EXTRACTION OF RELIABLE TRANSFORMATION PARAMETERS FOR UNSUPERVISED SPEAKER ADAPTATION

Jen-Tzung Chien a, Jean-Claude Junqua b and Philippe Gelin b

a Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
b Speech Technology Laboratory, Panasonic Technologies Inc., Suite #202, 3888 State Street, Santa Barbara, USA
jtcchien@mail.ncku.edu.tw & jcj, gelin@research.panasonic.com

ABSTRACT

Adaptation of speaker-independent hidden Markov models (HMM’s) to a new speaker using speaker-specific data is an effective approach to reinforce speech recognition performance for the enrolled speaker. Practically, it is desirable to flexibly perform the adaptation without any knowledge or limitation on the enrolled adaptation data (e.g. data transcription, length and content). However, the inevitable transcription errors on adaptation data may cause unreliability in model adaptation. The variable amount and content of adaptation data require the algorithm to control the transformation parameters, we exploit the reliability assessment criteria using the confidence measure and description length. Experiments show that the unsupervised speaker adaptation with reliability assessment can significantly improve the recognition performance for any lengths of adaptation data.

1. INTRODUCTION

It is well known that current speaker-independent (SI) speech recognizers without adaptation abilities are very sensitive to speaker variables such as speaker gender, age, accent and emotion. If we could extract some compensation vectors from speaker specific utterances and adapt the existing SI speech models to the new speaker, the resulting speaker-adaptive (SA) speech recognizer could be improved accordingly. Recently, an increasing number of research efforts have been focused on unsupervised adaptation (also known as adaptation without a teacher), in which there is no teacher needed to oversee the adaptation process [4][6-7]. This paradigm is very practical because transcriptions of adaptation data are not required prior to adaptation. However, the performance of unsupervised adaptation heavily depends on the transcription accuracy obtained on the adaptation data. If the transcriptions are exact, it is equivalent to supervised adaptation. But, in real cases, the inevitable transcription errors will cause the adaptation performance to vary greatly. Thus, how to evaluate the reliability of transcriptions becomes a crucial topic for unsupervised adaptation.

In the literature, the transformation-based adaptation, e.g. maximum-likelihood linear regression (MLLR) [5], is popular for speaker adaptation because all the HMM mean vectors can be efficiently adapted (or transformed) via the shared regression functions even though some HMM units are unseen in adaptation data. To obtain desirable recognition results for various amounts of adaptation data, we have to dynamically change the number of regression class to fit different occurrence rates of HMM units in adaptation data. The construction of a tree structure that records various degree of tying for HMM pdfs can provide an effective approach to this problem [1][10]. Using this approach, we could collect the adaptation tokens and transcription labels for each tree node to determine its associated transformation (or regression) parameters. The so-called hierarchical transformation is then performed using the parameters selected from the tree structure. However, in case of unsupervised adaptation, the transcription errors will cause the uncertainty of transformation parameters in each tree node. An autonomous model complexity control (AMCC) method [10] used minimum description length (MDL) [8] as a criterion for searching the transformation parameters in the tree structure. This criterion is feasible for reliability assessment in unsupervised adaptation. In this paper, we propose a novel selective unsupervised adaptation algorithm, where the extraction of reliable parameters is based on the confidence measure of adaptation tokens associated with the decoded word, state and mixture component transcriptions. From the unsupervised speaker adaptation experiments, we find that using the parameters extracted via the assessments of description length and confidence measure outperforms that captured in a fixed tree layer. The performance can be further improved by performing two-pass sequential adaptation according to confidence measure and description length.

2. UNSUPERVISED SPEAKER ADAPTATION

In continuous density HMM framework, we are given a set of parameters \( \lambda = \{ \lambda_\alpha \} = \{ \omega_\alpha, \mu_\alpha, r_\alpha \} \), where \( \omega_\alpha \), \( \mu_\alpha \) and \( r_\alpha \) are the mixture weight, mean vector, and precision matrix of the \( k \)th mixture component from the \( i \)th state. Let the HMM parameters be grouped into \( C \) clusters. The problem of unsupervised speaker adaptation is to transform the existing HMM parameters \( \lambda \) to a new speaker through some cluster-dependent functions \( G_n(\cdot) \), \( \eta = \{ \eta_j \} \), providing that the word sequence (or transcription) \( W = \{ w_t \} \) of adaptation data
\( X = \{ x_i \} \) is unknown. Consequently, there are two sets of free parameters, i.e., transformation parameters, \( \eta - \{ \eta_i \} \) and word sequence \( W \). In theory, two sets of parameters \( (\eta, W) \) can be jointly estimated by applying the maximum a posteriori \( [3] \) principle, which is expressed by

\[
(\hat{\eta}, \hat{W}) = \arg \max_{\eta, W} p(\eta, W | X) = \arg \max_{\eta, W} p(X | \eta, W) g(\eta, W) \tag{1}
\]

