ON THE USE OF FEATURE-SPACE MLLR ADAPTATION FOR NON-NATIVE SPEECH RECOGNITION

Yoo Rhee Oh and Hong Kook Kim

Department of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
{yroh, hongkook}@gist.ac.kr

ABSTRACT

In this paper, we address issues associated with a feature-space maximum likelihood linear regression (fMLLR) adaptation method applied to non-native speech recognition. In particular, fMLLR smoothing is proposed here to compensate for mismatches between adaptation and test data, caused by the various disfluencies of non-native speakers. The proposed fMLLR smoothing is performed with a Viterbi decoding procedure and implemented at two levels: a Gaussian mixture probability density function (mpdf) level and an observation probability density function (opdf) level. The mpdf-level smoothing is performed by comparing the pdf of each Gaussian mixture component of an original speech feature vector with that transformed by the fMLLR. On the other hand, the opdf-level smoothing compares the Gaussian mixture probabilities between the original and its fMLLR transformed feature vectors. It is shown from non-native automatic speech recognition experiments on a Korean-spoken English continuous speech corpus that an ASR system employing the proposed mpdf-level and opdf-level fMLLR smoothing methods can relatively reduce the average word error rate by 30.65% and 29.82%, respectively, when compared to a traditional fMLLR adaptation method.

Index Terms—Non-native speech recognition, feature-space maximum likelihood linear regression (fMLLR), feature compensation, acoustic model adaptation

1. INTRODUCTION

As an automatic service in an international society, the demand for non-native automatic speech recognition (ASR) in many applications is increasing, most notably in computer assisted language learning (CALL) systems and automatic response systems (ARS), etc. However, the performance of non-native ASR tends to significantly degrade due to mismatches between the training and test data, caused by differences between characteristics of the mother tongue and the disfluencies of non-native speakers in pronunciation [1]. In order to improve the performance of non-native ASR, a number of adaptation methods have been reported; these adaptation methods can be classified as acoustic modeling [2]-[4], pronunciation modeling [2][5]-[9], language modeling [10], and hybrid modeling [11] adaptation approach. Moreover, it has been shown that maximum likelihood linear regression (MLLR) and/or maximum a posteriori (MAP) adaptations applied after the adaptations described above can likely improve ASR performance for non-native speech even further [9][12]-[14].

In this paper, we attempt to apply a feature adaptation for non-native speech in order to compensate for a mismatch between training and test data in a feature domain. Note that feature adaptation methods have already been proposed in several fields, such as speaker adaptation [15], with feature-space MLLR (fMLLR) adaptation being used for speaker adaptation [16]. To efficiently reduce the mismatch between training and test conditions in the fMLLR framework, several smoothing methods for fMLLR adaptations have also been proposed, which non-linearly transform a feature vector [17]-[19]. For example, an fMLLR transformation matrix was smoothed by projecting it into a low dimensional subspace by using the concept of heteroscedastic discriminant analysis [17]. A non-linear feature space transformation method was proposed in which each feature vector was first Gaussianized, followed by an fMLLR transform [18]. In addition, a non-linear method based on the feature-space minimum phone error (fMPE) technique to discriminatively estimate improved features was presented in [19].

In order to apply an fMLLR adaptation to non-native speech, we propose new fMLLR smoothing techniques since significant mismatches exist between the adaptation and test data due to various disfluencies in the pronunciation of non-native speakers. That is, the proposed fMLLR smoothing is performed via a Viterbi decoding procedure and implemented at two levels: at a Gaussian mixture probability density function (mpdf) level and at an observation probability density function (opdf) level. The mpdf-level fMLLR smoothing is performed by comparing the pdf of each Gaussian mixture component of an original speech feature vector with that transformed by the fMLLR. Conversely, the opdf-level fMLLR smoothing is performed by comparing Gaussian mixture probabilities between the original and its fMLLR-transformed speech feature vectors. Then, for further improvement, a traditional MLLR adaptation [9][13][14] is also applied to an ASR system that employs an fMLLR adaptation based on the proposed fMLLR smoothing methods.

