A Posteriori and a Priori Transformations for Speaker Adaptation in Large Vocabulary Speech Recognition Systems

Driss Matrouf, Olivier Bellot, Pascal Nocera,
Georges Linares, Jean-François Bonastre

Laboratoire d’Informatique d’Avignon
LIA, Avignon, France
{driss.matrouf, olivier.bellot, pascal.nocera, georges.linares, jean-francois.bonastre}@lia.univ-avignon.fr

Abstract
The speaker-dependent HMM-based recognizers gives lower word error rates in comparison with the corresponding speaker-independent recognizers. The aim of speaker adaptation techniques is to enhance the speaker-independent acoustic models to bring their recognition accuracy as close as possible to the one obtained with speaker-dependent models. In this paper, we propose a method using training and test data for acoustic model adaptation. This method operates in two steps. The first one performs an a priori adaptation using the transcribed training data of the closest training speakers to the test speaker. This adaptation is done with MAP procedure allowing reduced variances in the acoustic models. The second one performs an a posteriori adaptation using the MLLR procedure on the test data, allowing mapping of Gaussians means to match the test speaker’s acoustic space. This adaptation strategy was evaluated in a large vocabulary speech recognition task. Our method leads to a relative gain of 15% with respect to the baseline system and 10% with respect to the conventional MLLR adaptation.

1. Introduction
The speaker-dependent HMM-based recognizers have lower Word Error Rates (WER) than speaker-independent ones. In fact, modeling inter-speaker variability is usually performed by training acoustic models with as large as possible population of speakers. This training manner leads to a relative high variance in acoustic models and hence reduces discriminatory capabilities between different phonemes, especially in the context of larger perplexity tasks. Nevertheless, in the speaker-dependent case, the large amount of required training data for each test speaker reduces the utility and portability of such systems.

To deal with inter-speaker variability, two classes of approaches have been studied. The first one consists in performing normalization in the feature space. This class contains the cepstral mean removal technique [1], the vocal track length normalization [2], a feature space normalization based on mixture density HMM [3], and a signal bias removal estimated by maximum likelihood estimation (MLE) [4].

The second class of approaches operates in acoustic model space. In this class a compact model for speaker-adaptive training (SAT) technique was introduced in [5]. This technique consists in modeling separately the speaker variation and removing its effect in the training data. Thus the variance of models is reduced and hence the overlap of the acoustic models. The most used techniques in the second class consist in adapting the speaker-independent models to a specific speaker so as to obtain a recognition accuracy as close as possible to the one obtained on speaker-dependent system. In this framework, many adaptation schemes have been proposed: in [6] Maximum A Posteriori (MAP) estimations techniques were proposed. It attempts to obtain a Bayesian estimate of the model parameters using adaptation data available from the test speaker. In [7] the speaker-independent system is transformed to come closer to the test speaker by applying a linear transformation on the means of speaker-independent Gaussians. The transformation is estimated so as to maximize the likelihood of the test speaker’s adaptation data.

Other adaptation schemes are based on the fact that the training data contains a number of training speakers, some of whom are acoustically closer, to the test speaker, than the others [8]. This technique use the adaptation data to find a subset of the training speakers which are closer to the test speaker. And then, it compute and apply a linear transformation to map the acoustic space of each selected training speaker closer to the test speaker’s acoustic space. The linear transformation is computed by using the MLLR procedure [9].

In this paper, we propose a method using training- and test- data for acoustic model adaptation. There are two steps in this method: the first one performs an a priori adaptation using transcribed training data with MAP adaptation. The second one performs an a posteriori adaptation using test data with MLLR adaptation. Both modifications have different goals: the former allows a
reduced variance in acoustic models whereas the latter allows a mapping of the acoustic models means to be closer to the test speaker’s acoustic space.

In the next section we present the proposed adaptation method. We describe the goals and the strategies for the a priori and the a posteriori speaker adaptations. In section 3 we describe two strategies for training-speakers selection. Section 4 shows results for several recognition experiments in large vocabulary task framework. The last section (5) is dedicated to some conclusions and comments.