Assuming the parameters \( \eta \) and \( W \) are independent, we may further divide the estimation into two stages and utilize the expectation-maximization (EM) algorithm \([2]\) to solve the problem. Namely, the first stage is to estimate the most likely word sequence \( \hat{W} \) by performing the following steps

\[
\hat{W} = \arg \max_{\hat{W}} E[\log p(X, S, L | \hat{W}) + \log g(\hat{W}) | X, \eta, \hat{W}] \tag{2}
\]

where \( S = \{ s_i \} \) is the state sequence, \( L = \{ l_i \} \) is the mixture component sequence and \( g(\hat{W}) \) corresponds to the language model of word sequence. Given the new word sequence \( \hat{W} \), the new transformation parameters \( \hat{\eta} \) are estimated by

\[
\hat{\eta} = \arg \max_{\eta} E[\log p(X, S, L | \eta) + \log g(\eta) | X, \hat{W}] \tag{3}
\]

After several iterations, we can find the optimal transformation parameters for unsupervised adaptation. Nevertheless, the N-best word sequence hypotheses is also feasible to determine the transformation parameters \( \hat{\eta} \) [6-7]. In this study, we only employ the most likely word sequence \( \hat{W} \). The HMM parameters are transformed by using the parameters \( \hat{\eta} = \{ \hat{\mu}_k, \hat{\theta}_k \} \) defined by

\[
\hat{\eta} = G_{\eta}(\lambda) = \{ \lambda_{\hat{\mu}_k}, \lambda_{\hat{\theta}_k} \} \tag{4}
\]

Herein, \( \lambda_{\hat{\mu}_k} \) is labeled by the \( k \)-th cluster membership \( \Omega_k \). Under the constraints of Gaussian prior \( g(\hat{\mu}_k) = N(\hat{\mu}_k | m, \tau) \) and non-informative prior \( g(\hat{\theta}_k) = \text{constant} \), the transformation parameters are derived by

\[
\hat{\mu}_k = \left( \sum_{i, k \in \Omega_k} \xi_{\alpha}^{(i,k)} (i, k) \hat{\theta}_k r_{ik} + \tau \right)^{-1} \left( \sum_{i, k \in \Omega_k} \xi_{\alpha}^{(i,k)} (i, k) \hat{\theta}_k r_{ik} (x_i - \mu_k) + \tau m_i \right), \tag{5}
\]

\[
\hat{\theta}_k = \left( \sum_{i, k \in \Omega_k} \xi_{\alpha}^{(i,k)} (i, k) \right)^{-1} \left( \sum_{i, k \in \Omega_k} \xi_{\alpha}^{(i,k)} (i, k) (x_i - \mu_k - \hat{\mu}_k) (x_i - \mu_k - \hat{\mu}_k)^T \right), \tag{6}
\]

where \( \xi_{\alpha}^{(i,k)} (i, k) = \text{Pr}(s_i = i, l_i = k | X, \eta, \hat{W}) \).

### 3. EXTRATION OF TRANSFORMATION PARAMETERS

As mentioned above, it is efficient to build a tree structure of HMM’s to dynamically vary the degree of tying for transformation-based adaptation. Usually, we build the tree by clustering the HMM pdfs using the K-means algorithm. In the built tree, the root node contains all HMM pdfs and leaf nodes are occupied by individual HMM pdfs. Having the tree, we obtain the node labels of HMM pdfs in each layer. Generally, the HMM pdfs connected to the same node possess similar acoustical behaviors and can be suitably transformed via the shared transformation parameters. Using the unsupervised adaptation technique addressed in section 2, we first estimate the transcriptions of word, state and mixture component sequences from adaptation data. Then, the adaptation tokens and their associated HMM pdfs are assigned to different tree nodes. The transformation parameters embedded in each tree node are accordingly calculated by using (5)/(6). To reinforce the transformation discriminability, the HMM parameters should be transformed by using the parameters nearest to the leaf layer [1]. Thus, we may automatically search the transformation factors for each HMM pdf layer by layer along its associated tree path based on a bottom-up strategy [1].

