Low-memory Fast On-line Adaptation for Acoustically Mismatched Children’s Speech Recognition

S Shahnawazuddin and Rohit Sinha

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati, India.

{s.syed,rsinha}@iitg.ernet.in

Abstract

This work focuses on the issues and the challenges in acoustic adaptation in context of on-line children’s speech recognition. When children’s speech is decoded on adults’ speech trained acoustic models, severely degraded recognition performance is noted on account of extreme acoustic mismatch. Though a number of conventional adaptation techniques are available, they are found to be undesirably latent for an on-line task. For addressing the same, in this work we have combined two low complexity fast adaptation techniques, namely acoustic model interpolation and low-rank feature projection. Two schemes for doing the same are presented in this work. In the first approach, model interpolation is done using weights estimated in unconstrained fashion. The other approach is a hybrid one in which a set mean supervectors are pre-estimated using suitable developmental data. Those are then optimally scaled using the given test data. Though the unconstrained approach results in better improvements over baseline, it has a higher complexity and memory requirements. In case of the hybrid approach, for interpolating M models, the number of parameters to be estimated and memory requirements are reduced by a factor of (M − 1).

Index Terms: Speech recognition, acoustic mismatch, feature projection, fast adaptation.

1. Introduction

Automatic speech recognition (ASR) systems trained on adults’ speech exhibit highly degraded performance when used for recognizing children’s data. The same has been established in earlier reported works [1, 2, 3]. The major reasons for the same are that the children’s speech has higher fundamental and formant frequencies and greater spectral variability than the adults’ case as reported in [4, 5]. Several techniques addressing the mismatch have been reported over the years [2, 6, 7, 8]. Among those, the spectral warping of the children speech features to the training speech space using vocal tract length normalization (VTLN) [9] is found to significantly improve the recognition performance [10]. Spectral smoothing of the speech features was suggested for mismatch reduction in [11, 12]. As argued in [12], the higher order cepstral coefficients are affected more by the pitch dependent distortions in case of children’s speech. Consequently, additional spectral smoothing through binary weighting (BW) of the higher order cepstral coefficients was suggested to overcome the same.

A major drawback of the BW scheme is the complete loss of information when a particular dimension is suppressed to achieve spectral smoothing. This can be addressed by a soft-weighting (SW) of cepstral coefficients in place of BW. In this work, SW has been implemented using a structured low-rank projection of cepstral features using principal component analysis (PCA) [13]. In comparison to the BW scheme, a greater amount of spectral information is conserved during the SW based projection. As a result, the SW approach is found to outperform the BW technique.

In this paper, we also explore the effectiveness of model interpolation based fast acoustic adaptation for an on-line children ASR. In such adaptation tasks, the available adaptation data is very low and the latency in system response is a major factor. Consequently, the employed adaptation technique should have a low complexity. The fast adaptation techniques are found to better suit such tasks as those are required to compute a very few parameters. Apart from exploring the effectiveness of fast adaptation approaches, we have also presented a scheme to further reduce the involved complexity. An efficient amalgamation of SW based feature projection and model interpolation technique is proposed. Given the test data, the proposed adaptation technique requires the estimation of only two parameters to derive the adapted model parameters. Despite the introduced constraints, significant improvements over the baseline system are obtained. In addition to that, it is ensured that the overall memory requirements remain amenable to the on-line system.

This paper is organized as follows: The experimental setup for the evaluation of the explored schemes is given in Section 2. In Section 3, a brief review of the model interpolation based fast adaptation is presented. The proposed on-line adaptation approach is outlined and experimentally evaluated in Section 4. Finally the paper is concluded in Section 5.

2. Experimental setup

An ASR system is developed on the WSJCAM0 British English speech corpus [14] using the HTK speech recognition toolkit [15]. The WSJCAM0 database consists of 15.5 hours of speech data from 92 adult speakers. A Hamming window of length 25 ms with frame rate of 100 Hz and a pre-emphasis factor of 0.97 is employed for speech data analysis. The 13-dimensional MFCC base features are computed employing 21-channel Mel-filterbank. The first and second order temporal derivatives are also appended. The speaker independent (SI) ASR system is trained using cross-word tri-phone acoustic modeling along with decision tree based state tying. Each triphone is modeled using a 3-states left-to-right HMM with 8 diagonal covariance Gaussian components per state. Similarly, a 3-states HMM with 16 diagonal covariance Gaussian components per state is used for modeling silence and a short-pause. PFs, the test set of PFSTAR British English children speech database [16], is used for evaluating the effectiveness of the explored approaches in the mismatched conditions. This set con-
contains 1.1 hours of speech data from 60 children which amounts to a total of 5067 words. To counter the significant differences in the word-list and word frequencies across the adults and children datasets, a 1.5k bi-gram LM is trained on the transcripts of speech data in PFSTAR excluding PFTs having a perplexity of 1.02% of OOV. The word error rate (WER) metric is used as a measure of recognition performance. The WERs for the unadapted SI system, under mismatched (children’s test data recognized using acoustic models trained on adults’ speech) condition turns out to be 64.24. It is to note that this poor WER has resulted in spite of the use of domain specific LM. To simulate on-line ASR task, unsupervised utterance-specific adaptation is performed. The first-pass hypothesis is derived by decoding the given test utterance using the SI models. On-line adaptation is performed under the constraints of the first-pass transcription. The developmental data is derived from the training portion of PFSTAR corpus. All speech data is re-sampled to 8 kHz rate.

