SEQUENTIAL UBM ADAPTATION FOR SPEAKER VERIFICATION

Jun Wang, Dong Wang, Xiaojun Wu, Thomas Fang Zheng*

Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology; Center for Speech and Language Technologies, Research Institute of Information Technology, Tsinghua University; Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China

* Corresponding Author: fzheng@tsinghua.edu.cn, Tel/Fax: +86-10-62796393

ABSTRACT

GMM-UBM-based speaker verification heavily relies on a well trained UBM. In practice, it is not often easy to obtain an UBM that fully matches acoustic channels in operation. To solve this problem, we propose a novel sequential MAP adaptation approach: by being sequentially updated with data from new enrollments, the UBM learns and converges to the working channel. Our experiments are conducted on a time-varying speech database, with two channel-mismatched UBMs as the initial model. The results confirm that the sequential UBM adaptation provides significant performance improvement, leading to a relative EER reduction of 6.3% and 14.8% for the two mismatched UBMs, respectively.

Index Terms— UBM, MAP, speaker verification

1. INTRODUCTION

The GMM-UBM framework is widely used in speaker verification [1]. This approach relies on a well-trained universal background model (UBM) to represent the general speaker space, and each enrolled speaker is represented by a Gaussian mixture model (GMM) which is adapted from the UBM via maximum a posterior (MAP) estimation [2].

A basic assumption of the GMM-UBM approach is that the UBM is able to represent all acoustic and phonetic variations in speech data, so that the deviation of a speaker GMM from the UBM reflects and only reflects speaker characteristics. On the one hand, this requires a large amount of data in UBM training, and on the other hand, the acoustic channels of the training data and the working circumstance should be consistent. In practice, however, it is often difficult, if not impossible, to collect sufficient channel-matched data to train a fully consistent UBM. Furthermore, some practical channels are changing, which fails a pre-trained UBM anyway.

A multitude of research has been conducted to address channel mismatch or session variation within the GMM-UBM framework. This can be categorized into feature transformation [3, 4, 5], model compensation [6, 7] and score normalization [1, 8]. A comprehensive statistical approach was proposed in [9], where the authors model speaker and channel variation as independent variables spanning in low-rank subspace, and then infer channel factors by factor analysis. [10] follows this line, though allows only channel factors spanning in low-rank subspace, leaving speaker space full-rank. This method is augmented in [11] where the authors presented a straightforward interpretation for the subspace method, together with a simple implementation. In [12], various feature and model compensation approaches were investigated, and the conclusion is that adaptation based on low-rank channel subspace (eigen-channels) is highly effective to deal with channel mismatch.

Besides GMM-UBM, channel mismatch was also studied for other verification approaches. For example, [13] proposed to reduce channel impact in neural network based verification systems by eliminating some hidden nodes of the network; [10] presented some feature mapping functions to mitigate channel discrepancy in SVM-based systems.

All the above research requests some prior data to learn certain compensation structures (transforms or eigen-subspace). In situations where the working channel is totally new, or the channel is time-variant, it is usually difficult to collect sufficient training data. This in turn fails most of the existing methods.

In this paper, we propose a sequential UBM adaptation approach to the problem. Specifically, the system starts from an initial UBM (probably channel mismatched); whenever a new speaker enrolles, his/her data are pooled in the form of sufficient statistics. These statistics are used to obtain a speaker-specific UBM via MAP from the initial UBM. The speaker model (a GMM) is then derived from his/her own UBM via the conventional MAP. By this approach, the UBM (and so speaker models adapted based on it) is updated sequentially and gradually, finally converging to the new or dynamic channel with a large amount of enrollments.

The rest of the paper is organized as follows: Section 2 presents the sequential UBM MAP algorithm, Section 3 presents the experimental results, and Section 4 concludes the paper, together with some ideas for future work.
2. SEQUENTIAL UBM ADAPTATION

The proposed sequential approach extensively relies on MAP-based parameter estimation. We therefore first review the MAP theory, and then present the sequential approach in detail.