#### 3.1 Minimum Description Length (MDL)

However, in unsupervised learning, the transcription errors will cause unreliability estimation of the transformation parameters. It is crucial to assess the reliability of transformation parameters for unsupervised adaptation. In the literature, the Bayesian information criterion (BIC) [9] and minimum description length (MDL) [8] were exploited to select or identify the statistical models and determine the number of parameters at the same time. They are very similar and permit to choose the parameters alleviating the problems of overtraining and unreliability. In the application of supervised speaker adaptation [10], the MDL principle was used to determine a tree cut in the tree structure and extract the transformation parameters by minimizing the description length written by

\[
\sum_{i=1}^{C} n_i \log \sigma^2_{\alpha, \alpha} + 2 \cdot C \cdot D \cdot \log n, \tag{7}
\]

where \( n_i \) is the number of adaptation tokens in the \( i \)-th cluster, \( D \) is the dimension of observation vector, \( n \) is total number of adaptation data and \( \sigma^2_{\alpha, \alpha} \) is the variance of difference vectors between adaptation tokens and their associated mean vectors, \( \{ x_i - \mu_k \} \), in the \( i \)-th cluster and \( \alpha \)-th dimension. This algorithm is herein applied to unsupervised adaptation.

#### 3.2 Maximum Confidence Measure (MCM)

Besides, this paper presents a verification scheme for evaluating the transformation reliability. Using this scheme, the transcriptions of adaptation tokens in each tree node are verified via hypothesis testing formula [11]. Let’s assume that the \( i \)-th tree node contains \( n_i \) i.i.d. adaptation tokens \( x' = \{ x'_1, x'_2, \ldots, x'_n \} \) and their corresponding HMM parameters \( \hat{x}' = \{ \hat{x}'_1, \hat{x}'_2, \ldots, \hat{x}'_n \} \). Using hypothesis testing for verifying transformation reliability, the likelihood ratio test is designed to evaluate whether or not the adaptation tokens \( x' \) are correctly transcribed by a set of HMM parameters \( \hat{x}' \). Hence, our
criterion is to test the null hypothesis, $H_0$, that adaptation tokens $\chi'$ are transcribed by target HMM parameters $\tilde{\lambda}$, against the alternative hypothesis, $H_1$, that $\chi'$ are transcribed by alternative (or anti) HMM parameters $\tilde{\lambda'}$, i.e. executing the following likelihood ratio test

$$LR(\chi',\tilde{\lambda'};\eta) = \frac{p(\chi' | H_0)}{p(\chi' | H_1)} = \frac{\prod_{i=1}^n p(x'_i | \tilde{\lambda}, \eta_i)}{\prod_{i=1}^n p(x'_i | \tilde{\lambda}', \eta_i)} .$$  \tag{8}

where $p(x'_i | \tilde{\lambda}, \eta_i)$ and $p(x'_i | \tilde{\lambda}', \eta_i)$ are the likelihoods of the null and alternative hypotheses given current estimate $\eta_i$, respectively. According to a decision threshold $\tau$, the transcriptions of adaptation tokens are accepted if $LR(\chi',\tilde{\lambda'};\eta) \geq \tau$ and rejected if $LR(\chi',\tilde{\lambda'};\eta) < \tau$. Generally, the likelihood ratio $LR(\chi',\tilde{\lambda'};\eta)$ measures how confident the adaptation tokens are transcribed in tree node $c$. Therefore, we are motivated to apply the confidence measure $LR(\chi',\tilde{\lambda'};\eta)$ to extract the reliable parameters in a tree structure. Our scheme is designed to capture the transformation parameters for HMM pdf $\tilde{\lambda}_d$ along its associated tree path $P_d$ as follows

$$\hat{c} = \arg\max_{c \in P_d} \frac{1}{n_c} \log LR(\chi',\tilde{\lambda'};\eta) = \arg\max_{c \in P_d} \frac{1}{n_c} \sum_{i=1}^n \left( \log p(x'_i | \tilde{\lambda}, \eta_i) - \log p(x'_i | \tilde{\lambda}', \eta_i) \right) .$$  \tag{9}

Namely, the new transformation parameters $(\hat{\mu}_d, \hat{\Sigma}_d)$ embedded in tree node $\hat{c}$ with the highest log likelihood ratio rate are extracted for adapting the HMM pdf $\tilde{\lambda}_d$. This algorithm is herein referred as maximum confidence measure. Notably, the confidence measure criterion and the model adaptation are sequentially performed in EM iterations. The calculation of likelihood ratio test is always done by using the newest adapted HMM parameters. In this study, the likelihood of alternative hypothesis is determined by a geometric mean term [11]

$$p(x'_i | \tilde{\lambda}', \eta_i) = \frac{1}{M} \sum_{m=1}^M p(x'_i | \tilde{\lambda}^{(m)}, \eta_i) .$$  \tag{10}

where $\{\tilde{\lambda}^{(m)}, m=1, \ldots, M \}$ denotes the cohort (or competing) HMM pdfs for $\tilde{\lambda}'$. It is easy to implement equation (10) because we don’t need to retrain the training utterances to obtain the anti model $\tilde{\lambda}'$.

