LATENT CORRELATION ANALYSIS OF HMM PARAMETERS FOR SPEECH RECOGNITION[1]

Zhijian Ou, *Jun Luo*

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
ozj@tsinghua.edu.cn

ABSTRACT

Correlation between HMM parameters has been utilized for various rapid speaker adaptation, e.g. eigenvoice adaptation. The covariance matrix of the supervector which is a concatenation of all the Gaussian means in HMM, is clearly a good measure of such parameter correlation. In this paper, we propose to treat the supervector as a latent variable under HMM, and perform estimation of the hidden supervector's covariance matrix directly from the acoustic frames using EM algorithm. In contrast to traditional methods which depend on using well-trained/adapted supervector samples, the proposed method is more theoretically sound and capable of dealing well with speaker-specific data sparseness. Moreover, the idea of conducting utterance-level correlation analysis, estimating utterance eigenvoices, and performing (unsupervised) utterance adaptation is explored. Experiments on the OGI Numbers database show that the proposed approach achieves better adaptation performance than the traditional methods, and the utterance-level correlation analysis is found to be useful.

Index Terms— correlation analysis, HMM, eigenvoice

1. INTRODUCTION

The most common acoustic model in current speech recognition systems is hidden Markov model (HMM). However, some independence assumptions associated with HMMs as they are used in speech recognition have been known to ignore certain types of correlation that exist in the speech signal. A well-known type is the temporal correlation between successive feature frames (ignored by the HMMs' state-conditional independence assumption). Another type, which has received much attention recently, is the correlation between different sounds, as a consequence of the constant or slowly changing characteristics of some underlying factors (e.g. the speaker, acoustic environment, speaking style, emotional state, etc.) [1]. In HMM-based acoustic modeling, such correlation is actually the correlation between the model parameters representing different sounds, which we refer to as parameter correlation in this paper. It is usually assumed that the parameters are independent in both model training and decoding. One example of the non-realistic result of such independence assumption is that the speech recognition system may assign models from different speakers or emotional states to different parts of an utterance.

[1] This work was supported by National Natural Science Foundation of China (No. 60402029)

Note that directly incorporating correlation between HMM parameters at the decoding phase is computationally intractable. Currently, people usually utilize such parameter correlation in speaker adaptation. By first analyzing the training corpus with various speakers, we can obtain the desirable parameter correlation. Then such a priori information about the inter-speaker variation can be used to derive constraints for rapid speaker adaptation. Various methods have been proposed with different ways to represent the parameter correlation, e.g. RMP [2], and eigenvoice [3], etc. Among them, the eigenvoice approach is more attractive, since it is based on the covariance matrix of the parameters, which is clearly a good measure of the parameter correlation.

In eigenvoice modeling for speaker adaptation, the parameters (usually the Gaussian means) in any speaker-dependent (SD) model are concatenated to form a (speaker) supervector. The supervector's covariance matrix is estimated simply as the sample covariance matrix from a set of training speaker supervectors, and then sent to principle component analysis (PCA) to obtain the dominant eigenvectors, namely eigenvoices. A basic assumption here is that we consider the parameters jointly to be an observable supervector and we have a set of well-trained speaker supervectors as its samples/observations (that can be separately obtained for every training speaker). In practice, people often resort to MLLR [4] adaptation to create the SD models [5], when there are less sufficient speaker-specific data (may having unseen phones).

In this paper, we first provide a new and more theoretically sound method to conduct correlation analysis of HMM parameters, capable of dealing well with speaker-specific data sparseness. Specifically, we consider the parameters jointly to be a latent (Gaussian) supervector under HMM, and perform estimation of the hidden supervector's covariance matrix directly from the acoustic feature frames using EM algorithm [6]. There is no need to have supervector samples explicitly. This results in a general latent correlation analysis of HMM parameters. Second, note that, as we know, parameter correlation arises from certain underlying factors that are consistent throughout an utterance [1]. In speaker adaptation, we solely consider the underlying factor to be speaker-related, conduct speaker-level correlation analysis of parameters, and obtain speaker eigenvoices. In this case, we in fact pool together all utterances of each speaker, as if there were just one utterance per speaker. A further idea is that we can conduct utterance-level correlation analysis, use PCA similarly to obtain utterance eigenvoices, and then perform (unsupervised) utterance adaptation. With the above general analysis procedure, it is straightforward to implement this idea.

Experiments are carried out on the OGI Numbers database [7], which is an English telephone speech corpus consisting of continuously spoken numbers. We provide figures that show how sparse the data is. The results show the effectiveness of the

proposed latent correlation analysis approach, and the utterance-level correlation analysis is found to be useful.

