PRIOR PARAMETER TRANSFORMATION FOR UNSUPERVISED SPEAKER ADAPTATION

Guoqiang Li, Limin Du and Ziqiang Hou

Institute of Acoustics, Chinese Academy of Sciences
17 zhongguancun Rd., Beijing 100080, China
ligq@eudoramail.com

ABSTRACT

In a strictly Bayesian approach, prior parameters are assumed known, based on common or subjective knowledge. But a practical solution for maximum a posteriori adaptation methods is to adopt an empirical Bayesian approach, where the prior parameters are estimated directly from training speech data itself. So there is a problem of mismatches between training and testing conditions in the use of prior parameters. We proposed a prior parameter transformation (PPT) adaptation approach that transforms the prior parameters to be more representative of the new speaker. In this paper we extend it to unsupervised mode. For easily confused speech units, different transformation matrices are applied to make them distinct. Initial experiments show that the PPT algorithm can get much improvement for a small amount of adaptation data even in the unsupervised mode.

1. INTRODUCTION

In real world applications of speech recognition techniques there are usually performance degrades due to mismatches between training and testing conditions. The mismatches may come from speaker, channel or environmental variability. It is impossible to collect a large amount of speech data to cover all these situations. So it is desirable to modify existing HMM’s after obtaining a small amount of speech data of a new speaker in a testing condition. Such adaptation techniques are important components in practical speech recognition applications.

Many adaptation techniques have been proposed recently. One kind is transformation-based methods, such as maximum likelihood linear regression (MLLR) [1]. It modifies existing HMM’s, usually speaker independent (SI) models, by performing linear transformations on the HMM mean vectors. It can influence unseen models by tying linear transformations across different models. It transforms each separate cluster of HMM mean vectors by a cluster specific transformation, which is estimated from the adaptation data in the cluster. Thus, all HMM’s can be modified at one time even though some models are not observed in the adaptation data. So its adaptation process is fast. But the transformation lacks of specificity and gives broader modification on HMM mean vectors.

Another is a kind of Bayesian methods [2], which combines adaptation data from a new speaker with the prior knowledge. The prior knowledge is embedded in prior densities of the HMM parameters. It provides an optimal structure to incorporate many sources of knowledge. It has a nice asymptotic property. But it can only modify the seen models in the adaptation data and then there must be enough examples in the adaptation data for each model before it can adapt all the models. So its adaptation process is slow. There also exists a serious likelihood imbalance problem when the set of adapted maximum a posteriori (MAP) speaker adaptive models and the set of unadapted ML SI models are mixed to evaluate the overall utterance likelihood in order to find the most likely sequence of words [3]. To obtain the efficiency of transformation-based adaptation methods and effectiveness of Bayesian methods some hybrid methods have been developed.

In [4], an adaptation scheme was proposed that retains the nice properties of Bayesian schemes for large amounts of adaptation data and has improved performance for small amounts of adaptation data. They achieved this by using their transformation-based adaptation as a pre-processing step to transform the SI models so that they better match the new speaker characteristics and improve the prior information in MAP estimation schemes. To combine the transformation and an approximate Bayesian method, they first transformed the SI counts using the transformation parameters estimated with the constrained ML method. The transformed counts were then combined with SD counts collected using the adaptation data. At last, the combined method was estimated from these counts. But the transformation parameters and HMM parameters are separately estimated. And it used an approximate MAP estimation scheme that linearly combines the SI and SD counts for each component density, where the weight is fixed and does not reflect the dynamic variation between SI and SD counts.

In [5], the HMM parameters are transformed by adding a stochastic bias of new estimate with Gaussian parameters to the current HMM parameters. Then a Bayesian estimation technique that incorporates prior knowledge into the simple transformation was applied for estimating the transformation bias parameters. In the proposed hybrid algorithm, two sets of parameters need to be estimated. One is the set of mixture Gaussian HMM parameters. The other is their corresponding set of bias transformation parameters. Given the adaptation data from a new speaker, the speaker-adaptive HMM parameters are generated by sequentially performing the transformation-based adaptation and MAP adaptation. The MAP estimate is then obtained by maximizing the posterior likelihood, which consists of a likelihood function and a prior density. But there are two drawbacks with it. First, the prior density of bias transformation parameters with much prior information is not easily chosen. So they are set to be simpler forms and then do not bring much prior information about the HMM parameter transformation. Second, it only introduces simple stochastic biases as transformation parameters. Such restricts the performance of the combined methods.