### 3.3 Hybrid MCM-MDL

The MCM criterion is aimed at extracting the transformation parameters for each HMM pdf along the associated tree path. The parameters with the highest log likelihood ratio rate are extracted for model adaptation. The MDL decides a tree cut and extracts the parameters with the minimum description length. Assuming the number of tree node $C$ in (7) is fixed, MDL corresponds to the lowest accumulated variance of difference vectors $\{x_i - \mu_d\}$. In general, the MCM and MDL criteria originate from different viewpoints and extract different sets of parameters. To further elevate the performance, we combine the measures of confidence score and description length into the unsupervised learning of HMM’s. Our method is to perform a two-pass adaptation in each EM iteration by sequentially performing the selective unsupervised adaptation using MCM criterion followed by that using MDL criterion. The reliability of adapted HMM’s could be doubly verified.

### 4. EXPERIMENTS

The unsupervised speaker adaptation experiments conducted in this paper are aimed at the recognition of 408 Mandarin syllables. The detail description of experimental setup was mentioned in [1]. Two speech databases were collected. The first one consisted of 5045 phonetically-balanced Mandarin words uttered by 51 males and 50 females. This database contained all acoustics of 408 Mandarin syllables. We applied this database to generate the SI HMM’s and resultingly build a HMM tree structure [1]. There were eight layers in the tree. The second database consisted of four repetitions of 408 isolated Mandarin syllables spoken by a single female speaker who was excluded from the first database. We used three repetitions for testing and the remaining one for adaptation. The contents of adaptation data were unknown beforehand. Two databases were severely mismatched because of different recording rooms, microphones and speakers. Without adaptation, the baseline result using SI speech models had a top five recognition rate of 73.8%. The number of adaptation data of N=0, 25, 50, 75, 100, 200, 250, 300, 350 and 408 were included for assessing the effects of various adaptation data lengths in unsupervised adaptation. In this study, the tree nodes with at least five adaptation tokens are sufficient for extracting the transformation parameters. Due to possible insufficient data, we only performed the adaptation of mean vectors. Covariance matrices were unchanged. Besides, we construct the cohort membership of each HMM pdf according to the divergence distance measure between HMM pdfs. Because the number of cohort components did not influence our experiments too much, we simply set ten cohort members for each HMM pdf. Furthermore, in the experiments we ignored the contribution of prior density of language models and transformation parameters $g(\eta, W)$.

Two sets of experiments were carried out. First of all, we investigated the effects of hierarchical transformation with different tree depths in unsupervised adaptation. The recognition comparison is given in Figure 1. Herein, the extraction of transformation parameters is simply based on a bottom-up search strategy [1]. In case of four tree layers, the transformation parameters are calculated for four layers at most. From the figure, we find that a tree structure with small amount of layers (e.g. depth=4) provides good results for smaller N. But, when larger N is involved, more layers are needed to improve the recognition results. This phenomenon is especially obvious in unsupervised adaptation with bad transcriptions of adaptation data. Therefore, how to compromise the tradeoff between tree layer and amount of adaptation data becomes important in unsupervised adaptation.
In this study, we present several approaches to automatically capture the reliable transformation parameters without presetting them in a fixed layer. As soon as the observations allocated in a tree node meet the specified criterion, the corresponding transformation parameters are therefore used for unsupervised adaptation. The recognition results using bottom-up search method with eight tree layers, MDL, MCM and MCM-MDL are illustrated in Figure 2. The averaged transcription rate of adaptation data is about 40%. All experiments are done while the parameters are extracted from an eight-layer tree structure. We can see that the unsupervised adaptation with reliability assessment (i.e. MDL and MCM algorithms) is always superior to that without assessment (i.e. bottom-up algorithm) no matter how much adaptation data is available. Furthermore, the MCM algorithm outperforms the MDL algorithm. If we perform two-pass adaptation (i.e. MCM-MDL algorithm), the recognition performance can be significantly raised. In case of N=50, the bottom-up algorithm obtains a top five recognition rate of 78%, while the MDL and MCM algorithms can reach, respectively, recognition rates of 81% and 81.4%. Using hybrid MCM-MDL, the recognition rate can be further increased to 82.5%.

5. CONCLUSION

This paper explored the problem of adaptation when the speech recognizer has no prior knowledge of adaptation data. We explained the theoretical principle of unsupervised adaptation as a Bayesian estimation problem of adaptation parameters and word sequence. We also presented several methods for evaluating the transformation reliability in unsupervised adaptation. Compared to a traditional method, which does not consider how reliable the transcriptions of the adaptation data are, our proposed methods provide significant improvements for various amounts of adaptation data.

6. ACKNOWLEDGMENT

The first author acknowledges the support of Drs. Brian A. Hanson and Hisashi J. Wakita during his visit to Speech Technology Laboratory, Panasonic Technologies Inc.

7. REFERENCES