The paper is organized as follows. In section II, we describe the proposed latent correlation analysis of HMM parameters. In section III, we outline the maximum a posteriori (MAP) eigenvoice adaptation, which is used to exploit the estimated parameter correlation. Section IV presents experimental results, followed by conclusions in the last section.

2. LATENT CORRELATION ANALYSIS OF HMM PARAMETERS

For the study of correlation between HMM parameters in this paper, we only consider the Gaussian means and concatenate them to form a supervector x, supposed to be randomly distributed with mean μ and covariance matrix $\sum$, and generally represent the underlying factors consistent throughout an utterance. $x=\left(x_1^T,\cdots,x_D^T\right)^T$ has D subvectors x_i, $i=1,\cdots,D$, where D denotes the total number of Gaussian components in the speaker-independent HMM. The generative model of an utterance incorporating the supervector variable x is shown in Fig.1, plotted as a Bayesian network [8]. Here q_t, y_t are respectively the state variable and the acoustic feature variable at frame t. To simplify notation, we assume that the value of q_t represents a combination of the HMM state index and Gaussian component index, and is used as the global index to the subvector in the supervector (i.e. the Gaussian pool). The conditional distributions at node x and y_t are respectively:

$$p(x)=\mathcal{N}(x\mid\mu,\Sigma) \tag{1}$$

$$p(y_t\mid x,q_t=i)=\mathcal{N}(y_t\mid x_i,C_i) \tag{2}$$

where C_i, $i=1,\cdots,D$, is the (diagonal) covariance matrix for the *i*-th Gaussian component. (Note that C_i is not the speaker-independent covariance matrix in the usual sense.) It turns out that to conduct correlation analysis for supervector x (i.e. obtain the covariance matrix $\sum$) is essentially translated to perform parameter estimation for the above generative model of speech, including μ, $\sum$ and $C_{1:D}$. (Here and also in the following, we use a set of subscripts to denote the corresponding set of variables, and $1:D$ represents the set $\{1,\cdots,D\}$.)

2.1. Parameter estimation using EM algorithm

The EM algorithm [6] provides a general approach to the problem of maximum likelihood parameter estimation in statistical models with hidden variables. In the EM algorithm, we need to compute the conditional distribution of all hidden variables in the model (i.e. x and q_t's here), given observations (i.e. y_t's). This inference problem is intractable, if we consider the Bayesian network representation of the model shown in Fig.1. After moralization, the hidden supervector x will be connected with all hidden state variables q_t's to form one big clique [8]. Thus we assume that we have observed q_t's. That is, each frame in the training data is supposed to have been aligned to a mixture component, using

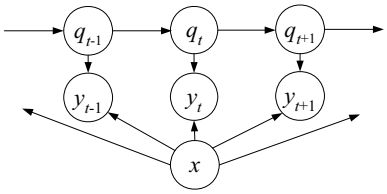

Fig.1: Bayesian network representation of the generative model of speech, incorporating the supervector variable *x*.

Viterbi alignment (though a forward-backward alignment can also be used) with the speaker-independent HMM. Denote the aligned training data by N utterances, $\left(q_{1:T_n}^{(n)},y_{1:T_n}^{(n)}\right)$, each with the length of T_n, $n=1,\cdots,N$. Denote the current parameter estimate as $\bar{\theta}=\left\{\bar{\mu},\bar{\Sigma},\bar{C}_{1:D}\right\}$. The auxiliary function $Q\left(\theta,\bar{\theta}\right)$ is defined as follows:

$$Q(\theta,\bar{\theta})=\sum_{n=1}^{N}\int_{x^{(n)}}p\left(x^{(n)}\mid\bar{\theta},q_{1:T_n}^{(n)},y_{1:T_n}^{(n)}\right)\log p\left(x^{(n)},q_{1:T_n}^{(n)},y_{1:T_n}^{(n)}\mid\theta\right)dx^{(n)}$$

Let $o(n)$, $h(n)=\{1:D\}\setminus o(n)$ denote the mixture components appearing and missing in $q_{1:T_n}^{(n)}$ respectively. For $i\in o(n)$, let $L_i^{(n)}$ be the number of frames aligned to mixture component i in the n-th utterance, and $m_i^{(n)}$, $B_i^{(n)}$ be the corresponding sample mean and sample covariance matrix of these frames. Let $\mathbf{L}^{(n)}$, $\mathbf{C}$ be the block diagonal matrix, with $L_i^{(n)}I$ and C_i as the $F\times F$ diagonal blocks respectively for $i=1,\cdots,D$, where F is the dimension of the acoustic frame, and I is the identity matrix. For $i\in h(n)$ (i.e. unseen phones), we set $m_i^{(n)}$=0.