3. Fast adaptation of acoustic models

The acoustic model interpolation based fast adaptation approaches, as discussed earlier, are quite promising in context of on-line ASR tasks. The basic assumption in all such approaches is that the adapted Gaussian means lie in a low dimensional sub-space. Therefore, the adapted GMM-HMM mean parameters can be derived by a linear combination of a set of predefined acoustic models (bases). In these approaches, the \( \theta \)th adapted Gaussian mean vector \( \mathbf{m}^\theta \) is derived as follows:

\[
\mathbf{m}^\theta = \alpha_0 \mathbf{1} + \beta_1 \mathbf{a}_1 + \ldots + \beta_M \mathbf{a}_M
\]

(1)

where \( \alpha_0 \) is an optional bias vector and \( \{ \mathbf{a}_j \}_{j=1}^M \) is the set of predefined bases. The remaining GMM-HMM parameters do not change and are borrowed from the SI system. Consequently, unlike MAP/MLLR, only a few interpolation weights (direction coordinates \( \{ \beta_j \} \)) are required to be estimated. These interpolation weights are global in nature, i.e., the same set of weights are applied to all the Gaussians. Due to these constraints, even a small amount of adaptation data happens to be sufficient. Reduction in the number of parameters also implies low latency which is highly desirable in on-line ASR tasks. In the following we have discussed a few of the fast adaptation approaches reported in literature. We have also discussed their comparative effectiveness in context of on-line children’s ASR task under mismatched condition.

3.1. Review of explored adaptation techniques

In the Eigenvoices (EV) based fast adaptation approach [17], a set of universal speaker models called eigenvoices are created using the training data. First, a speaker adapted (SA) model is created for each of the speakers in the training set by MAP/MLLR adaptation of the SI model means. The mean vectors of all the Gaussians are stacked in form of a mean supervector of very high dimensionality. One mean supervector is created from each of the SA models. The eigenvoices are then obtained by choosing the top \( M \) eigenvectors through PCA performed on the correlation matrix constructed out of the adapted mean supervectors. If \( \mathbf{A} \) denotes the matrix of the extracted \( M \) eigenvoices, \( \{ \mathbf{a}_j \}_{j=1}^M \), the Gaussian mean for a test speaker is then modeled using (1). In this case, \( \alpha_0 \) and \( \alpha_j \) represent the \( \theta \)th Gaussian component in SI mean supervector and \( j \)th eigenvoice, respectively. The interpolation weights are estimated using maximum likelihood eigen-decomposition (MLED) [17].

Another model interpolation based approach reported in literature is the reference speaker weighting (RSW) technique [18] which is an extension of the work reported in [19]. In RSW, the mean adapted SA models, derived as explained earlier, are used. Given the test data, a maximum-likelihood (ML) approach using Viterbi-based alignment is employed for selecting the SA models prior to their interpolation. The mean vectors of the models are interpolated without the use of bias.

A low complexity model interpolation approach employing sparse representation (SR) is presented in [20]. In that approach, two dictionaries, \( D_1 \) and \( D_2 \), are created. The columns of \( D_1 \) are the mean supervectors extracted from the SA models. PCA is done on the correlation matrix derived out of \( D_1 \) and the resulting eigenvectors constitute the columns of \( D_2 \). A target supervisor is derived by adapting the SI mean parameters using the given test data and denoted as \( z \). To obtain the adapted Gaussian means, two separate sparse codings (\( z_1 \) and \( z_2 \)) of \( z \) are derived over \( D_1 \) and \( D_2 \) and the sparse coded targets are then jointly scaled in ML sense. The sparse coding helps in compressing the relevant acoustic information captured by the dictionaries into two supervectors. The scaling ensures that the adapted means are optimal in ML sense since the employed greedy SR techniques are not optimized using ML criteria. This technique greatly reduces the complexity of estimating interpolation weights using the ML criterion as the number of bases reduces to two after sparse coding.

