CORRELATION BASED PREDICTIVE ADAPTATION OF HIDDEN MARKOV MODELS

Mohamed Afify[1] *Yifan Gong*[1,2] *Jean-Paul Haton*[1]

[1]CRIN/CNRS-INRIA-Lorraine,B.P. 239 54506 Vandeouvre,Nancy,France
[2] Media Technologies Laboratory, Texas Instruments, P.O.BOX 655303 MS 8374, Dallas TX 75265, U.S.A.

ABSTRACT

Hidden Markov model (HMM) adaptation is currently of interest, to overcome the degradation effect of speaker and/or channel mismatches in practical speech recognition systems. The Bayesian framework provides a theoretically optimal formulation for combining adaptation data and prior knowledge, but it suffers from the drawback of being incapable of adapting parameters of the models that have no observations in the adaptation speech. In this article we present a new predicitve (in the sense of influencing unobserved distribution parameters) adaptation algorithm for the mean vectors of an HMM. We also point out some theoretical relationships between the proposed method and other techniques used in the context of predictive model adaptation. The efficacy of the proposed approach is demonstrated in speaker adaptation experiemnts for both an isolated word task, and TIMIT phonetic recogntion.

1. INTRODUCTION

The performance of hidden Markov model based speech recognizers may degrade significantly when a mismatch occurs between their training and testing environments. This mismatch is often encountered in practice, and can be due to speaker and/or channel difference. It has been shown that adapting the model parameters using some data observed in the test environment, is an effective way to overcome this degradation. The Bayesian framework provides a theoretically optimal formalism for the combination of both adaptation data and prior knowledge, and hence it enables obtaining robust parameter estimates from limited adaptation data (e.g. [4]). However, a major drawback of this approach is its incapability of influencing parameters of the models that have no observations in the adaptation speech. Hence, for very short calibration speech, Bayesian techniques will only adapt a small fraction of the system parameters, and the need arises for predictive adaptation algorithms that can potentially estimate parameters not observed in the adaptation data.

Recently several predictive model adaptation techniques were proposed, where **predictive** refers to their ability of adapting parameters for which there exist no adaptation data. These techniques can be broadly classified as being Bayesian or transformation based. Bayesian techniques use extended maximum aposteriori (EMAP) estimation where a joint prior distribution is adopted in a MAP framework [6, 11]. Also some interesting approximations to the joint prior were recently proposed, using pairwise correlations [5], and using Markov random fields [8]. On the otherhand, tranformation based algorithms either use a set of linear transformations trained to predict target distributions from a set of source distributions during adapation [3, 1], or estimate unobserved distributions means by utilizing the transfer vectors of distributions in their neighbourhoods [9, 10].

Generally, the application of an exact EMAP algorithm to a system having even moderate number of parameters is prohibitive both from the computation and storage points of view. While transformation based methods are generally heuristic, and lack a well defined probabilistic criterion. [1] In this article we present the theoretical foundations as well as experimental evaluation of a new correlation based predictive adaptation technique for the mean vectors of a continuous density hidden Markov model. The basic idea is to predict the mean of an unobserved distribution as a combined estimate of a set of pairwise MMSE estimates using observed distribution means in its neighbourhood. We also point out some relationships between this approach and both the linear transformation, and the transfer vector field methods.

The article is organized as follows. Section 2 presents the basic principle of the technique, while Section 3 discusses its application to HMM mean adaptation. Relationships between the proposed method and both linear regression, and transfer vector field adaptation are outlined in Section 4. Experimental evaluation and conclusion are given in Sections 5 and 6 respectively.

2. BASIC PRINCIPLES

In this section we give the basic principles of the proposed method. The following discussion concerns the distribution level, where speech is assumed to be properly aligned to distributions using the Viterbi algorithm and a set of initial models (e.g., for speaker adaptation, speaker independent models). Also for simplicity we consider scalar observations, while the generalization to vector observations can be trivially obtained. The basic idea of the proposed method is to identify a neighbourhood (in correlation sense)with each distribution. Then during adaptation pairwise MMSE

[1] In fact, we show in Section 4 that transformation based methods can be obtained from the proposed approach by making some approximations.

estimates of the means of an unobserved distribution are obtained from those obsereved ones in its neighbourhood. Finally, a unique estimate of the mean is obtained by combining the pairwise estimates in a ML sense.

