A Fast Speaker Adaptation Method using Aspect Model

Seongjun HAHM¹, Akinori ITO¹, Shozo MAKINO¹ and Motoyuki SUZUKI²

¹Graduate School of Engineering, Tohoku University, Japan
²Institute of Technology and Science, The University of Tokushima, Japan

{branden65, aito, makino}@makino.ecei.tohoku.ac.jp, moto@m.ieice.org

Abstract

We propose a fast speaker adaptation method using an aspect model. The performance of speaker independent (SI) model is very sensitive to environments such as microphones, speakers, and noises. Speaker adaptation techniques try to obtain near speaker dependent (SD) performance with only small amounts of specific data and are often based on initial SI model. One of the most important purposes for adaptation algorithms is to modify a large number of parameters with only a small amount of adaptation data. The number of free parameters to be estimated from adaptation data can be reduced by using aspect model. In this paper, we introduce an aspect model into an acoustic model for rapid speaker adaptation. A formulation of probabilistic latent semantic analysis (PLSA) is extended to continuous density HMM. We carried out an isolated word recognition experiment on Korean database, and the results are compared to those of conventional expectation maximization (EM) algorithm, maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR).

Index Terms: Speaker Adaptation, Aspect Model, PLSA, SD Model, SI Model

1. Introduction

When developing a speaker independent (SI) automatic speech recognition system, it is important to account for the wide variability that can be present in any speech waveform. This variability can result from changes in the individual speaker, the noises, the microphone and channel of the recording device. There can be a significant gap in performance between these systems and their speaker adaptive (SA) or speaker dependent (SD) counterparts. The reduction in a system’s error rate between its SI mode and its speaker dependent mode can be more than 50% [1]. Therefore, how to adapt a large number of parameters effectively with only a small amount of data can often be a problem to be faced in adaptation systems.

A method of providing speaker constraint to speech recognition systems that has proven successful is hierarchical speaker clustering [2]. In this approach, similar reference speakers are grouped together into a speaker cluster for which one model is trained. When using speaker clustering, there is a trade-off between robustness and specificity. Large clusters are more general but can be trained more robustly. Smaller clusters can represent more specific speaker types but may lack a sufficient amount of training data required for accurate density function estimation. To overcome this kind of disadvantages the speaker cluster weighting (SCW) [1] models set used for recognition uses weighted combinations of the models from the predetermined set of L different modes sets. In this method, speaker clustering and weighting of speaker clusters are really important.

In this paper, we will examine the Bayesian adaptation method that exploits an aspect model, which is “a mixture of mixture model.” An aspect model is used for clustering and weighting methods and reducing the free parameters to be estimated. We, then, formulate and discuss the potential of the techniques using PLSA [3]. Finally we will show the experimental results using maximum a posteriori (MAP) [4], maximum likelihood linear regression (MLLR) [5], and the proposed model.

2. Review of PLSA

Probabilistic Latent Semantic Analysis (PLSA) is a novel statistical technique for the analysis of two- mode and co-occurrence data, which has applications in information retrieval and filtering, natural language processing, machine learning from text, and in related areas [3]. The basic idea of the Latent Semantic Analysis (LSA) is to map high-dimensional count vectors, such as the ones arising in vector space representations of text documents, to a lower dimensional representation in a so-called latent semantic space. The goal of LSA is to find a data mapping which provides information well beyond the lexical level and reveals semantical relations between the entities of interest. PLSA is the probabilistic approach compared to LSA. PLSA is based on a mixture decomposition derived from a latent class model. Fig. 1 shows the graphical representation of the PLSA.

![Figure 1: Graphical representation of the aspect model in the asymmetric (a) and symmetric (b) parameterization.](image-url)

The model of Fig. 1(a) is represented by following expression.

\[
P(d, w) = P(d)P(w|d) \\
P(w|d) = \sum_{z \in Z} P(w|z)P(z|d)
\] (1)

In this formula, \(w\) is a word, \(d\) is a document, and \(z\) is a latent class. The model of Fig. 1(b) is defined by following expression using Bayes’ rule.

\[
P(z|d) = \frac{P(d|z)P(z)}{P(d)}
\] (2)

\[
P(d, w) = \sum_{z \in Z} P(z)P(d|z)P(w|z)
\] (3)
The standard procedure for maximum likelihood estimation in latent variable models is the expectation maximization (EM) algorithm. Standard calculations yield the E-step equation

\[ P(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')} \]  

(4)

as well as the following M-step formulation:

\[ P(w|z) = \frac{\sum_{d,w} n(d, w) P(z|d, w)}{\sum_{d,w} n(d, w')} P(z|d, w') \]  

(5)

\[ P(d|z) = \frac{\sum_{d,w} n(d, w) P(z|d, w)}{\sum_{d,w} n(d', w) P(z|d', w)} \]  

(6)

\[ P(z) = \frac{1}{R} \sum_{d,w} n(d, w) P(z|d, w) \]  

(7)

\[ R = \sum_{d,w} n(d, w) \]

where \( n(d, w) \) denotes the term frequency, i.e., the number of times \( w \) occurred in \( d \).