2. Adaptation process

Because of the inter-speaker variability modeling, the speaker-independent models have a relative large variance in comparison with the corresponding speaker-dependent models. By using the MLLR adaptation we only adapt the Gaussians means, so the resulting acoustic models still have a relative high variance and hence an high overlap among different speech units, resulting in reduced discriminative capabilities. To reduce variances, one may use the MAP (Maximum A Posteriori) adaptation [6]. But this process requires a relative large amount of adaptation data to re-estimate all Gaussians variances.

In this paper, we propose a strategy resulting in adapted acoustic models with reduced variances. The adaptation is performed in two steps (see Figure 1). The first adaptation step, that we term the a priori adaptation is based on selecting a cluster of training speakers who “are” acoustically close to the test speaker. Then the speaker-independent acoustic models are adapted by using the transcribed training data corresponding to those selected speakers. This first adaptation is done with the MAP procedure, which transforms the means, variances and gains of Gaussians: let \( \mu_g \), \( \Sigma_g \) as mean and variance in the speaker-independent acoustic models. The new mean \( \bar{\mu}_g \) and variance \( \bar{\Sigma}_g \) of the Gaussian \( g \) are given by:

\[
\bar{\mu}_g = \frac{\eta_g + \tau_g \mu_g}{c_g + \tau_g}, \quad (1)
\]

\[
\bar{\Sigma}_g = \frac{1}{c_g + \tau_g} \left[ \gamma_g + \tau_g [\Sigma_g + \mu_g \mu_g^T] \right] - \bar{\mu}_g \bar{\mu}_g^T, \quad (2)
\]

Where

\[
c_g = \sum_t c_g(t), \quad (3)
\]

\[
\eta_g = \sum_t c_g(t|x_t), \quad (4)
\]

\[
\gamma_g = \sum_t c_g(t|x_t)c_t^T, \quad (5)
\]

The parameter \( \tau_g \) is usually chosen to be constant. \( c_g(t) \) is the a posteriori probability of the Gaussian \( g \) at time \( t \), conditioned on all acoustic observations \( x_{t-1}...T \).

This first processing step adapts all Gaussians parameters, however, only variances and gains Gaussians adaptations can effectively improve the modeling capabilities of the system for a specific test speaker. In fact, really adapting the Gaussians means compared to a specific speaker, only relying on training speakers requires a very large population of speakers. The selected training speakers cluster is then accordingly closer to the test speaker. By using a not so large training speakers population (120 speakers), the spectral variation caused by the inter-speaker variability in each speech unit is reduced, but the Gaussians means remains unadapted to the test speaker.

The second processing step consists in adapting the Gaussians means of the acoustic models resulting from the first adaptation step. This last adaptation is done by using the MLLR procedure [9]: the test data is decoded by using the reduced variances acoustic models (the a priori adapted models). Then the resulting frame/state alignment is used to estimate a global linear transformation, which is applied to the Gaussians means of the a priori adapted acoustic models. We term this second step adaptation the a posteriori adaptation.

3. Speaker clustering

To perform the a priori adaptation we need to find a subset of the training speakers who are the closest to the test speaker. This is done with the LIA speaker recognizer, AMIRAL [10], based on Gaussian Mixture Models (GMM). For this task we used a GMM with 128 Gaussians for each training speaker. The system compares all the training speakers to the test speaker, and then order these speakers from nearest to farest. So, the transcribed training data of the \( n \) nearest speakers are selected to adapt the speaker-independent acoustic models with the MAP procedure.

Another strategy to select the training speakers based on HMM instead of GMM was tested. Firstly, we constructed 120 training-speaker dependent HMMs. However, the data available from each training speaker are usually not sufficient to obtain robust estimations of the speaker dependent model parameters. So, we used MAP procedure [6] to adapt the speaker-independent models to each training speaker, and hence obtain 120 HMMs representing each of the training speakers. It was then required to find the subset of the closest training speakers to the test speaker. The test data are decoded using a speaker-independent system leading to frame/state alignment. Then the acoustic likelihood of test data, conditioned on this alignment, is computed using each of training speaker dependent HMM. The top \( n \) speakers are then selected as the acoustically closest training speakers to the test speaker.