Assume that two distribution means μ_k and μ_l are jointly normal, and that their joint distribution is given by

$$p(\mu_k, \mu_l) \sim N(\mu_k^{SI}, \mu_l^{SI}, \sigma_{kk}^2, \sigma_{ll}^2, r_{kl}) \qquad (1)$$

where $\mu_{k(l)}^{SI}$ is the speaker independent mean, $\sigma_{kk(ll)}^2$ is the variance, and r_{kl} is the correlation coefficient.
The minimum mean square error (MMSE) estimate of μ_k given μ_l is given by [2]

$$\hat{\mu}_{k,l} \overset{\text{def}}{=} E[\mu_k|\mu_l] = \mu_k^{SI} + \frac{r_{kl}\sigma_{kk}}{\sigma_{ll}}(\mu_l - \mu_l^{SI}) \qquad (2)$$

and its associated variance is given by

$$\hat{\sigma}_{k,l}^2 = \sigma_{kk}^2(1 - r_{kl}^2) \qquad (3)$$

As we don't have perfect knowledge of μ_l, we use an estimate μ_l^*, a MAP estimate is used in this work, which is given by

$$\mu_l^* = \frac{N_l}{N_l + \tau}\bar{x}_l + \frac{\tau}{N_l + \tau}\mu_l^{SI} \qquad (4)$$

where $\bar{x}_l$ is the sample average of observations of l,N_l is the number of observations belonging to l, and τ is a parameter controlling the relative weight of the prior and the adaptation data.
In this case the estimate of μ_k given μ_l can be written as

$$\hat{\mu}_{k,l} = \mu_k^{SI} + \frac{r_{kl}\sigma_{kk}}{\sigma_{ll}}(\mu_l^* - \mu_l^{SI}) \qquad (5)$$

and it can be shown that its associated variance is given by

$$\hat{\sigma}_{k,l}^2 = \sigma_{kk}^2\left(1 - r_{kl}^2\left(1 - \frac{\eta_l}{(N_l + \tau)}\right)\right) \qquad (6)$$

where $\eta_l \overset{\text{def}}{=} \frac{\sigma_{x_l}^2}{\sigma_{ll}^2}$, and $\sigma_{x_l}^2$ is the sample variance of distribution l. In this work we take $\tau = \eta_l = 5.0 \;\; \forall l$.
To obtain a unique estimate of μ_k we combine the set of pairwise estimates in the neighbourhood ($\mathcal{N}(k)$) of k. In this work we use a maximum likelihood (ML) based combination given by

$$\hat{\mu}_k = \frac{\sum_{l=1}^{|\mathcal{N}(k)|}\hat{\mu}_{k,l}/\hat{\sigma}_{k,l}^2}{\sum_{l=1}^{|\mathcal{N}(k)|}1/\hat{\sigma}_{k,l}^2} \qquad (7)$$

Substituting (5) and (6) into (7), and after some simplification we get

$$\hat{\mu}_k = \mu_k^{SI} + \frac{\sum_{l=1}^{|\mathcal{N}(k)|}\frac{r_{kl}\sigma_{kk}}{\sigma_{ll}}(\mu_l^* - \mu_l^{SI})/(1 - r_{kl}^2(1 - \frac{\eta_l}{(N_l+\tau)}))}{\sum_{l=1}^{|\mathcal{N}(k)|}1/(1 - r_{kl}^2(1 - \frac{\eta_l}{(N_l+\tau)}))} \qquad (8)$$

3. APPLICATION TO MODEL ADAPTATION

To perform mean prediction of distribution k using (8), we need estimates of $\{\sigma_{ll}, r_{kl} \;\; l \in \mathcal{N}(k)\}$. These estimates can be obtained using the moment method from a training set consisting of N groups (e.g. speakers) as shown below.

$$\sigma_{ll}^2 = \frac{1}{N}\sum_{i=1}^{N}(\bar{x}_{l,i} - \mu_l^{SI})^2 \qquad (9)$$

$$r_{kl} = \frac{\sum_{i=1}^{N}(\bar{x}_{k,i} - \mu_k^{SI})(\bar{x}_{l,i} - \mu_l^{SI})}{\sqrt{\sum_{i=1}^{N}(\bar{x}_{l,i} - \mu_l^{SI})^2}\sqrt{\sum_{i=1}^{N}(\bar{x}_{k,i} - \mu_k^{SI})^2}} \qquad (10)$$

where $\bar{x}_{l,i}$ denotes the sample average of the l^{th} distribution in the i^{th} group.