In addition, in a strictly Bayesian approach, prior parameters are assumed known, based on common or subjective knowledge about the stochastic process involved. But a practical solution is to adopt an empirical Bayes approach [6], where the prior parameters are estimated directly from training speech data. So there is still a problem of mismatches between training and testing conditions. So it is desirable to move the prior parameters estimated from speech data in the training condition...
to be more suitable for the new speaker in a testing condition. Although there are many ways to achieve this goal we proposed a prior parameter transformation (PPT) adaptation approach [7] to modify the prior parameters to be more representative of the new speaker in a testing condition. It can influence unseen models by tying prior parameter transformations across different models according to amount of adaptation data available. Based on the improved prior information better HMM parameters can be obtained even with small amount of adaptation data.

The paper is organized as the following. In section 2, we review the PPT adaptation approach. Then we will give some discussions in section 3, and present some comparison experiments in section 4 and conclusion in section 5.

2. THE PPT APPROACH

In this section we only consider a simple case that mean vectors of HMM output p.d.f. are only adapted and other HMM parameters are left the same as the SI models and then transformation matrices apply to prior mean vectors only. PPT takes some adaptation data from a new speaker and then estimates the transformation parameters and finally updates HMM mean parameters with the transformed prior mean parameters.

The joint p.d.f. of observations \( X = \{X_t\}_{t=1:L,T} \) is specified by the following equation:

\[
P(X \mid \lambda) = \sum_{n=1}^{N} \sum_{h=1}^{H} P(X, s, l \mid \lambda)
\]

where

\[
P(X, s, l \mid \lambda) = a_{n,h} \prod_{t=1}^{T} a_{s_i,s_t} \omega_{s_t, r_{s_t}} N(x_t \mid m_{s_t, r_{s_t}})
\]

where \( \lambda \) is a set of unknown HMM parameters and here only includes HMM mean vectors. \( s \) and \( l \) are hidden state sequence and mixture sequence of HMM. \( \omega_{nk} \) denotes the mixture weight for the \( k \)-th mixture component of \( n \)-th state subject to the constrain \( \sum_{k=1}^{K} \omega_{nk} = 1 \). And \( N(x \mid m_{nk}, r_{nk}) \) is the normal density function denoted by:

\[
N(x \mid m_{nk}, r_{nk}) \propto r_{nk}^{1/2} \exp[-\frac{1}{2}(x-m_{nk})^T r_{nk}^{-1} (x-m_{nk})]
\]

where \( m_{nk} \) is a \( D \) dimensional mean vector to be estimated and \( r_{nk} \) is a \( D \times D \) known precision matrix. Since we adapt only HMM mean vectors, the Gaussian mixture parameter vector to be estimated is

\[
\lambda' = \{m_{nk} \}_{n=1:1,N,L=1:1,K}.
\]

The conjugate prior density for the vector parameter \( m_{nk} \) is a normal density of the form[8]:

\[
g(m_{nk} \mid \mu_{nk}) = N(m_{nk} \mid \mu_{nk}, \tau_{nk})
\]

where \( \mu_{nk} \) is a prior mean vector of \( D \) dimension, which is to be estimated and transformed, and \( \tau_{nk} \) is a \( D \times D \) known precision matrix. Here we only consider a linear transformation of the following form on an old prior mean vector \( \mu_{nk} \):

\[
\hat{\mu}_{nk} = W_{nk} \xi_{nk}
\]

where \( \xi_{nk} = \left( \begin{array}{c} 1 \\ \mu_{nk} \end{array} \right) \) is an extended mean vector and \( \hat{\mu}_{nk} \) is the transformed prior mean vector and \( W_{nk} \) is \( D \times (D+1) \) transformation matrix to be estimated. Then PPT algorithm calculates the following estimate of the transformation matrix and HMM mean vectors by maximizing the posterior likelihood of the adapted models generating the adaptation data and iteratively employing expectation maximization (EM) algorithm[9].