2.1. Review of MAP estimation

Suppose that a random vector \( x \) follows a normal distribution \( N(\mu, \Sigma) \), where \( \mu \) is the mean vector and \( \Sigma \) is the covariance matrix. Further assuming that the elements of \( x \) are mutually independent, \( \Sigma \) can be represented by its diagonal vector \( \sigma \). Given a set of training data \( X = \{x_i\} \) where the samples are independent, \( \mu \) and \( \sigma \) can be estimated by maximizing the following objective function:

\[
\mathcal{L}(\mu, \sigma) = \log P(\mu, \sigma | X) \\
\propto \log \left\{ \prod_i \mathcal{N}(x_i; \mu, \sigma) P(\mu, \sigma) \right\}.
\]

where \( P(\mu, \sigma) \) is the prior distribution which is often assumed to be Gaussian-Wishart. This is the well-known MAP estimation. In the case where \( \mu \) is predominantly important (such as in speaker verification), \( \sigma \) sticks to the mode of the prior and thus is omitted from the estimation equation. Further assuming the prior over \( \mu \) is diagonal Gaussian:

\[
P(\mu) = N(\mu; \hat{\mu}, \hat{\sigma})
\]

where \( \hat{\mu} \) is the mean vector and \( \hat{\sigma} \) is the diagonal vector of the covariance \( \Sigma \). This leads to the objective function given by:

\[
\mathcal{L}(\mu) \propto \sum_i \log \mathcal{N}(x_i; \mu, \sigma) + \log N(\mu; \hat{\mu}, \hat{\sigma}).
\]

Maximizing this objective with respect to \( \mu \) leads to the following MAP estimation:

\[
\mu = \frac{\sum_i x_i + \frac{1}{N} \hat{\mu}}{N + \frac{1}{\sigma}} \tag{1}
\]

where \( N \) is the number of training samples in \( X \), and both the multiplication and division are element-wise.

When extending to Gaussian mixture models (GMM), each training sample \( x_i \) contributes to a particular Gaussian component \( N_c \) according to the posterior probability that \( x_i \) belongs to \( N_c \):

\[
r_i(c) = \frac{\mathcal{N}(x_i; \mu_c, \sigma_c)}{\sum_m \mathcal{N}(x_i; \mu_m, \sigma_m)}. \tag{2}
\]

Define the following sufficient statistics:

\[
r_c = \sum_i r_i(c) \tag{3}
\]

\[
z_c = \sum_i r_i(c)x_i, \tag{4}
\]

the MAP estimation is given by:

\[
\mu_c = \frac{z_c + \frac{1}{r_c} \hat{\mu}}{r_c + \frac{1}{\sigma}} \tag{5}
\]

where \( r_c \) is a vector with all its elements equal to \( r_c \). Note we have assumed identical priors for all mixture components.

2.2. Sequential UBM adaptation

The motivation of sequential UBM adaptation is simple: whenever a new enrollment occurs, use the enrollment speech data to update the UBM. We hope that the updated UBM better represents the working channel, and with a large amount of enrollment data, the update leads to convergence to the new channel.

In practice, however, this is not as simple as the first glance. Since the new enrollment data involve not only channel properties, but also speaker characteristics, it is difficult to tell whether the adaptation goes to channel or speaker characteristics. In order to deal with this difficulty, we start from a ‘pool and re-estimation’ procedure.

First of all, assume that we have reserved all the data that were used to train the UBM, denoted by \( \hat{X} = \{\hat{x}_i\} \). With a new enrollment \( X = \{x_i\} \), the original and the new data can be pooled together and re-train the UBM, leading to the following result:

\[
\mu_c = \frac{z_c + \hat{\mu}_c}{r_c + \hat{r}_c} \tag{6}
\]

\[
= \frac{z_c + \hat{r}_c \hat{\mu}_c}{r_c + \hat{r}_c} \tag{7}
\]

where \( r_c \) and \( z_c \) have been defined in (3) and (4) respectively. Comparing (6) and (5), we notice that the re-training process can be implemented as an MAP adaptation, by setting the prior distribution such that the mean equals to the mean of the UBM, and the covariance as follows:

\[
\hat{r}_c = \frac{\sigma}{\hat{\sigma}}. \tag{8}
\]