The organization of the remainder of this paper is as follows. The baseline ASR for native English, including its acoustic, pronunciation, language models, and speech database, is briefly described in Section 2. In Section 3, we explain the fMLLR adaptation method and compare fMLLR adaptation methods applied to speaker adaptation and to non-native speech. In Section 4, we propose fMLLR smoothing methods for non-native speech. In Section 5, the performance of a non-native ASR system employing the proposed methods is evaluated and compared to that using a traditional fMLLR adaptation method. Finally, we conclude our findings in Section 5.
2. BASELINE ASR SYSTEM

2.1. Speech database

In order to construct a baseline ASR for native speech, we used a subset of the Wall Street Journal database (WSJ0), where WSJ0 was a 5,000-word closed-loop task used to evaluate the performance of a large vocabulary continuous ASR system [20]. The WSJ0 training set consisted of 7,138 sentences (130,507 words in total) by 83 native English speakers, which were recorded using a Sennheiser close-talking microphone and several far-field microphones, in which each utterance was sampled at a rate of 16 kHz.

In order to enable non-native speakers to use such a baseline ASR system, we used a subset of the Korean-Spoken English Corpus (K-SEC) [21] to adapt the acoustic models and then to evaluate the performance of the non-native ASR system. The adaptation set was composed of 14,759 isolated word utterances, spoken by 14 Koreans. In addition, the test set was composed of 784 sentence utterances (686 and 98 sentences were spoken by 49 Korean and 7 native English speakers, respectively) at an average of 10.4 words per sentence (8,176 words in total).

2.2. Baseline ASR system

As an ASR feature, we extracted 12 mel-frequency cepstral coefficients (MFCCs) with logarithmic energy for every 10 ms analysis frame and concatenated their first and second derivatives, resulting in a 39-dimensional feature vector. During training and testing, we applied cepstral mean normalization and energy normalization to the feature vectors. In this paper, the acoustic models were based on 3-state left-to-right, context-dependent, 4-mixture, and crossword triphone models, which were trained using the HTK version 3.4.1 Toolkit [22]. All of the triphone models were expanded from 41 monophones, which also included silence and pause models, and the states of the triphone models were tied using a decision tree [23]. As a result, the acoustic models consisted of 8,360 triphones and 5,356 states.

In order to explore discrepancies in the behavior of the acoustic models due to the differences between the target language and the speaker’s mother tongue, only the texts from the test set were used to construct a back-off bigram language model. The pronunciation of each word was built from the Carnegie Mellon University (CMU) pronunciation dictionary [24] and any words missing from the CMU dictionary were transcribed manually.

Table I shows the performance evaluation results of the baseline ASR system using the test data set for native and non-native speech. As shown in the table, the average word error rates (WERs) of the baseline ASR system for native speech and non-speech. As shown in the table, the average word error rates (WERs) of the baseline ASR system for native speech and non-native speech were 0.68% and 19.92%, respectively.

<table>
<thead>
<tr>
<th>Speech</th>
<th>Native</th>
<th>Non-native</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.68</td>
<td>19.92</td>
</tr>
</tbody>
</table>

where \( x_t \) is the \( d \)-dimensional input speech feature vector at time \( t \), and \( \omega_k \) and \( K \) are the weight of the \( k \)-th Gaussian mixture component and the number of Gaussian mixtures, respectively. In addition, \( m_g(g), S_g(g), A(g), \) and \( b(g) \) are the mean vector, covariance matrix, transformation matrix, and bias vector for a regression class \( g \), where the same genome is mapped for a set of HMM states since the set of HMM states is shared with the same Gaussian mixture component. The estimation of the transformation matrix \( A \) and the bias vector \( b \) is achieved by using an expectation-maximization (EM) algorithm; a more detailed description is explained in [21].