3. PLSA Formulation in Acoustic Space

Fig. 2 shows the transition from language model to acoustic model. In language model, \( d \) means a document. The \( d \) of acoustic model means a state trained from a specific speaker and \( x \) means the feature vector.

![Transition from language models to acoustic models.](image)

This transition can be represented by the following formulae.

\[ P(d, x) = P(d)P(x|d) \]

\[ P(x|d) = \sum_{z \in Z} P(x|z)P(z|d) \]  

(8)

\[ P(z|d) = \frac{P(d|z)P(z)}{P(d)} \]  

(9)

\[ P(d, x) = \sum_{z \in Z} P(z)P(d|z)P(x|z) \]  

(10)

The E-step equation for acoustic model is formulated as follows:

\[ P(z|d, x) = \frac{P(z, d, x)}{\sum_{z'} P(z', d, x)} \]  

(11)

\[ = \frac{P(z|d)P(x|z)}{\sum_{z'} P(z'|d)P(x|z')} \]

where \( x \) is a observed feature parameter for estimation, \( d \) is each state, and \( z \) is a latent class. The \( n(d, w) \) of the language model is changed into \( P_c(x^d_i|\lambda_d) \) under the specific acoustic model.

\[ P_c(x^d_i|\lambda_d) = \frac{b_{q^d_i}^{x^d_i} \lambda_d}{\sum_{d'=1}^{N_d} b_{q^d_i}^{x^d_i} \lambda_{d'}} \]  

(12)

In this expression \( x^d_i \) is the \( i \)th feature vector of certain state, \( d, b(\cdot) \) is the observation symbol probability distribution, \( q^d_i \) stands for the state information of obtained by the Viterbi algorithm, \( N_d \) is the number of states across the model, and \( \lambda_d \) is the model under the specific state, \( d \). Followings are M-step formulae. In these formulae, \( x^d_i \) is the \( i \)th feature vector under certain state, \( d \), and \( P_c(\cdot) \) means the probability under empirical acoustic models.

\[ P(x_i|z) = \frac{\sum_d P(z, d, x) P(z|d, x_i)}{\sum_{d, x'} P(z, d, x')} \]  

(13)

\[ = \sum_d \sum_{x'} P_c(x^d_i|\lambda_d) P(z|d, x^d_i) \sum_d \sum_{x'} P_c(x^d_i|\lambda_d) P(z|d, x^d_i) \]  

(14)

\[ P(d|z) = \frac{\sum_d P_c(x^d_i|\lambda_d) P(z|d, x^d_i)}{\sum_d \sum_{x'} P_c(x^d_i|\lambda_d) P(z|d, x^d_i)} \]  

(15)

![Figure 3: The example of the (a) 4-mixture model and (b) aspect models trained from (a).](image)

Fig. 3 shows the example of 4-mixture model and aspect models trained from 4-mixture model. PLSA can be viewed as one of the methods for mixture decomposition. After decomposition, original distribution is not changed. Fig. 4 shows the example of calculating aspect models.

![Figure 4: The example of calculating aspect models.](image)

Fig. 5 shows a comparison of adaptation procedure using mixture and aspect models. Here the weighting of each aspect
model, \( P(z) \), is the unit for adaptation whereas the each mixture is the unit for adaptation in mixture models. For adaptation of the aspect models, EM algorithm is applied for estimating \( \tilde{P}(z) \), updated \( P(z) \).

For estimating global \( P(z_k) \), the following equation is used.

\[
\tilde{P}(z_k) = \frac{\sum_{j=1}^{S} \sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{k=1}^{M} \sum_{j=1}^{S} \gamma_t(j,k)}
\]

(16)

where \( \gamma_t(j,k) \) is the probability of being in state \( j \) at time \( t \) with \( k \)th aspect model. If the global \( P(z) \) is used for adaptation, the \( P(z) \) is the same across all phonemes and states.

For estimating state-dependent \( P_t(z_k) \), the following expression is applied.

\[
P_t(z_k) = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)}
\]

(17)

where \( \gamma_t(j,k) \) is the probability of being in state \( j \) at time \( t \) with \( k \)th aspect models.

The formula for recognition is as follows:

\[
P(x_i|\lambda) = \sum_x P(x_i|d) P(z) P(d|z)
\]

\[
= \sum_x P(x_i|d) P(z) \sum_{d',x'} P(x_i|d') P(z|x')
\]

(18)

This model can be viewed as the combination of the Reference Speaker Weighting (RSW) [1] and Speaker Cluster Weighting (SCW). RSW is an interpolation of models from “reference speakers.” SCW is the cluster’s mixture models. \( P(d|z) \) can be thought of as weighting of reference speakers and \( P(z) \) could be referred to weighting of speaker clusters if we assume that \( P(d|z) \) is the speaker cluster.