This means that the strength of the prior depends on the data volume used to train the initial UBM. Since \( \hat{r}_c \) is usually large, \( \hat{\sigma} \) should be set to a small value. In other words, the prior is highly strong. This strong prior is important, which ensures that the UBM can be adapted to the working channel in a stable manner. This strong prior also distinguishes the UBM MAP from the conventional speaker MAP used to grow a speaker model: the later usually relies on a much weaker prior to learn speaker idiosyncracy quickly. Note that if the prior is significantly strong, i.e., \( \sigma \rightarrow 0 \), we have \( \mu \rightarrow \hat{\mu} \), which is the conventional static UBM scheme.

The above pool and re-estimation (and equally the UBM MAP adaptation) conducts whenever a new speakers enrolls, leading to a sequential adaptation approach. At each enrollment, the new speaker data are pooled to the sufficient statistics \( r_c \) and \( z_c \), and then (6) is invoked to perform re-estimation. We highlight that the pool and re-estimation approach is a non-bias estimation for \( \mu \) and converges to the
working channel with unlimited enrollment data, so does the
dual sequential UBM MAP.

![Diagram of Sequential UBM MAP adaptation.](image)

Fig. 1. Sequential UBM MAP adaptation.

Fig. 1 illustrates the sequential MAP process. For sim-
plicity, we set the covariance of the prior for UBM and
speaker MAP are both isotropic, i.e., the elements of \( \hat{\Sigma} \)
are equal to a constant, say \( 1/k_u \) for the UBM MAP and \( 1/k_s \)
for the speaker MAP, respectively. Starting from the initial
UBM0, the first enrollment contributes \( z_c \) and \( r_c \), and (6)
is applied to grow UBM1 with \( k_u \). From UBM1, a speaker
model GMM1 is grown via MAP with \( k_s \). For the next en-
rollment, the speech data are pooled to \( r_c \) and \( z_c \), and (6)
is applied again to grow UBM2, based on which the speaker
model GMM2 is grown. This process continues for all the
new enrollments, and the pair (UBM_1, GMM_1) is reserved for
verifying speaker i.

### 3. EXPERIMENTS

#### 3.1. Database and configurations

We conduct the experiments on a time-varying database [14]
which involves 60 speakers (30 males and 30 females). For
each speaker, 100 Chinese sentences were recorded from Jan-
uary 2010 to 2012. The sampling rate of the signals is 8 kHz
and the sample size is 16 bits. The enrollment segments are
about 20 seconds in length, and the verification segments are
of 5-10 seconds. The 16-dimensional Mel frequency cepstral
coefficients (MFCCs) plus their first order derivatives are used
as acoustic features. Both the UBM and speaker models in-
volve 1024 Gaussian components.

In order to test the sequential approach in learning new
channels, we start the experiments with two initial UBMs
which were trained with databases that mismatch the one
used for enrollment and verification: UBM_a which was trained
with three hours of desktop microphone speech data
(45 males and 38 females), and UBM_b which was trained
with six hours of telephone speech data (150 males and 150
females).

#### 3.2. Experimental results

We treat the static UBM approach, i.e., without any adaptation
once the UBM is delivered, as the baseline. The sequential
adaptation approach proposed in Section 2 is then applied.
Note that these two approaches differ only in UBM treatment;
the speaker model adaptation strategy is identical.

According to the discussion in Section 2, the UBM and
speaker MAP are primarily impacted by the adaptation factors
\( k_u \) and \( k_s \). A large \( k_u \) is essential to ensure a stable
UBM update, while \( k_s \) should be small for effective learning
of speaker idiosyncracy. In this experiment we choose four
value for \( k_u \) (90, 180, 270, 360) and three smaller value for
\( k_s \) (0.5, 1 and 2).