With the above accumulated statistics, it can be shown that the posterior distribution of $x^{(n)}$ given $y_{1:T_n}^{(n)},q_{1:T_n}^{(n)}$ is Gaussian with mean and covariance matrix as follows:

$$\mu^{(n)post}=\Sigma^{(n)post}\left(\mathbf{L}^{(n)}\bar{\mathbf{C}}^{-1}m_{1:D}^{(n)}+\bar{\Sigma}^{-1}\bar{\mu}\right) \tag{3}$$

$$\Sigma^{(n)post}=(\mathbf{L}^{(n)}\bar{\mathbf{C}}^{-1}+\bar{\Sigma}^{-1})^{-1} \tag{4}$$

So the auxiliary function $Q\left(\theta,\bar{\theta}\right)$ can be rewritten as:

$$\begin{aligned}Q\left(\theta,\bar{\theta}\right)=&\sum_n\left\langle\log p\left(x^{(n)}\mid\theta\right)\right\rangle_n\\&+\sum_n\left\langle\log p\left(y_{1:T_n}^{(n)}\mid\theta,x^{(n)},q_{1:T_n}^{(n)}\right)\right\rangle_n+\text{constant}\end{aligned} \tag{5}$$

where we use the notation $\left\langle f\left(x^{(n)}\right)\right\rangle_n$ to denote the expectation for arbitrary function $f\left(x^{(n)}\right)$ of $x^{(n)}$, under the posterior distribution $p\left(x^{(n)}\middle|\bar{\theta},y_{1:T_n}^{(n)},q_{1:T_n}^{(n)}\right)$.

To maximize the auxiliary function (5) with respect to μ and $\sum$, only the first term in the right-hand side of (5) is involved. By setting the corresponding derivatives to zero, we can obtain the re-estimate of μ and $\sum$ as:

$$\hat{\mu}=\mu^{post},\quad \hat{\Sigma}=\Sigma^{post}+\frac{1}{N}\sum_n\Sigma^{(n)post} \tag{6}$$

where μ^{post}, Σ^{post} are computed respectively as the sample mean and sample covariance matrix over the data $\mu^{(n)post}$, $n=1,\cdots,N$.

To maximize the auxiliary function (5) with respect to $C_{1:D}$, only the second term in the right-hand side of (5) is involved. By setting the corresponding derivative to zero, we can obtain the re-estimate of $C_{1:D}$ as:

$$\hat{C}_i = \frac{\sum_{n:\, i\in o(n)} L_i^{(n)} \left[B_i^{(n)} + \left(\mu_i^{(n)post} - m_i^{(n)}\right)\left(\mu_i^{(n)post} - m_i^{(n)}\right)^T + \Sigma_i^{(n)post} \right]}{\sum_{n:\, i\in o(n)} L_i^{(n)}}$$

where $\Sigma_i^{(n)post}$ is the covariance block at the diagonal of $\Sigma^{(n)post}$ corresponding to subvector x_i.

2.2. Discussions

There have been some methods to address the issue of how to estimate the eigenvoices when there are less sufficient speaker-specific data (may having unseen phones). A practical method is to use MLLR adaptation to establish SD supervectors. Classical MAP adaptation is not suitable for this purpose, since it only transforms the observed parameters, which leaves unseen phones still unchanged. EMAP adaptation may be used [9]. Note that the step of computing (3) is essentially to perform EMAP adaptation to obtain the posterior mean $\mu^{(n)post}$. However, the re-estimate formula (6) tells us that simply using the sample covariance matrix of the adapted supervectors, like Σ^{post}, as the estimate of Σ is insufficient, which ignores the posterior covariance $\Sigma^{(n)post}$. Thus, the traditional methods (treating supervector as observed and using sample covariance matrix alone) seem to be more heuristic motivated and not theoretically sound.

The method by iteratively interleaving eigenvoice adaptation and eigenspace estimation in [10] also introduces latent variable for analysis. However, it is required that we know the desirable number of eigenvoices beforehand and keep it fixed. In practice, the number of eigenvoices may be adjusted according to the amount of adaptation data. Additionally, the resulting basis vectors after iterative estimations are not guaranteed to be orthogonal, and so they are not strictly eigenvectors. Moreover, there are no corresponding eigenvalues which measure the importance of each basis vector. In our approach, the computationally intensive steps of supervector covariance matrix estimation and PCA are carried out once, and then all eigenvectors of interest along with their corresponding eigenvalues are obtained.

