
<td>99.03</td>
<td>98.32</td>
<td>97.23</td>
<td>95.30</td>
<td>91.44</td>
</tr>
<tr>
<td>Mismatch</td>
<td>99.50</td>
<td>99.18</td>
<td>98.49</td>
<td>95.69</td>
<td>85.30</td>
<td>74.95</td>
<td>57.72</td>
<td>46.36</td>
</tr>
<tr>
<td>Adapted*</td>
<td>98.96</td>
<td>98.51</td>
<td>97.60</td>
<td>95.45</td>
<td>90.74</td>
<td>81.31</td>
<td>71.04</td>
<td></td>
</tr>
</tbody>
</table>

nized as the state-of-art feature in most speech recognition systems. We can see the superior performance of MFCC_0, D, A features over the baseline MFCC_0 from Figure 2.

Table 3 shows that MFCC_0 features perform better than MFCC_0, D, A features when the adaptation scheme is utilized. The 13 cepstral coefficients MFCC_0 features are able to surpass the performance of MFCC_0, D, A features, which composed of 13 cepstral coefficients and 26 regression features, with the aid of adaptation scheme. The MFCC_0 Adapt* in Table 3 refers to the optimally adapted speech recognition accuracy which has the best adapted result. The state-of-art MFCC_0, D, A features did not outperform the adapted baseline MFCC_0 in heavily reverberant conditions. This is also shown in Figure 2. The experiments establish the strength of adaptation for speech recognition particularly in severe reverberant environments.

We had also performed speech recognition with both MLLR adaptation and MFCC_0, D, A features. This combination yielded the best results in the experiments. Table 4 shows the performance of matched, mismatched and optimally adapted speech recognition in reverberant environments. We are able to achieve recognition accuracy of more than 90% up to RT60 of 0.5s for the adapted system. The results also indicate evident improvements (bold) for severe reverberation such as 0.5s to 0.7s.

6. CONCLUSION

The amount of adaptation data used is trivial compared to the vast training data used in the experiment. Though the adapted results are not as good as the matched acoustic results, substantial improvements in the recognition accuracy have been achieved with just a small amount of adaptation data. These improvements are significant and it is more apparent for reverberation times such as 0.4s which are common in real-world scenario.

It is also important to note that only isolated digit utterances were used for adaptation data and normalization strategies were not implemented. The performance can be further enhanced if the adaptation data comprises some connected digits utterances or more utterances and the features are normalized.

It has been shown that with matched environments, it is possible to maintain the recognition accuracy above 90% up to severe reverberations. The use of adaptation scheme with MFCC_0, D, A enabled us to maintain the recognition accuracy above 90% even for reverberant conditions up to 0.5s. Significant improvements have been achieved with the use of adaptation data in compensating the reverberant effects in speech recognition systems.

Therefore, we conclude that the combination of robust feature extraction scheme and adaptation scheme offers the optimal performance for speech recognition in reverberant environments.

7. REFERENCES