The neighbourhood $\mathcal{N}(k)$ of distribution k is constructed from the mostly correlated distributions (i.e those having highest r_{kl} values). This neighbourhood construction procedure works for 1-dimensional observations, and a possible measure of distribution correlation for P-dimensional vectors can be calculated as:

$$\bar{r}_{kl} = \frac{1}{P}\sum_{j=1}^{P}|r_{kl}(j)| \qquad (11)$$

where each $r_{kl}(j)$ in (11) is calculated as in (10). Possibly phonetic knowledge, or any other constraints can be applied to reduce the computational complexity of the neighbourhood construction process. However, no such attempt was made in this paper.

3.1. Summary of the adaptation algorithm

Up to this point the discussion was carried out at the distribution level. The extension to HMMs is straightforward. The distributions of all mixture components of all states are considered as a large pool. During training (correlation structure and parameter estimation), and adaptation the speech is assigned to distributions using the the Viterbi algorithm and a set of initial models. Once the speech is assigned to appropriate distributions, the algorithm proceeds as described in the previous sections. A summary of the training and adaptation algortihms is given below.

1. Training Phase
   - Assign the training speech to distributions using the Viterbi algorithm, and a set of initial models.
   - Using the assigned speech , calculate the parameters σ_{kk}, and $\{r_{kl} \; \forall l\}$ for each distribution k (equations (9)-(10)).
   - For each distribution k construct the neighbourhood $\mathcal{N}(k)$ from the mostly correlated distributions (those having highest r_{kl} values).
2. Adaptation Phase
   - Assign the adaptation speech to distributions using the Viterbi algorithm and a set of initial models (possibly speaker independent models).

- For distributions having observations use MAP estimation (4) to calculate their means.
- Use (8) to predict the means of unobserved distributions from the means of observed distributions in their neighbourhoods.

4. RELATIONSHIPS WITH EXISTING APPROACHES

In this section we present some relationships between the proposed algorithm and both vector field correlation (VFC)[9], and linear regression prediction [3].

In (8) if we make the assumptions that distributions k and l are perfectly correlated (i.e. $r_{kl} \approx 1$ and $\sigma_{kk} \approx \sigma_{ll}$), and that $\eta_l \overset{\text{def}}{=} \eta \;\; \forall l$, we get

$$\hat{\mu}_k \approx \mu_k^{SI} + \frac{\sum_{l=1}^{|\mathcal{N}(k)|}(N_l+\tau)(\mu_l^* - \mu_l^{SI})}{\sum_{l=1}^{|\mathcal{N}(k)|}(N_l+\tau)} \tag{12}$$

which is similar to vector field correlation [9].
Also in (8) if we make the assumption that $(N_l + \tau) \gg \eta_l$ (i.e. asymptotic case),and $\eta_l \overset{\text{def}}{=} \eta = \tau \;\; \forall l$, we get

$$\begin{aligned}\hat{\mu}_k &\approx \mu_k^{SI} + \frac{\sum_{l=1}^{|\mathcal{N}(k)|} \frac{r_{kl}\sigma_{kk}}{\sigma_{ll}}(\mu_l^* - \mu_l^{SI})/(1-r_{kl}^2)}{\sum_{l=1}^{|\mathcal{N}(k)|} 1/(1-r_{kl}^2)} \\ &= \frac{\sum_{l=1}^{|\mathcal{N}(k)|} \frac{((r_{kl}\sigma_{kk}/\sigma_{ll})\mu_l^* + (\mu_k^{SI} - (r_{kl}\sigma_{kk}/\sigma_{ll})\mu_l^{SI}))}{(1-r_{kl}^2)}}{\sum_{l=1}^{|\mathcal{N}(k)|} 1/(1-r_{kl}^2)}\end{aligned} \tag{13}$$

This is the average of a set of linear transformations weighted by the factors $1/(1-r_{kl}^2)$, and is similar to that used in linear regression based prediction [3]. In the latter case the parameters of the linear transformation are estimated using linear regression. It is well known that for the normal assumption used in this paper the coefficients of the linear transformations in (13) coincide with those obtained using linear regression. However, the proposed approach allows for more general distribution assumptions, that could result in other estimates which are possibly nonlinear.