Remarkably, the idea of conducting utterance-level correlation analysis, estimating utterance eigenvoices, and performing (unsupervised) utterance adaptation is proposed and experimentally evaluated, showing its usefulness.

3. MAP EIGENVOICE ADAPTATION

Currently, parameter correlation is mainly used in (speaker) adaptation. Once we obtain the supervector covariance matrix Σ, we can perform PCA to get the desirable eigenvoices, say, $e_1,\cdots,e_R$, and the corresponding eigenvalues $\lambda_1,\cdots,\lambda_R$, where R is the dimension of the eigenspace. Suppose that every supervector x can be written in the form:

$$x = \mu + \sum_{r=1}^{R} w_r \cdot e_r \tag{7}$$

where $w_{1:R}$ is the combination weights. The adaptation problem then reduces to the estimation of the combination weights from the adaptation data.

MAP estimation is a good choice for this purpose [5][11]. Denote the adaptation data as $y_{1:T}$, and suppose that each frame has been aligned to mixture component i with the occupation probability $\gamma_i(t)$. The MAP estimate of the weights $w_{1:R}$ is:

$$\hat{w}_{1:R} = \arg\max_{w_{1:R}} \left[p\left(y_{1:T} \mid w_{1:R}\right) p_0\left(w_{1:R}\right) \right] \tag{8}$$

Here $p_0\left(w_{1:R}\right)$ is the a priori distribution, which can be easily derived from the result of eigen-analysis of the supervector covariance matrix as [11]:

$$\log p_0\left(w_{1:R}\right) = -\frac{1}{2}\sum_{r=1}^{R}\frac{w_r^2}{\lambda_r} \tag{9}$$

By taking derivative to $w_{1:R}$, the maximization problem in (8) can be solved as follows, for $r=1,\cdots,R$:

$$\begin{aligned} &\sum_i \sum_t \gamma_i(t)\, e_{r,i}{}^T C_i^{-1}\left(y_t - \mu_i\right) \\ &= \sum_{k=1}^{R} w_k \cdot \left[\sum_i \sum_t \gamma_i(t)\, e_{r,i}{}^T C_i^{-1} e_{k,i} + \delta_{k,r}\frac{1}{\lambda_r} \right] \end{aligned} \tag{10}$$

where $e_{r,i}$ is the subvector corresponding to mixture component i in the r-th eigenvoice e_r, and $\delta_{k,r}=1$ iff. $k=r$, otherwise it equals to zero.

In theory, we can iterate the occupation probability computation and the weight estimation. In the first iteration, the occupation probabilities are computed using an SI model. In subsequent iterations, they are computed using the adapted model. In the experiment, we perform only one iteration, since further iterations are observed to give minor differences.

4. EXPERIMENTAL RESULTS

Experiments are carried out on the OGI Numbers database [7], which is an English telephone speech corpus consisting of naturally spoken numbers with 30-word vocabulary[2]. We use 6049 utterances spoken by 3059 speakers for training and 2061 utterances by 1044 speakers for testing.

The acoustic feature is 39-dimensional, formed by 12 MFCCs with normalized log-energy and their first and second order differentials. Cepstral mean subtraction (CMS) is applied to the feature vector. There are 26 monophone models, a silence model, and a short-pause model. The silence and all monophones are modeled with three emitting states each, and the short-pause has only one state which is tied to the middle state of the silence model. Gaussian mixture model (GMM) with diagonal covariance matrices is used for state-output distributions.

[2] This is a relatively simple task. It will be shown below how sparse the data is in this task, and thus is suitable for our study.

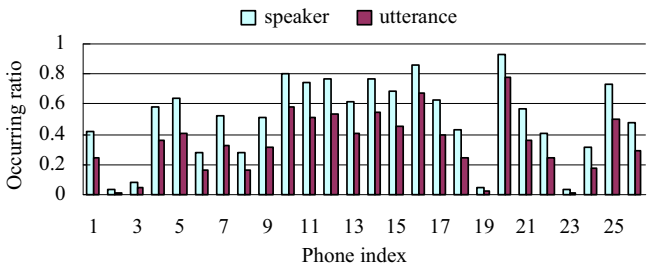

Fig.2: The occurring ratio for each of the 26 phones over all training speakers and all training utterances separately.