\[
\Delta_{nk} \hat{W}_{nk} \xi_{nk} \xi'^{nk} = \Delta_{nk} \hat{\tau}_{nk} \xi_{nk}
\]

\[
\hat{m}_{nk} = (\tau_{nk} + c_{nk} r_{nk})^{-1}(\tau_{nk} \hat{W}_{nk} \xi_{nk} + c_{nk} r_{nk} \xi')
\]

where

\[
c_{nk}(t) = \omega_{nk} N(x_t \mid m_{nk}, r_{nk}) / \sum_{k} \omega_{nk} N(x_t \mid m_{nk}, r_{nk})
\]

\[
C_{nk} = \sum_{t=1}^{T} c_{nk}(t)
\]

\[
\bar{x}_{nk} = \sum_{t=1}^{T} c_{nk}(t) x_t
\]

\[
\Delta_{nk} = \tau_{nk} C_{nk} r_{nk} (\tau_{nk} + c_{nk} r_{nk})^{-1}
\]

When each mixture component has an individual transformation matrix the above formula can be rewritten as the following:

\[
\hat{\mu}_{nk} = \hat{W}_{nk} \xi_{nk} = \bar{x}_{nk}
\]

\[
\hat{m}_{nk} = \bar{x}_{nk}.
\]

We can see that prior mean vectors are transformed to be specific for the test speaker and then the PPT algorithm is equivalent to standard maximum likelihood re-estimation formulae for the HMM mean vectors.

A better approach is to pool the information from a number of distributions into a single matrix, which is then used to transform all the prior mean vectors of the contributing distributions. The tying of degree is determined by the amount of adaptation data available. The case of small amounts of training data is of special interest, and this obviously needs a large degree of tying.

To generate the tied transformation matrices the summation should be performed over all tied distributions. If a transformation matrix \( W \) is shared by a class of mixtures \( \{s, l\}_{s=1:L, r} \), where \( s, l \) represents \( l \)-th mixture of \( s \)-th state in the tying class, the above equation becomes:

\[
\hat{W}_{nk} = \sum_{s=1}^{S} \sum_{l=1}^{L} \omega_{nk} N(x_t \mid m_{nk}, r_{nk}) / \sum_{k} \omega_{nk} N(x_t \mid m_{nk}, r_{nk})
\]

\[
\hat{m}_{nk} = \sum_{s=1}^{S} \sum_{l=1}^{L} \omega_{nk} x_t / \sum_{s=1}^{S} \sum_{l=1}^{L} \omega_{nk}
\]
\[
\sum_{r=1}^{R} \Delta_{x_r} \hat{w}\xi_{x_r} \xi'_{x_r} = \sum_{r=1}^{R} \Delta_{x_r} \hat{w_{x_r}} \xi'_{x_r} .
\] (8)

The above equation can be solved using Gaussian elimination or LU decomposition methods to calculate the transformation matrix on a row-by-row basis.

3. DISCUSSION

In practical application unsupervised speaker adaptation is desirable. In order to improve its performance in this case we take two passes. In the first pass adaptation data are recognized by SI models. For easily confused speech units the second pass is taken to pick up correct speech units. Different transformation matrices are applied to their prior parameters to make these confused speech units more distinct and their resulting likelihood functions are compared. It is important to choose proper transformation matrices for the easily confused speech units only.

3.1 Methods to improve prior information

There are some other ways to improve prior information for MAP estimation. First the use of the speaker clustering can surely improve the prior information and then the performance of the PPT algorithm as a pre-processing.

Second method is to perform an iteration of MLLR on adaptation data to obtain estimates for all HMM mean vectors, which is set as prior mean parameters for next MAP. Then perform an iteration of MAP on the improved HMM parameters. But two steps of estimation are separately performed. In addition this kind adaptation method is the same as MLLR for first few adaptation data since MAP cannot update any model parameters when the amount of adaptation data is not enough. So this adaptation strategy is not suitable for fast adaptation.

