ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

REGRESSION TRANSFORMATION OF PRIOR MEANS FOR SPEAKER ADAPTATION

Guoqiang Li, Limin Du and Ziqiang Hou

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

ABSTRACT

Maximum a posteriori adaptation method combines the prior knowledge with adaptation data from a new speaker, which has a nice asymptotical property, but has a slow adaptation rate for not modifying unseen models. In a strictly Bayesian approach, prior parameters are assumed known, based on common or subjective knowledge. But a practical solution 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 this paper we propose a prior parameter transformation (PPT) adaptation approach that transforms the prior parameters to be more representative of the new speaker. 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 model parameters can be obtained even with small amount of adaptation data.

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 is transformation-based method [1]. It modifies existing HMM's, usually speaker independent (SI) models, by performing transformations on the mean vectors of models. It can influence unseen models by tying transformations across different models. It transforms different clusters of speech HMM's 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.

Another is a kind of Bayesian methods [2], which combines adaptation data from a new speaker with the prior knowledge. The prior knowledge consists of prior densities of the HMM parameters. It provides an optimal structure to incorporate many sources of knowledge. It has a nice asymptotical property. But it can only modify the seen models 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. To obtain the efficiency of transformation-based adaptation methods and effectiveness of Bayesian methods some hybrid methods have been developed.

In [3], a Bayesian estimation technique that incorporates prior knowledge into the transformation was applied for estimating the transformation 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 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 maximum a posteriori (MAP) adaptation. The MAP estimate is then obtained by maximizing the posterior likelihood, which consists of a likelihood function and a prior density. There are two drawbacks with it. First, the prior density of 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 stochastic biases as transformation parameters. Such restricts the performance of the combined methods.

In another hybrid method [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 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. And transformation parameters and HMM parameters are separately estimated.

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 [7], 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 we propose a prior parameter transformation (PPT) adaptation approach to modify the prior parameters to be more representative of the new speaker in a new condition.

The paper is organized as the following. In section 2, we introduce the general case of our new approach. Then in section 3, we focus on a simple case and give a formula. In section 4 some discussion will be given. In section 5, we will present some initial experimental results and in section 6 give conclusion.

2. A GENERAL PPT APPROACH

The PPT approach to speaker adaptation requires an initial SI continuous HMM system and a set of prior parameters. As many researchers do a set of prior parameters may be chosen as the mean vectors of a SI system. Here, the PPT approach for a general condition is given that a set of prior parameters is not limited to the SI HMM mean vectors.

If θ is a HMM parameter vector to be estimated from a sequence of observations, $X = \{x_1, \cdots, x_T\}$ with probability density function (p.d.f.) $p(X|\theta)$ and its prior density is $g(\theta|\varphi)$, where φ is a prior parameter vector, then the MAP estimate can be obtained by:

$$\theta_{MAP} = \arg\max_{\theta} p(X|\theta) g(\theta|\varphi). \quad (1)$$

But there is no sufficient statistic of a fixed dimension for the parameter vector θ because of the underlying hidden process [2], i.e. the state mixture component and the state sequence of a Markov chain for an HMM. Therefore, no joint conjugate prior density can be specified. However, $p(X|\theta)$ can be seen as a margin p.d.f. of the joint p.d.f., $p(Y|\theta)$, of the parameter θ expressed as the product of a multinomial density and multivariate Gaussian densities, where Y is the complete data including hidden state and mixture component sequence and observations X. Then a practical candidate to model the prior knowledge about the mixture weight parameter vector is a Dirichlet density while the joint conjugate prior density for each Gaussian mixture component is normal-Whishart density. After this choice for the prior density family the expectation maximum (EM) algorithm [5] can be applied to MAP estimation problem if the prior density belongs to the conjugate family of the complete-data density.

In a strictly Bayesian approach, the parameter φ of this family of p.d.f.'s $G(\cdot|\varphi)$ is also assumed known, based on common or subjective knowledge about the stochastic process involved. However, in a real situation it is usually difficult to have such subjective knowledge, especially in such cases where the parameters are continuous and multi-dimensional [6].

An alternate solution to a strictly Bayesian approach is to adopt an empirical Bayesian approach [6], where the prior parameters are estimated directly from data itself. The estimation is based on the marginal distribution of the data given the estimated prior parameters. To estimate the parameters of the marginal p.d.f., one simple method is a standard method of moments, which equates the first few samples moments with the corresponding population moments. The initial set of model parameters for prior parameter estimation is derived from either the parameters of a currently available SI system or by polling from several speaker dependent (SD) trained systems.