Table 1: %WER results for various methods

Mixture num per state		1	2	4
Baseline		20.86	16.85	13.34
Speaker adaptation	MLLR	20.71	16.79	13.25
	MAP	20.75	16.83	13.32
	MLLR+EV	20.79	16.27	12.59
	EM+EV	**18.42**	**15.76**	**12.44**
Utterance adaptation	MLLR	20.71	16.80	13.29
	MAP	20.75	16.86	13.24
	MLLR+EV	20.81	16.62	13.20
	EM+EV	**18.31**	**15.20**	**11.97**

To alleviate computation cost, we split the acoustic feature vector into 1-dimension streams. And for each stream separately, we use the algorithm in section II to estimate the set of parameters $\theta = \{\mu, \Sigma, C_{1:D}\}$, and then perform PCA on $\sum$ to obtain the desirable eigenvoices. For EM initialization, we copy the means and variances of the Gaussian components in the SI model to construct μ and (diagonal) $\sum$ respectively.

There are two cases. One is that we treat all utterances from a speaker as just one utterance, conduct speaker-level correlation analysis of parameters, and obtain speaker eigenvoices, in the traditional sense. Another case is that we process each utterance individually, conduct utterance-level correlation analysis, and obtain utterance eigenvoices. For both cases, 5 eigenvoices per stream are used in the experiments. Then we conduct unsupervised MAP eigenvoice adaptation for the 1044 test speakers (speaker adaptation) and 2061 test utterances (utterance adaptation) separately.

Before giving the recognition results, we show some statistics in Fig.2. A phone's occurring ratio over speakers is defined as the ratio of the total counts of the phone to the number of speakers, where we use binary counting of the occurrence of the phone for each speaker. A phone's occurring ratio over utterances is similarly defined. To sum up over all phones, we can see that on average, a speaker observes only 13.14/26≈50.5% of the 26 phones, and an utterance observes fewer phones with 8.76/26≈33.7%.

The word error rate (WER) results of the baseline and various adaptation methods are listed in Table 1, for both cases of speaker adaptation and utterance adaptation. "MLLR+EV" refers to using MLLR adaptation to establish speaker/utterance-specific models for eigenspace estimation. "EM+EV" refers to using the proposed latent correlation analysis via EM algorithm for eigenspace estimation. We experiment with varying number of Gaussian mixture components for HMM states. The results show that for both cases of speaker adaptation and utterance adaptation, the EM+EV system outperform the MAP, MLLR and MLLR+EV systems consistently. Moreover, utterance adaptation using "EM+EV" seems to be more effective than speaker adaptation using "EM+EV", though we have sparser phone occurrence in utterance-level analysis. This may be explained by the fact that there exist certain consistent underlying factors throughout an utterance, besides speaker information.

5. DISCUSSIONS AND CONCLUSIONS

In this paper, we first provide a new method to conduct correlation analysis of HMM parameters (i.e. estimate the supervector covariance matrix), capable of dealing well with speaker-specific data sparseness. Examination of the estimation formula in the proposed method makes clear the deficiency of the traditional methods which depend on using well-trained/adapted supervector samples. Second, the idea of utterance supervector (in contrast to speaker supervector) is proposed, and applied in estimating utterance eigenvoices and performing (unsupervised) utterance adaptation. Furthermore, we can consider a new way of speaker modeling, which is to use a speaker-specific distribution of the latent utterance supervector to model the utterances from a speaker. The result here is an encouraging step toward this end. In future, we also plan to apply the proposed method in large vocabulary experiments.

6. REFERENCES

[1] M. Blomberg, "Within-utterance correlation for speech recognition," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 2479–2482.

[2] S. Ahadi, P. Woodland, "Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density HMMs," *Comput. Speech Lang.*, vol. 11, pp. 187–206, 1997.

[3] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech and Audio Processing*, no. 6, pp. 695–707, 2000.

[4] 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.

[5] H. Botterweck, "Anisotropic MAP defined by eigenvoices for large vocabulary continuous speech recognition," in *Proc. ICASSP*, Salt Lake City, UT, May 2001, pp. 353–356.

[6] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, pp. 1–38, 1977.

[7] R. M. Fanty, M.Noel, and T.Lander, "Telephone speech corpus development at CSLU," in *Proc. ICSLP*, 1994.

[8] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, D.J. Spiegelhalter, *Probabilistic Networks and Expert Systems*, Springer-Verlag, 1999.

[9] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in *Proc. ICASSP*, Detroit, 1995, pp. 676–679.

[10] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on Speech and Audio Processing*, no. 3, pp. 345–354, May 2005.

[11] J. Luo, Z. Ou, and Z. Wang, "Fast eigenspace-based MAP adaptation within correlation subspace," *Journal of Tsinghua University (in Chinese)*, pp. 829–832, 2004.