The original formulation is intended for each feature vector, \( x \). We, however, used the sum of probability for feature vectors across each word at this time because the memory of computer is not enough for calculating the matrix across entire feature vectors. The expressions for this calculation is as follows:

\[ W = \{ w_1, w_2, \cdots, w_N \}, \]

\[
y_i = \begin{bmatrix} y_{1,1}(i) \\ \vdots \\ y_{1,M}(i) \\ y_{2,1}(i) \\ \vdots \\ y_{N,M}(i) \end{bmatrix},
\]

\[
y_{j,m}(i) = \sum_{t=1}^{y_{j,m}(i)} \gamma_t(i, j, m), 1 \leq i \leq N, 1 \leq j \leq S, 1 \leq m \leq M
\]

(19)

where \( w \) is a word, \( N \) is the number of words for training, \( y(i) \) is the sum of probability for feature vectors across each word, \( S \) is the total number of states, \( M \) is the number of mixtures of each state, and \( \gamma_t(i, j, m) \) is the probability of being in state \( j \) at time \( t \) with the \( m \)th mixture using the feature vector of the \( i \)th word.

4. Experimental Results

4.1. Experimental Conditions

We used the Korean isolated word databases, such as KLE452 databases. 35 males are used for training. For testing and adaptation, 3 speakers not included in training are used. For adaptation, EM, MAP and MLLR are used in addition to the proposed method. For EM, only the weighting value for each mixture is updated. For MAP, weighting value, mean, and variance of each mixture are updated and adjustment parameter \( \gamma \) is set to 50 which is decided empirically. For MLLR, the global transformation matrix is used for adaptation. The number of states, \( D \), is 3885 (i.e., 35 speakers \times 37 phonemes \times 3 states ). The number of the sum of probability for feature vectors across each word is 15820 (i.e., 35 speakers \times 452 words). We added the probability of each feature vector across the words. The conditions for experiments are shown in Table 1 and 2.

Table 1: Analysis Conditions.

<table>
<thead>
<tr>
<th>Feature extraction method</th>
<th>MFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Rate</td>
<td>16kHz</td>
</tr>
<tr>
<td>Pre-emphasis coefficient</td>
<td>0.97</td>
</tr>
<tr>
<td>Window Type</td>
<td>Hamming</td>
</tr>
<tr>
<td>Frame length</td>
<td>25ms</td>
</tr>
<tr>
<td>Feature vector dimension</td>
<td>MFCC + Power (13)</td>
</tr>
<tr>
<td>Cepstral Mean Normalization</td>
<td>Not used</td>
</tr>
</tbody>
</table>

4.2. Speaker Adaptation using global weighting

Using the model of each speaker we made 35-mixture mono-phone model which is the speaker independent model for speaker adaptation. 35-mixture mono-phone model is composed of each speaker model (i.e., \( d = 35 \)). After PLSA, we applied EM algorithm for global \( P(z) \). After adaptation, all phoneme models have the same weighting for aspect model. The experimental results are shown in Fig. 6. We used 2, 3, and 4 aspect models for proposed method. Although the amount of adaptation data is small, the word recognition rate (WRR) of proposed method is better than other methods.
Table 2: Database and Model information.

<table>
<thead>
<tr>
<th>Database</th>
<th>KLE452 (38 speakers/ Phone-balanced 452 isolated words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training data</td>
<td>1st utterances of 35 speakers/ 452 words</td>
</tr>
<tr>
<td>Adaptation data</td>
<td>2nd utterances of 3 test speakers/ 1, 5, 10, 50, 100, 150, 200 words</td>
</tr>
<tr>
<td>Test data</td>
<td>1st utterances of 3 test speakers/ 452 words</td>
</tr>
<tr>
<td>Model type</td>
<td>Mono-phone, 3 states, 35-mixture left-to-right HMM</td>
</tr>
</tbody>
</table>

Figure 6: Fixed word recognition rate using global weighting.

4.3. Speaker Adaptation using state-dependent weighting

In this experiment, we applied EM algorithm independently for each state of each phoneme model. It means that we cannot update weighting value for the phoneme model like EM, MAP in case that the feature vector is not appeared in adaptation phase. The experimental results are shown in Fig. 7.

Figure 7: Fixed word recognition rate using state-dependent weighting.

In case that the number of words for adaptation is more than 50, the word recognition rate is not better than other methods. Using small number of aspect models compared to mixture model, we could obtain improved adaptation results. MLLR and EM methods had worse results than SI model in less than 10 and 50 words, respectively. Poor MLLR performance for the tiny amounts of adaptation data was already reported in [5]. The experimental results show that for small amount of adaptation data PLSA decomposes the mixture models effectively. We could obtain the improved results using only weighting values. However, the number of speaker used database in this paper was not enough relatively. We need to perform experiments using database including larger population and check the results.

5. Conclusions

In this paper, we formulated Probabilistic Latent Semantic Analysis (PLSA) as an acoustic model and evaluated the performance in speaker adaptation. We expected that PLSA have the power of solving these kinds of problems and modeling effectively. Although we use only weighting values by PLSA, we could obtain better results for small amount of adaptation data. Future work is expected to perform experiments on many speakers and various noisy environments.

6. References