The two strategies give always the same five first
close training speakers to the test speaker. This adaptation is done with the MAP procedure. In the Table 1, we term this adaptation Adapt. 1. The a posteriori adaptation is performed by using the test signal with the MLLR procedure. In the Table 1, we term this adaptation Adapt. 2. Both adaptations will be compared to the MLLR-only one, applied on the test data. We term this last adaptation Adapt. 3. For MLLR, we used 1 global linear transformation with an offset.

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Baseline</th>
<th>Adapt. 1</th>
<th>Adapt. 2</th>
<th>Adapt. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER</td>
<td>26.2</td>
<td>25.4</td>
<td>22.4</td>
<td>24.9</td>
</tr>
</tbody>
</table>

Table 1: Word Error Rate (%) Comparisons: Adapt. 1: a priori adaptation with MAP, Adapt. 2: a posteriori adaptation with MLLR on models obtained in Adapt. 1, Adapt. 3: MLLR on the baseline acoustic models.

We can see that the a priori adaptation using MAP (adapt. 1 in Table 1) doesn’t improve significantly the word error rate (only 3% relative gain with respect to the baseline system). However, this step is important because its conjunction with the MLLR procedure leads to a relative gain of 15% with respect to the baseline system (compare adapt. 2 and adapt. 1). This fact is due to a smaller variance of acoustic models in a priori adaptation (MAP). Then the MLLR better maps the means of acoustic Gaussians model to match the test data.

In our experiments, the relative gain obtained by using the MLLR (1 global transformation) with respect to the baseline system is about 5% (from 26.2% to 24.9%). This gain is 3 times smaller than the one obtained by conjunction of the a priori and a posteriori adaptations. The relative gain obtained by the a priori and a posteriori adaptations with respect to the conventional MLLR is about 10%.

Figure 1: Schematic diagram of adaptation process.

4. Experimental results

In this section, we present the results of several recognition experiments. These experiments were conducted using SPEERAL, a large vocabulary speech recognition system, developed at the LIA. The lexicon size is about 20k words with 3.6% out-of-vocabulary words. This system uses a trigram language model. The baseline system is speaker and gender independent. The acoustic model contains 38 phonemes. Each phoneme is a 3-state left-to-right context-independent CDHMM (Continuous Density HMM). Each state is a mixture of 64 Gaussians. The signal speech is parameterized using 13 coefficients, 12 mel-warped cepstral coefficients plus energy. The first and second order derivatives parameters are also used.

To estimate the acoustic and linguistic models, we have used a training data extracted from Bref [2], with 120 speakers (66 females and 54 males). The training data contain 66.5k sentences. The test data were provided for ARC B1 of AUPELF [1], with 20 speakers and 299 sentences. The sentences are articles published in the french newspaper “Le Monde”.

The a priori adaptation is performed by using the 5
5. Conclusion
We have presented a new acoustic model adaptation strategy using test and training data. This adaptation strategy is divided in two steps. The first one performs an \textit{a priori} adaptation using transcribed training data with MAP procedure. The second one performs an \textit{a posteriori} adaptation using the test data with MLLR procedure. The goal of the \textit{a priori} adaptation is to reduce the variance in acoustic models whereas the goal of the \textit{a posteriori} adaptation is to map the acoustic model Gaussians means to be closer to the test speaker’s acoustic space.

We have evaluated the proposed adaptation strategy in large vocabulary continuous speech recognition task. We conducted experiments demonstrating that the conjunction of the \textit{a priori} and \textit{a posteriori} adaptations is more efficient than the \textit{a priori} adaptation (MAP) alone or the \textit{a posteriori} adaptation (MLLR) only. The relative gain obtained by the conjunction of the two adaptations is about 10% with respect to the conventional MLLR, 15% with respect to the baseline error rate, and 12% with respect to the \textit{a priori} adaptation alone.

In these experiments we have chosen arbitrarily five selected training speakers. Further studies must be done to determine the optimal amount of training data to be used for \textit{a priori} adaptation.

6. References