To address the acoustic mismatch, we explored the effectiveness of the discussed adaptation approaches. To simulate an on-line ASR task, unsupervised utterance-specific adaptation experiments are performed. The first-pass hypothesis generated using the SI system is used for computing the statistics required for basis selection and weight estimation. The number of bases interpolated is varied and the best case WERs for each of the techniques are given in Table 1. In case of children ASR, VTLN is reported to be greatly effective. So, the explored approaches are combined with VTLN to further reduce the WERs. VTLN warped MFCC features corresponding to frequency warping factors lying in the range of 0.88-1.12 and in steps of 0.02 are computed for the given test data. To find the optimal warp factor, an ML grid search is then performed by aligning the different warped test features against the SI models under the constraints of the first-pass hypothesis. The optimally warped MFCC features are then employed during the second-pass decoding using the adapted models derived through the model interpolation techniques. The WERs for this study are also enlisted in Table 1.

It is to note that significant improvements in recognition performances over the corresponding baselines are obtained in both the cases. Moreover, the EV and the SR based approaches result in almost similar performances. Though the SR technique
has the least complexity in terms of interpolation weight estimation, the memory requirement is more in comparison to the EV approach. In case of EV, along with the SI model parameters, $M$ eigenvectors are required to be stored. For the SR based technique, all the mean supervectors as well as the corresponding eigenvectors are to be stored. In the following, we present an on-line adaptation approach to handle both the issues of complexity and memory.

4. Proposed on-line adaptation technique

As already discussed, the BW/ SW scheme suppresses the higher order cepstral coefficients to reduce the acoustic mismatch. Let $P_{K \times D}$ denote such a transformation matrix to be applied to the $D$-dimensional base MFCC feature vector such that the contributions from the higher order coefficients get deemphasized in the resulting $K$-dimensional feature vector. Since, derived velocity and acceleration components of the base MFCC features also exhibit a similar nature, the same transform is applied to those components as well. The structure of the transformation matrix $P$ applied to the 3D-dimensional test speech feature vector is given as

$$
P_{3K \times 3D} = \begin{bmatrix}
P & 0 & 0 \\
0 & P & 0 \\
0 & 0 & P
\end{bmatrix}.
$$

(2)

In case of SW, $P_{K \times D}$ is estimated by performing PCA on the correlation or covariance matrix derived from the $D$-dimensional base features corresponding to the entire adults’ training speech. The mean vectors and the covariance matrices of the SI system are transformed using $P$ to derive the set of corresponding model parameters $\Lambda$ representing a lower dimensional subspace. The Gaussian mixture-weights and the state transition matrices in $\Lambda$ remain the same as the SI system. In the proposed adaptation technique, SW based feature projection is combined with the EV technique. This can be achieved in two different ways. In the following, the first of the two techniques is outlined in context of an on-line ASR task. It is followed by the description of the second technique to further reduce the overall complexity as well as the memory.

4.1. Unconstrained scheme

In the work presented in [10], it is shown that the rank of the BW transform is correlated with VTLN warp factor of the given test data with respect to the SI system. This relation can be employed in case of an on-line ASR task to automatically select the rank of the transformation to be applied. Motivated by that, we have also made a similar study in case of SW. Given the VTLN warp factor for the children’s test data, the rank of the transformation is selected as per the look-up table shown in Table 2. During the training phase, the SI system is transformed using different transformation matrices $\tilde{P}_{(\alpha_i)}$ corresponding to warp factor $\alpha_i$ values varying from 0.88 to 0.98 in steps of 0.02 to derive $\Lambda_{(\alpha_i)}$. A set of $M$ eigenvoices $\{a_j^{(\alpha_i)}\}_{j=1}^M$ is derived from $\Lambda_{(\alpha_i)}$ using whole training data transformed using corresponding $\tilde{P}_{(\alpha_i)}$ as explained in Section 3.1. During testing, warping factor $\bar{\alpha}_i$ is determined for the given test data under the constraints of the first-pass hypothesis. Using the look-up table, the corresponding set of eigenvoices are selected. Finally, the mean vector for the $r^{th}$ Gaussian in the adapted acoustic model is derived as

$$
m^{(\alpha_i)} = a_0^{(\alpha_i)} + \sum_{j=1}^{M} \beta_j^{(\alpha_i)} a_j^{(\alpha_i)} \tag{3}
$$

