DERIVATION OF EIGENTRIPHONES BY WEIGHTED PRINCIPAL COMPONENT ANALYSIS

Tom Ko and *Brian Mak*

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong

{tomko, mak}@cse.ust.hk

ABSTRACT

Last year we proposed a new acoustic modeling method called *eigentriphones* in which all triphones are distinct (with no tied states) so that they may be more discriminative. In our method, frequent triphones are used to derive an eigenbasis using PCA, and the infrequent triphones are then "adapted" as a linear combination of the eigenvectors which are also called *eigentriphones*. Although the eigentriphones method compares favorably with traditional tied-state triphones, the PCA procedure has two limitations: (1) only the frequent triphones are employed, and (2) they are considered "equal" even though some are more robust than the others. In this paper, weighted PCA is proposed to solve both problems so that *all* triphones — frequent and infrequent triphones — may contribute to the derivation of the eigentriphones, each at a different extent depending on its sample count. Experimental evaluation on the WSJ 5K-vocabulary speech recognition task shows that weighted PCA produces better models than simple PCA, and its performance is fairly independent of the number of eigentriphones once more than 20% of them are used. As a consequence, all triphones may be represented by fewer eigentriphones, resulting in a more compact model.

Index Terms: Eigentriphones, eigenvoice adaptation, context-dependent acoustic modeling, weighted PCA.

1. INTRODUCTION

In acoustic modeling, one has to strike a balance between detailed context-dependency (CD) modeling [1] and robust training. The number of context-dependent modeling units (e.g. triphones or quinphones) grows exponentially with the contexts while the training samples for the contexts distribute unevenly. For instance, it is found that on the Wall Street Journal corpus, 80% of the data are contributed by 20% of the triphones observed in the corpus. How to train the infrequent context-dependent units robustly is one of the major problems in acoustic modeling.

There are three common approaches:

- Parameter tying (sharing) as done in generalized triphones [1], state tying [2], shared distributions or senones [3], and tied subspace Gaussian distributions [4].
- Model interpolation: detailed models are interpolated with models of lower resolution as in generalized triphones [1] and back-off discriminative acoustic model [5].
- Basis approach: a common basis is constructed, and all models are represented by some combination of the basis components. This includes subspace Gaussian mixture model [6], Bayesian sensing HMM [7], and canonical state model [8].

Last year we proposed a new acoustic modeling method called *eigentriphone* [9, 10] in which all triphones are distinct (with no tied states) so that they can be more discriminative. In eigentriphone modeling, frequent triphones, which can be robustly trained, are used to derive an eigenbasis using *principal component analysis* (PCA), and the infrequent triphones are then derived as a linear combination of its eigenvectors which are also called *eigentriphones*. Although the eigentriphone method compares favorably with traditional tied-state triphones, the PCA procedure has two limitations: (1) only a subset of triphones — the frequent triphones — may be employed to derive the eigenbasis, and (2) all the frequent triphones employed for PCA are considered "equal" although some are considered more robust than the others. In this paper, weighted PCA [11] is used to solve both problems so that all triphones — frequent and infrequent triphones — may contribute to the derivation of eigentriphones, each at a different extent depending on its frequency.

This paper is organized as follows. In Section 2, we will review our eigentriphone modeling method. The proposed improvement using weighted PCA is described in Section 3. That is followed by experimental evaluation in Section 4 and conclusions in Section 5.

2. REVIEW OF EIGENTRIPHONE MODELING

We will first briefly review our current derivation procedure of model-based eigentriphones as described in [10], and then describe the proposed improvements in the next Section.

The eigentriphone approach for acoustic modeling is inspired by the eigenvoice method [12] in speaker adaptation. Speaker-dependent models in eigenvoice are replaced by triphone models in eigentriphone, and the derivation of eigentriphones is repeated for each base phoneme (or monophone). Thus, since there are 39 base phonemes in our systems, 39 sets of eigentriphones have to be derived.

At the core of eigentriphone modeling is the use of PCA to derive an eigenbasis. To make sure the derived basis is reliable, in the past, triphones of a base phoneme are divided into 2 groups: the rich set whose triphones have sufficient training samples for robust training, and the poor set whose triphones have to be adapted using the eigentriphones.

This work was supported by the Research Grants Council of the Hong Kong SAR under the grant number DAG05/06.EG43.