There is still a problem of mismatches between training and testing conditions for empirical Bayesian methods to estimate the prior parameters. So under certain circumstances, some better-estimated prior parameters can lead to better HMM model parameter estimations. In the PPT approach, the prior parameters are transformed by the following transformation to be more suitable for the new speaker:

$$\pi = f_\eta(\varphi) \quad (2)$$

where $f_\eta(\cdot)$ is a transformation function with a set of transformation parameters η and π is the transformed prior parameter. The transformation functions are tied across different models and the degree of tying is adjusted by the amount of adaptation data available. Then the HMM parameters θ and transformation parameters of prior parameters η can theoretically jointly estimated using the MAP framework:

$$(\theta, \eta) = \arg\max_{\theta,\eta} p(X|\theta) g(\theta | f_\eta(\varphi)). \quad (3)$$

Usually the MAP estimation is not easily solved for incomplete data such as HMM. Thus, we employ the EM algorithm to iteratively increase the posterior likelihood $p(\theta, \eta | X)$ of the current estimates (θ', η') and obtain the new estimates (θ, η) in an optimal manner. We perform first step of EM algorithm (E-step) by computing the following auxiliary function:

$$Q(\theta, \eta | \theta', \eta') = E\{\log p(Y|\theta, \eta) + \log g(\theta | f_\eta(\varphi)) | X, \theta', \eta'\} \quad (4)$$

where Y is the complete data including hidden state and mixture component sequence and observations X. Then, in the second step (M-step), we find new estimates (θ, η) by maximizing $Q(\theta, \eta | \theta', \eta')$ over θ and η. It can be shown that if

$$Q(\theta, \eta | \theta', \eta') \geq Q(\theta', \eta' | \theta', \eta') \quad (5)$$

then

$$p(\theta, \eta | X) \geq p(\theta', \eta' | X). \quad (6)$$

By iteratively applying the EM algorithm it can be guaranteed that the posterior likelihood does not decrease.

3. PPT APPROACH FOR MEANS ONLY

In this section we only consider a simple case that mean vectors of HMM system are only adapted and other HMM parameters are not adapted and then regression transformations only apply to prior mean vectors. The derivation of the PPT estimates is given below for Gaussian mixtures. PPT takes some adaptation data from a new speaker and updates the transformation parameters and HMM mean parameters.

The joint p.d.f. of observations X is specified by the following equation:

$$p(X \mid \theta) = \prod_{t=1}^{T} \sum_{k=1}^{K} \omega_k N(x_t \mid \mu_k, r_k) \tag{7}$$

where ω_k denotes the mixture weight for the *k*-th mixture component subject to the constrain $\sum_{k=1}^{K} \omega_k = 1$.

$N(x \mid m_k, r_k)$ is the *k*-th normal density function denoted by:

$$\begin{aligned} & N(x \mid m_k, r_k) \\ & \propto |r_k|^{1/2} \exp[-\tfrac{1}{2}(x - m_k)' r_k (x - m_k)] \end{aligned} \tag{8}$$

where m_k is the D dimensional mean vector to be estimated and r_k is $D \times D$ known precision matrix. Since we adapt only HMM mean vectors, the Gaussian mixture parameter vector to be estimated in the equation (7) is $\theta = (m_1, \cdots, m_K)$.

The conjugate prior density for the vector parameter m_k is a normal density of the form:

$$g(m_k \mid \hat{\mu}_k) = N(m_k \mid \hat{\mu}_k, \tau_k) \tag{9}$$

where $\hat{\mu}_k$ is a prior mean vector of D dimension, which is to be estimated and transformed, and τ_k is $D \times D$ known precision matrix. In this section we only consider linear regression transformation of the following form on an old prior mean vector μ_k:

$$\mu_k = W \xi_k \tag{10}$$

where $\xi_k = \begin{pmatrix} 1 \\ \hat{\mu}_k \end{pmatrix}$ is an extended mean vector and μ_k is the transformed prior mean vector and W is $D \times (D+1)$ transformation matrix to be estimated. Then

$$\begin{aligned} & E[\log p(Y \mid \theta) \mid X, \hat{\theta}] \\ & \propto -\tfrac{1}{2} \sum_{k=1}^{K} c_k (m_k - \bar{x}_k)' r_k (m_k - \bar{x}_k) \end{aligned} \tag{11}$$

where $c_{kt} = \dfrac{\hat{\omega}_k N(x_t \mid \hat{m}_k, \hat{r}_k)}{p(x_t \mid \hat{\theta})}$, $c_k = \sum_{t=1}^{T} c_{kt}$,

$\bar{x}_k = \sum_{t=1}^{T} c_{kt} x_t / c_k$.

And the joint log prior density of Gaussian mixture mean vectors is given:

$$\begin{aligned} & \sum_{k=1}^{K} \log g(m_k) \\ & \propto -\tfrac{1}{2} \sum_{k=1}^{K} (m_k - W\xi_k)' \tau_k (m_k - W\xi_k) \end{aligned} . \tag{12}$$

All mixture components of a state are transformed by a transformation function, i.e. they are tied to employ a common transformation.