During the Viterbi decoding, the Gaussian mixture density of \( x_t \) at the \( k \)-th Gaussian mixture component of an HMM state \( s \) can be obtained using

\[
\rho_{s,k}(x_t) = \hat{p}_{s,k}(x_t) = \hat{p}_{s,k}(\mathbf{x}(\mathbf{h}) + b(g)) = [\mathbf{d}(\mathbf{h}) | A(g)h + b(g), \mu_k, \Sigma_k] (2)
\]

where \( \hat{x}_t \) is the transformed speech feature vector and \( |A(g)\) is a determinant of the Jacobian of the \( g \)-th regression class for transforming \( x \) to \( \hat{x}_t \).

3. fMLLR ADAPTATION FOR NON-NATIVE SPEECH

3.1. Overview of fMLLR adaptations

Similar to the transformation matrix for MLLR adaptations, a \( d \) by \( d \) transformation matrix \( A \) and a \( d \)-dimensional bias vector \( b \) are estimated by aiming to maximize the Gaussian mixture density \( P(x_t) \), in the form of

\[
P(x_t) = \sum_{k=1}^{K} \omega_k N(x_t; A(g)m_g(g) + b(g), A(g)S_g(g)A(g)^T(g)) (1)
\]
tion. In the adaptation procedure, a set of transformation matrices for each speaker is estimated by using an EM algorithm. In the Viterbi decoding procedure, the input speech feature vectors are transformed using the transformation matrix corresponding to each speaker. On the other hand, Fig. 1(b) shows the main procedure for the fMLLR adaptation of non-native speech. In this procedure, a set of transformation matrices is estimated using an EM algorithm for non-native speech, not for each non-native speaker. In the Viterbi decoding procedure, each input speech feature vector is then transformed using one set of transformation matrices, even though the non-native speaker changes. It should be noted that the set of transformation matrices for non-native speech is generalized based on the adaptation data spoken by non-native speakers, whereas each set of transformation matrices is optimized for the corresponding speaker in speaker adaptation. Therefore, the fMLLR adaptation for non-native speech should be selectively applied to the input of non-native speech due to the broad range of possible fluency levels according to different non-native speakers.

4. PROPOSED fMLLR SMOOTHING FOR NON-NATIVE SPEECH RECOGNITION

After applying an fMLLR adaptation for speaker recognition, we can estimate the probability density function (pdf) for the speech feature vector transformed from Eq. (2), $\hat{z}_t$, as follows.

$$
\log P_t(\hat{z}_t) = \log \sum_{k=1}^{K} \omega_{t,k} \hat{p}_{t,k}(\hat{z}_t) \\
= \log \sum_{k=1}^{K} \omega_{t,k} g_b(x_t|\theta_{t,k}) + b(x_t) \mu_{t,k} \Sigma_{t,k} \\
= \log \sum_{k=1}^{K} \omega_{t,k} \hat{p}_{t,k}(\hat{z}_t)
$$

However, it was noted from Section 3.2 that applying the fMLLR adaptation for non-native speech can also provide a mismatch for some non-native test speech if the test data are characterized differently from the adaptation data. Therefore, it would be better if the fMLLR adaptation method should not be performed for mismatched non-native speech data. Based on this phenomenon, we proposed two fMLLR smoothing methods to be performed with the Viterbi decoding procedure: a Gaussian mixture pdf (mpdf)-level fMLLR smoothing and an observation pdf (opdf)-level fMLLR smoothing.

4.1. Mixture pdf (mpdf)-level fMLLR smoothing

The mpdf-level fMLLR smoothing is performed by comparing the pdf of each Gaussian mixture component, which is evaluated by an original speech feature vector, $x_t$, with that of the fMLLR-transformed component, $\hat{z}_t$, using the equation of

$$
\log P_t(\hat{z}_t) = \log \sum_{k=1}^{K} \omega_{t,k} \max\{p_{t,k}(x_t), \hat{p}_{t,k}(\hat{z}_t)\}
$$

where $s$ denotes the HMM state during the Viterbi decoding.