5. EXPERIMENTAL RESULTS

In this section we evaluate the proposed approach in speaker adaptation for TIMIT phonetic recognition, and for a confusable isolated word recognition database. In all the adaptation experiments, only the mean vectors are adapted, and no attempt was made to adapt the variances or mixture weights of the models.

5.1. TIMIT phonetic recognition

48 phone models representing 39 classes as in [7] are used. Each phone is modelled by a 3 state left to right HMM. The feature vector consists of 12 MFCC with cepstral mean normalization applied at the sentence level. There are 3696 sentence from 462 speakers in the training set. The test set consists of 192 sentences from 24 speakers (core test set). No phone grammar is used in the test. The 2 SA sentences of each test speaker are used for adaptation. The baseline recognition system is implemented using HTK. 16 speaker groups corresponding to male and female speakers in 8 dialect reigons are used to estimate the correlation structure (as in equations (9)-(10)) .
Phone recognition results (% correct) for the speaker independent system (SI), conventional MAP adaptation (MAP), and MAP adaptation in conjunction with correlation based prediction, for neighbourhood size 8 (CMAP8), and neighbourhood size 4 (CMAP4) are shown in Table 1. The three rows of the table show results for mixture size 1,2, and 4 respectively.

Mixture size	SI	MAP	CMAP4	CMAP8
1	47.42	49.65	49.99	49.95
2	47.53	50.21	50.87	51.03
4	47.73	50.30	50.13	50.20

Table 1. Percent correct phonetic recognition results on TIMIT database. For speaker indpendent (SI) models, speaker adapted models using MAP adaptation (MAP), and speaker adapted models using MAP adaptation in conjunction with correlation based prediction neighbourhood size 4 (CMAP4), and neighbourhood size 8 (CMAP8).

Slight improvement (in fact degradation for mixture size 4) compared to conventional MAP adaptation can be observed from the table. We attribute this performance to the relatively small amount of training data which is not sufficient to estimate an accurate correlation structure. In addition, the parameters are estimated from speaker groups (because there is no sufficient data to estimate a robust model for each speaker), and thus the estimated structure doesn't faithfully represent parameter movements for individual speakers. We also suspected that using very simple acoustic models may result in poor segmentation and hence inaccurate estimation of the correlation structure. To assess these points we performed numerous experiments; using the segmentation provided with TIMIT in correlation structure estimation, increasing the number of speaker groups (N) by using an automatic speaker clustering procedure, and using models from individual speakers and replacing sample averages in (9) and (10) by MAP estimates, but no improvement in results compared to Table-1 was observed. An explanation may be that neither of these trials resulted in a good balance between estimating robust speaker models, and using a sufficient number of groups to estimate an accurate correlation structure, and we are currently experimenting with a tying mechanism that considers the correlation between groups of distributions.

5.2. Isolated word recognition

The purpose of this experimental setting is to test the prediction power of the proposed method. A vocabulary V is divided into two subsets $V1$ and $V2$, where a closed test is carried on each subvocabulary. During model adaptation of a subvocabulary (say $V1$), only examples of the other subvocabulary (say $V2$) are presented for adaptation. These examples are used to obtain MAP estimates of the means of $V2$, which are used to predict the means of $V1$ (using equation (8)). The predicted means of $V1$ are then used in testing. The same steps are repeated by reversing the roles of both subvocabularies, and average results are reported.

In our experiments a 21 word vocabulary V consisting of 5 subsets of confusable words given by:

- A, J, K.
- B, C, D, E, G, P, T, V, Z, THREE.
- M, N.
- GO, NO, OH.
- F, S, X.

The first subvocabulary $V1$ consists of 11 words given by:

- A, J.
- B, D, G, T,Z.
- M.
- GO, NO.
- F.

The other subvocabulary $V2$ consists of the remaining 10 words. It should be noted that attempt was done to get an even distribution of words of the confusable subsets among the two subvocabularies.

The vocabulary is uttered by 24 speakers (12male/ 12female), each word is uttered two times by each speaker. The speakers are divided into two balanced groups $S1$ and $S2$ each containing 12 speakers. Hidden Markov models as well as correlation structure (equations (9)-(10)) estimated from one group are used in processing the other group. Thus each test consists of 4 smaller subtests resulting from the combinations of the speaker groups and the subvocabularies, and consists of 1008 trials.