The following procedure is repeated for each base phone i using its triphones that appear in the training corpus.

STEP 1: Monophone hidden Markov model (HMM) of base phoneme i is first estimated from the training data. Each monophone is a 3-state strictly left-to-right HMM, and each state is represented by an M-component Gaussian mixture model (GMM).

STEP 2: The monophone HMM is then cloned to initialize *all* its triphones. *No state tying is performed.*

STEP 3: Categorize each triphone q of base phoneme i into one of the following two (possibly overlapping) sets based on its training sample counts n_{iq} and two thresholds θ_m^R and θ_m^P:

- the rich triphone set Ω_i^R if $n_{iq} \geq \theta_m^R$, or
- the poor triphone set Ω_i^P if $n_{iq} < \theta_m^P$.

STEP 4: Only Gaussian means of the rich triphones are re-estimated. Their Gaussian covariances, mixture weights, and transition probabilities are copied from their base phoneme HMM.

STEP 5: For each rich triphone $r \in \Omega_i^R$, create a triphone supervector $\mathbf{v}_{ir}$ by stacking up all Gaussian mean vectors from its three states as below

$$\mathbf{v}_{ir} = \begin{bmatrix} \boldsymbol{\mu}_{ir11}, & \boldsymbol{\mu}_{ir12}, & \cdots, & \boldsymbol{\mu}_{ir1M}, \\ \boldsymbol{\mu}_{ir21}, & \boldsymbol{\mu}_{ir22}, & \cdots, & \boldsymbol{\mu}_{ir2M}, \\ \boldsymbol{\mu}_{ir31}, & \boldsymbol{\mu}_{ir32}, & \cdots, & \boldsymbol{\mu}_{ir3M} \end{bmatrix}, \quad (1)$$

where $\boldsymbol{\mu}_{irjm}, j = 1, 2, 3$, and $m = 1, 2, \ldots, M$ is the mean vector of the mth Gaussian component at the jth state of triphone r.

STEP 6: Derive an eigenbasis from the correlation matrix of all rich triphone supervectors $\mathbf{v}_{i1}$, $\mathbf{v}_{i2}$, ..., $\mathbf{v}_{i|\Omega_i^R|}$ using *principal component analysis* (PCA).

STEP 7: The supervector $\mathbf{v}_{ip}$ of any poor triphone $p \in \Omega_i^P$ is assumed to lie in the eigenbasis as follows:

$$\mathbf{v}_{ip} = \mathbf{e}_{i0} + \sum_{k=1}^{|\Omega_i^R|} w_{ipk}\mathbf{e}_{ik} , \quad (2)$$

where $\mathbf{e}_{i0}$ is the mean of all triphone supervectors of phoneme i, $\{\mathbf{e}_{ik}, k = 1, 2, \ldots, |\Omega_i^R|\}$ are the eigenvectors arranged in descending order of their eigenvalues λ_{ik}, and $\mathbf{w}_{ip} = [w_{ip1}, w_{ip2}, \ldots, w_{ipK_i}]$ is the eigentriphone coefficients vector of triphone p in the "eigentriphone space".

STEP 8: Estimate the eigentriphone coefficient vector $\mathbf{w}_{ip}$ of any poor triphone p by maximizing the following penalized log likelihood function:

$$Q(\mathbf{w}_{ip}) = L(\mathbf{w}_{ip}) - \beta \left(\sum_{k=1}^{|\Omega_i^R|} \frac{w_{ipk}^2}{\lambda_{ik}} \right) , \quad (3)$$

where $L(\cdot)$ is the log likelihood of the training data, and β is the regularization factor.

STEP 9: The Gaussian mean of the mth mixture at the jth state of poor triphone p can be obtained from $\mathbf{v}_{ip}$ as

$$\boldsymbol{\mu}_{ipjm} = \mathbf{e}_{i0jm} + \sum_{k=1}^{K_i} w_{ipk}\mathbf{e}_{ikjm} . \quad (4)$$

STEP 10: The Gaussian covariances, mixture weights, and transition probabilities of triphones in the rich set are then re-estimated.

3. IMPROVEMENT WITH WEIGHT PCA

In this paper, we investigate further improvement to the above eigentriphones derivation procedure in two aspects:

- To avoid the ad hoc categorization of triphones into the rich or poor set. Instead, *all* triphones may contribute to the derivation.
- Due to the uneven distribution of training data, some triphones are trained more robustly than the others. It is desirable to incorporate some notion of triphone reliability in the construction of the eigenbasis.