3.2 Influence of prior precision matrices

In the equation (5), if the prior precision is large enough, that is \( \tau_{nk} \gg c_{nk} r_{nk} \), then \( \Delta_{nk} \approx c_{nk} r_{nk} \) while if the prior precision is small enough, that is \( \tau_{nk} \ll c_{nk} r_{nk} \), then \( \Delta_{nk} \approx \tau_{nk} \). If the prior precision of some distribution is large enough, we can get the prior information that the mean vector of the corresponding HMM distribution varies very small across different speakers. At this time PPT algorithm accumulates the contribution from the distribution when calculating the transformation matrix in equation (8), which is the same as MLLR. In the other hand, if the prior precision of some distribution is small enough, the prior information is that the mean vector of the corresponding HMM distribution varies very large across different speakers. At this time PPT algorithm lowers the contribution from the distribution, from \( c_{nk} r_{nk} \) to a smaller value \( \tau_{nk} \), when accumulating the statistics in the equation (8). When accumulating the statistics to estimate transformation matrices and then prior parameters more suitable for the test speaker, the PPT algorithm collects as much the statistics as possible from robust distribution across speakers. Thus the resulting transformation matrices can be more robust.

In the equation (4), if \( \tau_{nk} \) is large enough, the resulting estimates of HMM mean vectors can depend on transformed prior mean vectors. Otherwise, they can be mainly calculated from statistics of speech data for the test speaker.

![Figure 1. Comparison of PPT and MLLR. Test file contains 400 sentences. SI's performance is 86.72%. For PPT and MLLR only one transformation is used. For PPT prior variance of HMM mean vectors is set fixed 10.](image-url)

4. EXPERIMENTAL RESULTS

In the experiments all adaptations are performed in a static supervised manner using labelled adaptation data. We use a continuous speech database to evaluate the new algorithm. Speech is parameterised by using 12 MFCCs plus log energy and their first and second time derivatives.

A set of SI models is trained on the speech from 123 speakers, female 57, male 66, each speaking 500 to 600 sentences of continuous speech. The models are state clustered cross-word triphones, containing a total of 8975 states. Each state has six mixture components, which are modelled by a normal distribution with diagonal covariance matrix. The basic phone set consists of 46 phone symbols plus a silence. Bigram language is used.

We perform three sets of experiments. In the first set of experiment, we select 50 sentences from speech data of 8 testing speakers as adaptation data and select 400 sentences as test data to obtain more accurate testing results. The SI system gives 86.72% word correct rate. But since the speech data for each testing speaker is limited to be about 500—600 sentences other speech data for the testing speaker cannot be obtained to train each SD system. For MLLR and PPT, only one transformation is used and for PPT \( \tau_{nk} \) in the equation is set fixed and a diagonal matrix with each diagonal element being 10. From the Figure 1, we can see that PPT is better than MLLR for different amount of adaptation data and it is interesting that PPT is much better than MLLR when number of adaptation data is less than 7. When there is only one sentences of adaptation data, both is much worse than SI 68.55% for PPT and 69.17% for MLLR.

In the second experiment, many sets of SD models are trained for each speaker with about 500 sentences as training data to estimate prior parameters. Another 8 speakers are selected for adaptation tests. For each testing speaker about 500 sentences are selected as training data to get SD performance, 50
sentences as testing data and 100 sentences from the training data as adaptation data.

From Figure 2 we can see that when adaptation data is small PPT is better than MLLR+MAP, i.e. performing alternately MLLR and MAP and when adaptation data is large both is similar.

In the third experiment, we perform unsupervised adaptation of PPT and MLLR. Selection of speech data is same as the second experiment. In Figure 3 when adaptation data is small PPT performance of unsupervised adaptation is better than MLLR.

5. CONCLUSION

In this paper we make some comparisons between PPT and MLLR. The PPT adaptation method transforms prior parameters to be more representative of a new speaker. Based on the improved prior information better model parameters can be obtained. Experiments show that PPT is better than MLLR, especially when adaptation data is small. In addition it can influence unseen models by employing prior parameter transformations, which are tied across different models according to the amount of available adaptation data. So the PPT adaptation approach is more effective with small amount of adaptation data while with a large amount of adaptation data it is equivalent to usual re-estimation algorithm.

It jointly estimates transformation matrices and HMM parameters to be more representative of a new speaker. Based on the improved prior information better model parameters can be obtained. Experiments show that PPT is better than MLLR, especially when adaptation data is small. In addition it can influence unseen models by employing prior parameter transformations, which are tied across different models according to the amount of available adaptation data. So the PPT adaptation approach is more effective with small amount of adaptation data while with a large amount of adaptation data it is equivalent to usual re-estimation algorithm.

PPT's aptitude for very fast adaptation makes it suitable for practical application. In the unsupervised speaker adaptation, it chooses different transformation matrices for easily confused speech units to make them more distinct.

6. REFERENCES