We use 5 state left to right HMMs, having one Gaussian distribution per state. 12 MFCC with and without using the difference coefficients are used in the experiments. The speaker independent word recognition accuracies with and without using dynamic coefficients are 84.0% and 79.5% respectively. Table 2 shows the results of applying the scenario described at the beginning of this subsection for speaker adaptation. In MAP estimation of each word 2 repititions of the word are used. The table shows the results when varying the neighbourhood size from 4-16.

Neighbourhood size	4	8	12	16
12 MFCC	80.8	82.4	84.8	83.7
12 MFCC + 12 ΔMFCC	85.5	87.1	87.3	87.1

Table 2. Word recognition rate (%) using predictive adaptation for isolated word recognition task with and without using difference coefficients for different neighbourhood sizes.

As can be seen from the table, the speaker adapted results outperform the speaker independent ones for all neighbourhood sizes when using both static and static+dynamic coefficients. Also the improvement increases with increasing the neighbourhood size until it slightly degrades at neighbourhood size 16, probably due to inclusion of weakly correlated distributions in the prediction.

6. CONCLUSION

We have presented a predictive adaptation technique for the mean vectors of a hidden Markov model, which is potentially capable of influencing parameters of the models that have no observations in the adaptation speech. The technique is based on the principle of MMSE estimation to predict unobserved mean vectors. The predicition is a combined estimate from a set of mean vectors in the neighbourhood (in correlation sense) of the unobserved distribution. For the normal assumption used in this paper, some interesting theoretical relationships between the proposed method and both vector field correlation and linear regression based prediction are pointed out. The adaptation algorithm was successfully tested in speaker adaptation experiments for both TIMIT phonetic recognition, and an isolated word recognition task.

REFERENCES

[1] S.M.Ahadi,P.C.Woodland,"Rapid speaker adaptation using model prediction," Proc. IEEE ICASSP-95, pp. 684-687.

[2] P.J.Bickel, and K.A.Doksum, Mathematical statistics: Basic ideas and selected topics,Holden-Day Inc, 1977.

[3] S.J.Cox,"A speaker adaptation technique using linear regression," Proc. IEEE ICASSP-95, pp. 700-703.

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

[5] Q.Huo, and C.H.Lee,"On-line adaptive learning of the correlated continuous density hidden Markov model for speech recognition," Proc. ICSLP-96, Oct. 1996.

[6] M.Lasry, and R.Stern,"A posteriori estimation of correlated jointly Gaussian Mean vectors," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 6, No. 4, pp.530-535, July 1984.

[7] K.F.Lee, and H.W.Hon,"Speaker independent phone recognition using hidden Markov models," IEEE Trans. Acoustics, Speech, and Signal processing, pp. 1641-1648, Nov. 1989.

[8] B.M.Shahshahani,"A Markov random field approach to Bayesian speaker adaptation," Proc. IEEE ICASSP-96, pp.697-700, 1996.

[9] S.Takahashi, and S. Sagayama, "Tied structure HMM based on parameter correlation for efficient model training," Proc. IEEE ICASSP-96, May 1996.

[10] M.Tonomura, T.Kosaka, S.Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation," Computer Speech and Language, pp. 117-132, Jan. 1996.

[11] G.Zavaliagkos, R.Schwartz, and J.McDonough, "Maximum a posteriori adaptation for large scale HMM recognizers," Proc. IEEE ICASSP-96, pp.725-728, 1996.

CORRELATION BASED PREDICTIVE ADAPTATION OF HIDDEN MARKOV MODELS

Mohamed Afify[1] , *Yifan Gong*[1,2] *and Jean-Paul Haton*[1]

[1]CRIN/CNRS-INRIA-Lorraine,B.P. 239 54506 Vandeouvre,Nancy,France

[2] Media Technologies Laboratory, Texas Instruments, P.O.BOX 655303 MS 8374, Dallas TX 75265, U.S.A.

Hidden Markov model (HMM) adaptation is currently of interest, to overcome the degradation effect of speaker and/or channel mismatches in practical speech recognition systems. The Bayesian framework provides a theoretically optimal formulation for combining adaptation data and prior knowledge, but it suffers from the drawback of being incapable of adapting parameters of the models that have no observations in the adaptation speech. In this article we present a new predicitve (in the sense of influencing unobserved distribution parameters) adaptation algorithm for the mean vectors of an HMM. We also point out some theoretical relationships between the proposed method and other techniques used in the context of predictive model adaptation. The efficacy of the proposed approach is demonstrated in speaker adaptation experiemnts for both an isolated word task, and TIMIT phonetic recogntion.