Prior mean vectors of Gaussian mixture mean vectors are transformed to maximize the posterior density function of adaptation data that consists of transformed prior density and likelihood function. The transformation parameter and Gaussian mixture parameter vectors can be estimated by maximizing the following function:

$$\begin{aligned} & \log R(\theta, W) \\ & = E[\log h(Y \mid \theta) \mid X, \hat{\theta}] + \sum_{k=1}^{K} \log g(m_k) \end{aligned} \tag{13}$$

First, making the differentiation of the above function (13) with m_k and equating it to zero can find the estimate of the Gaussian mixture mean vector.

$$m_k = (\tau_k + c_k r_k)^{-1} (\tau_k W \xi_k + c_k r_k \bar{x}_k). \tag{14}$$

Then the transformation parameter can be estimated by substituting the above estimated means into the function (13) and maximizing (13) over W. If r_k and τ_k are all diagonal matrices then we can obtain the following formula for W:

$$\sum_{k=1}^{K} F_k W \xi_k \xi_k' = \sum_{k=1}^{K} F_k \bar{x}_k \xi_k' \tag{15}$$

where $F_k = c_k \tau_k r_k (\tau_k + c_k r_k)^{-1}$.

The above development for a mixture of multivariate Gaussian densities can be easily extended to the case of HMM with Gaussian mixture state observation densities. And its extension to the case of tying transformation across different states is also straightforward. This can be achieved by just summing formula (15) over all mixture components in the tied regression class in the formula.

4. DISCUSSION

4.1 Other ways to improve prior information

There are some other ways to improve prior information for MAP estimation.

First, speaker clustering can achieve this. In [8], speaker clustering was used to improve the initial model parameters for prior parameter estimation. The speakers of the SI speech database are divided into a few clusters according to the similarity of their SD model parameters. For each of these speaker clusters, a cluster specific

HMM system is trained using all the speech data from the speakers of the cluster. When adaptation data for a new speaker is available, these adaptation data are used to select the closest cluster specific HMM to the new speaker for the prior parameter estimation. But it needs great computation load and disk space.

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. Then perform an iteration of MAP on the improved HMM parameters. But it is limited to a special case that the prior mean parameters are assumed as the same as the mean vectors of HMM system. In addition two steps of estimation are separate.

4.2 Not tying transformation

When a large amount of adaptation data is available we can have a separate transform for each mixture component which leads to the following formula:

$$\hat{\mu}_k = W_k \xi_k = \bar{x}_k,\ \hat{m}_k = \bar{x}_k \qquad (16)$$

Thus, the prior mean information has been degenerated to be totally suitable for the new speaker. And it is equivalent to a complete re-estimation of the mean vectors. When there is a large amount of adaptation data for the new speaker a separate transform can be used for each mixture component. Then there is no use transforming prior parameters. So the PPT approach is more suitable for small amount of adaptation data.

5. 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 2576 states. Each state has 4 streams and each stream has only one mixture component, which is modelled by a normal distribution with diagonal covariance matrix. The basic phone set consists of 47 phone symbols plus a silence.

Many sets of SD models are trained for each speaker with 500 sentences as training data and the rest sentences as testing data. To evaluate the new algorithm 10 speakers are selected for adaptation tests.

The SI system gives 30.18% word error rate (WER) and the SD system gives 22.00% WER (averaged over all 10 testing speakers)

From Table 1 we can see that the PPT adaptation approach is effective in small amount of adaptation data. We use single mixture HMM's as SI baseline system here. It is believed that the performances of MAP and PPT will be increased when using multiple mixture Gaussian components HMM's. We will do this experiment in near future.

Table 1
Performance comparison between MAP and PPT (WER, %). SI (WER=30.18%), SD (WER=22%)

No. of sentences	10	40	200
MAP	30.91	40.44	24.73
PPT	29.27	28.73	25.46

6. CONCLUSION

In this paper we proposed a prior parameter transformation (PPT) adaptation method that transforms prior parameters of HMM parameters to be more representative of a new speaker. It jointly estimates transformation and HMM parameters. Based on the improved prior information better model parameters can be obtained. 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. Initial experiments show that it is effective. In the future work we will extend to transforming other prior parameters.

7. REFERENCES

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Comput. Speech Lang. , pp.171-185, 1995.

[2] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations for Markov chains," IEEE Trans. Speech Audio Procession, vol.2, pp.291-298, 1994.

[3] J. T. Chen, C. H. Lee and H. C. Wang, "A hybrid algorithm for speaker adaptation using MAP transformation and adaptation," IEEE Signal Processing Letter, vol.4, pp.167-169, June 1997.

[4] V. V. Digalakis and L. G. Neuneyer, "Speaker adaptation using combined transformation and Bayesian methods," IEEE Trans. Speech Audio Processing, vol.4, pp.294-300, July 1996.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood estimation from incomplete data," J. R. Stat. Soc. B, vol.39, pp.1-38, 1977.

[6] J. O. Berger, Statistical decision theory and Bayesian analysis. Springer-Verlag, New York, 2nd edition, 1985.

[7] H. Robins. The empirical Bayes approach to statistical decision problems. Annals Math. Stat., 35(1): 1-20, March 1964.

[8] S. M. Ahadi, "Bayesian and predictive techniques for speaker adaptation," Ph.D. thesis, 1996, Cambridge University Engineering Department.