where $a_0^{(\alpha_i)}$ is the corresponding mean vector in the SI system transformed using $\tilde{P}_{(\alpha_i)}$. The global weights $\beta_j^{(\alpha_i)}$ are estimated using MLED. In model interpolation based approaches, the complexity of interpolation weight estimation process depends on the number of bases being interpolated and dimensionality of the feature vectors. If $K$ is the number of Gaussians in the acoustic model, $M$ is the number of bases being used and $D$ is the dimensionality of MFCC features, the computational complexity is $O(P[K^2D + MD])$. SW based feature projection reduces the value of $D$. For the PFs test set, the warping factors for most of the test utterances are found to lie in the range of 0.88-0.92. Consequently, the employed MFCC feature vectors and model parameters happen to be 21-27 dimensional. This, in turn, implies huge reduction in computation during the estimation of $\beta_j^{(\alpha_i)}$ using MLED. At the same time, the outlined scheme requires a lot more memory for storing the model parameters than the unadapted SI system. The $M$ eigenvoices corresponding to the various values of $\alpha_i$ are required to be saved. Such an increase in memory requirements is not suitable for an on-line ASR system. In addition to that, $M$ interpolation weights are also required to be estimated for each of the test utterances. This introduces latency as already discussed. In the following we describe an scheme to address both these issues.

4.2. Hybrid scheme

In order to further reduce the latency, we explored the possibility of deriving the interpolation weights a priori using some developmental data and exploit them during testing. To do so, the PFSTAR training data set is split into groups corresponding to their optimal warp factor estimated with respect to the SI system. For each warp factor $\alpha_i$, the interpolation weight vector $\beta^{(\alpha_i)}$ is then derived via MLED using labeled children’s developmental data corresponding to that warp factor. The adapted mean vector for the $r^{th}$ Gaussian is then modeled as

$$
m^{(\alpha_i)} = a_0^{(\alpha_i)} + \sum_{j=1}^{M} \beta_j^{(\alpha_i)} a_j^{(\alpha_i)} \tag{4}
$$

The pre-estimated mean vectors for all the Gaussians in the acoustic model, corresponding to each $\alpha_i$, are concatenated in the form of a supervector $m^{(\alpha_i)}$ and stored. During testing, a mean supervector is chosen on the basis of the warping factor for the given test data. This supervector is used along with the other parameters in $\Lambda_{(\alpha_i)}$ during decoding. Employing the warp factor based indexing of mean supervectors needs the estimation of a single parameter, i.e., the optimal value of $\alpha_i$ for the test data. With the proposed scheme, the entire process of deriving the model parameters becomes off-line. Moreover, only

<table>
<thead>
<tr>
<th>VTLN Warp factor, $\alpha$</th>
<th>Order of proj. matrix</th>
<th>VTLN Warp factor, $\alpha$</th>
<th>Order of proj. matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.98</td>
<td>36 $\times$ 39</td>
<td>0.92</td>
<td>27 $\times$ 39</td>
</tr>
<tr>
<td>0.96</td>
<td>33 $\times$ 39</td>
<td>0.90</td>
<td>24 $\times$ 39</td>
</tr>
<tr>
<td>0.94</td>
<td>30 $\times$ 39</td>
<td>0.88</td>
<td>21 $\times$ 39</td>
</tr>
</tbody>
</table>

*Table 2: A look-up table for selecting the order of the feature projection matrix based on the VTLN warp factor.*
In this work, we have explored the issues of acoustic mismatch in context of an on-line children’s ASR task. Towards addressing the high latency with existing adaptation approaches, a couple of low complexity on-line adaptation schemes have been proposed. Acoustic model interpolation and low-rank feature projection are combined to perform adaptation. The proposed approach not only reduces the latency but also ensures that the overall storage requirements are kept low. Given the test data, only two parameters are required to be estimated. Even with such stringent constraints, a relative improvement of 11.4% over the baseline is obtained for the on-line ASR task.

4.3 Evaluation results

The recognition performances for the proposed approaches with varying amount of children’s developmental data used in deriving the pre-estimated model parameters are plotted in Figure 2. Recognition performances are shown for the unconstrained as well as the hybrid schemes along with those of the two feature projection approaches (BW and SW) without any model interpolation. Its evident that the SW technique is superior to the BW approach. Moreover, the combination of SW with EV is quite efficient in both the proposed schemes. It is to note that when the amount of development data used is very low, using the predefined Gaussian means without scaling does not lead to any improvement over the SW technique. A global scaling of the pre-estimated model means using the test data, leads to consistent improvements over the unscaled case. Obviously, the unconstrained interpolation yields the best performance.

5. Conclusion

In this work, we have explored the issues of acoustic mismatch in context of an on-line children’s ASR task. Towards addressing the high latency with existing adaptation approaches, a couple of low complexity on-line adaptation schemes have been proposed. Acoustic model interpolation and low-rank feature projection are combined to perform adaptation. The proposed approach not only reduces the latency but also ensures that the overall storage requirements are kept low. Given the test data, only two parameters are required to be estimated. Even with such stringent constraints, a relative improvement of 11.4% over the baseline is obtained for the on-line ASR task.
6. References