The proposed mpdf-level smoothing method selects the more probable Gaussian mixture component from between the original and fMLLR-transformed components.

4.2. Observation pdf (opdf)-level fMLLR smoothing

The observation pdf (opdf)-level fMLLR smoothing is performed by comparing the Gaussian mixture probabilities, evaluated using both the original and fMLLR-transformed speech feature vectors,

$$
\log P_t(\hat{z}_t) = \log \sum_{k=1}^{K} \omega_{t,k} \max\{p_{t,k}(x_t), \hat{p}_{t,k}(\hat{z}_t)\}
$$

where $s$ denotes the HMM state during the Viterbi decoding.

The proposed mpdf-level smoothing method selects the more probable Gaussian mixture component from between the original and fMLLR-transformed components.

4.3. Combination of fMLLR and MLLR

In order to perform the feature adaptation, we first estimated a set of transformation matrices by using non-native speech adaptation data, where the number of classes for the transformation matrix estimation was set to 128. We then performed a traditional fMLLR adaptation method with/without the proposed mpdf-level and opdf-level fMLLR smoothing methods.

Table II compares the average WERs of the baseline ASR system employing a traditional fMLLR adaptation method with/without the proposed smoothing methods. Table II compares the average WERs of the baseline ASR system employing a traditional fMLLR adaptation method with/without the proposed smoothing methods. Table II compares the average WERs of the baseline ASR system employing a traditional fMLLR adaptation method with/without the proposed smoothing methods. Table II compares the average WERs of the baseline ASR system employing a traditional fMLLR adaptation method with/without the proposed smoothing methods. Table II compares the average WERs of the baseline ASR system employing a traditional fMLLR adaptation method with/without the proposed smoothing methods.
speech [9][13][14]. Table III compares the average WERs of an ASR system employing the traditional MLLR adaptation and the traditional fMLLR adaptation with/without the proposed mpdf-level and opdf-level fMLLR smoothing methods. It was shown from the table that an ASR system adapted with the MLLR adaptation provided a relative WER reduction of 64.51% for non-native speech when compared to the baseline ASR system. However, the WER of the ASR system was increased after combining FMLLR. This result could be explained by the fact that the ASR system was over-fitted to the non-native speech adaptation data when the traditional FMLLR and MLLR adaptations were combined. Conversely, ASR systems employing a combination of traditional MLLR and FMLLR with the proposed opdf-level and mpdf-level FMLLR smoothing methods achieved relative WER reductions of 11.60% and 10.61%, respectively, for non-native speech when compared to the ASR system employing only the traditional MLLR adaptation method.

### 6. CONCLUSION

In this paper, we proposed FMLLR smoothing methods applied to non-native speech recognition in order to mitigate potential mismatches between adaptation and test data, caused by various influences in the pronunciation of non-native speakers. The proposed smoothing methods were performed with a Gaussian mixture probability density function (mpdf) and an observation pdf (opdf) during the Viterbi decoding. It was shown from non-native speech recognition experiments with a Korean-spoken English continuous speech corpus that ASR systems employing a traditional fMLLR adaptation with the mpdf-level and opdf-level FMLLR smoothing methods could relatively reduce the average WERs by 30.65% and 29.82%, respectively, when compared to a traditional fMLLR adaptation method. Moreover, when combining FMLLR and MLLR, an ASR system using the proposed smoothing methods provided relative WER reductions of around 11.6% and 31.7% compared to using only the MLLR adaptation and only the fMLLR adaptation with no smoothing, respectively.

### 7. ACKNOWLEDGEMENTS

This work was supported in part by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-314-D00245), and by the Ministry of Knowledge and Economy, Korea, under the ITRC support program supervised by the National IT Industry Promotion Agency (NIPA-2009-C1090-0902-0010).

### 8. REFERENCES