The verification performance is evaluated in terms of
equal error rates (EER). The results with UBM_a and UBM_b
as the initial model are presented in Table 1 and Table 2 re-
spectively, where ‘SUBM’ denotes the sequential adaptation
approach. We observe that with a large \( k_u \), i.e., a strong prior,
the sequential adaptation approach outperforms the baseline
(static UBM). In the experiment with UBM_a, the average
relative EER reduction is 6.3\%, and in the experiment with
UBM_b, the reduction is 14.18\%. The relatively larger EER
reduction with UBM_b can be attributed to the fact that UBM_b
is trained with telephone data and so mismatches the working
channel (carbon-button microphone) in a more significant
way.

We also observe that the value of \( k_u \) impacts the sequen-
tial approach significantly. If the value is too small (e.g. 90
in the experiment with UBM_a), the performance is hurt; with a
larger \( k_u \), the sequential approach generally outperforms the
static approach. When comparing UBM_a and UBM_b, we find
UBM_b requests a smaller \( k_u \). This seems conflict our the-
tory since UBM_b is trained with a larger database than UBM_a
therefore \( k_u \) should be larger according to (8). However, con-

<table>
<thead>
<tr>
<th>EER%</th>
<th>( k_u = 0.5 )</th>
<th>( k_u = 1 )</th>
<th>( k_u = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBM (baseline)</td>
<td>11.75</td>
<td>11.92</td>
<td>11.75</td>
</tr>
<tr>
<td>SUBM(( k_u = 90 ))</td>
<td>12.69</td>
<td>12.42</td>
<td>12.05</td>
</tr>
<tr>
<td>SUBM(( k_u = 180 ))</td>
<td>11.67</td>
<td>11.20</td>
<td>11.47</td>
</tr>
<tr>
<td>SUBM(( k_u = 270 ))</td>
<td>11.19</td>
<td>11.05</td>
<td>11.07</td>
</tr>
<tr>
<td>SUBM(( k_u = 360 ))</td>
<td>11.12</td>
<td>11.02</td>
<td>11.05</td>
</tr>
</tbody>
</table>

Table 1. Results with UBM_a as the initial.

<table>
<thead>
<tr>
<th>EER%</th>
<th>( k_u = 0.5 )</th>
<th>( k_u = 1 )</th>
<th>( k_u = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBM (baseline)</td>
<td>10.04</td>
<td>10.22</td>
<td>10.44</td>
</tr>
<tr>
<td>SUBM(( k_u = 90 ))</td>
<td>9.36</td>
<td>9.20</td>
<td>8.83</td>
</tr>
<tr>
<td>SUBM(( k_u = 180 ))</td>
<td>8.77</td>
<td>8.88</td>
<td>8.82</td>
</tr>
<tr>
<td>SUBM(( k_u = 270 ))</td>
<td>8.72</td>
<td>8.84</td>
<td>8.79</td>
</tr>
<tr>
<td>SUBM(( k_u = 360 ))</td>
<td>8.73</td>
<td>8.82</td>
<td>8.79</td>
</tr>
</tbody>
</table>

Table 2. Results with UBM_b as the initial.
sidering the more significant channel mismatch with \( UBM_b \), we notice that it requires a weaker prior to adapt to the new channel quickly.

3.3. Quality of sequential UBM

Another interesting experiment is to investigate quality of the sequentially adapted UBMs themselves. This means that for each new UBM, we retrain all the speaker models and conduct the verification test. The results are shown in Fig. 2. We can see that the quality of the sequentially adapted UBM is improved gradually with more and more enrollment data available. Note this experiment does not reflect any practical scenario, since the enrollment data are usually not reserved for later speaker model retraining. However if we do so, we can obtain substantial gain.

![Fig. 2. Quality of sequentially adapted UBM.](image)

4. CONCLUSIONS

We proposed a novel sequential UBM adaptation method to address the UBM channel mismatch in speaker verification. By adapting an initial UBM with a strong prior whenever a new enrollment is available, the UBM learns and converges to the working channel gradually, leading to improved verification performance. In our experiments, this sequential approach provides a relative EER reduction of 6.3% and 14.8% for two mismatched UBMs, respectively.

This work focuses on enrollment-phase UBM adaptation, though the same approach can be applied to the test phase. Additionally, we are considering to update speaker models in a sequential way.

5. REFERENCES