Weighted PCA is a natural solution to both problems.

3.1. Weighted PCA (WPCA)

There are generally two ways to perform weighted PCA [11]:

- weigh each variable in the feature vector differently. This also may be considered as a generalization of using the correlation matrix for PCA, and normalize the covariance matrix by weights other than the standard deviations of the variables.
- weigh each observation differently. This may also be considered as a re-sampling of the observations.

Here we adopt the second form of weighted PCA, and weigh each triphone supervector by its sample count. In other words, we assign a reliability measure to each triphone supervector that is directly proportional to its sample count. Thus, the PCA procedure in STEP 6 of Section 2 is replaced by *weighted PCA*. That is, the new correlation matrix for PCA is computed from the following covariance matrix:

$$\frac{1}{n_i} \sum_q n_{iq}(\mathbf{v}_{iq} - \bar{\mathbf{v}}_i)(\mathbf{v}_{iq} - \bar{\mathbf{v}}_i)' , \quad (5)$$

where n_{iq} is the sample count of the triphone q of base phoneme i, $n_i = \sum_q n_{iq}$, and $\bar{\mathbf{v}}_i$ is the new weighted mean of the triphone supervectors.

3.2. Pruning of Eigentriphones

As will be seen in the next Section on experimental results, the use of weighted PCA has the additional benefit that the eigenspectrum is concentrated more in the eigenvectors with higher eigenvalues. As a result, fewer eigentriphones (i.e., a smaller subspace) may be employed to represent *all* triphones[1] with little or no performance degradation, resulting in a much more compact but distinctive set of triphone models.

3.3. Other Changes

Lastly, we also modify STEP 10 of Section 2 as follows. Firstly, instead of using the covariances from monophones, covariances from the tied-state triphones are copied to the "eigentriphone-adapted" triphones. Then the covariances, mixture weights, and transition probabilities of those triphones with sample counts exceeding the thresholds θ_v, θ_w, and θ_t are re-estimated. Secondly, in the past, unseen triphones adopted the HMM parameters from the monophones. Now their parameters are determined from the tied-state triphones.

[1] In the past, only the poor triphones were represented as points in the eigentriphone space, and the rich triphones had their own conventional HMM descriptions.

Table 1. Information of various data sets.

Data Set	#speakers	#utterances	vocab size	OOV
SI284	283	37,413	13,646	11.95%
si_dt_05.odd	10	248	1,260	0
Nov'93	10	215	1,004	0.29%

4. EXPERIMENTAL EVALUATION

4.1. Speech Corpora and Experimental Setup

The standard SI-284 Wall Street Journal (WSJ) training set was used for training the speaker-independent model. It consists of 7,138 WSJ0 utterances from 83 WSJ0 speakers and 30,275 WSJ1 utterances from 200 WSJ1 speakers. Thus, there is a total of about 70 hours of read speech in 37,413 training utterances from 283 speakers. All the training data are endpointed.

The standard Nov'93 5K non-verbalized test set were used for evaluation using the standard 5K-vocabulary bigram that came along with the WSJ corpus. The set si_dt_05.odd contains alternate sentences from the 1993 WSJ 5k Hub development test set after sentences with OOV words were removed. It was used to tune the system parameters. A summary of these data sets is shown in Table 1.

There were altogether 18,777 cross-word triphones based on 39 base phonemes. Each triphone model was a strictly left-to-right 3-state continuous-density hidden Markov model (CDHMM), with a Gaussian mixture density of at most $M = 16$ components per state. In addition, there were a 1-state short pause model and a 3-state silence model. The traditional 39-dimensional MFCC vectors were extracted at every 10ms over a window of 25ms.

Recognition was performed using the HTK toolkit [13] with a beam search threshold of 350.

4.2. Baseline Systems

Three baseline systems were trained for comparison.

- Baseline1: Eigentriphone modeling result using conventional PCA as in [10].
- Baseline2: A conventional tied-state triphone system. There were totally 6,481 tied states which were derived from a phonetic decision tree. The number of tied-states was selected to maximize the accuracy on the development set.
- Baseline3: A triphone system with no tied states. The corresponding monophone system was first trained and then cloned to initialize the triphones. Then the Gaussian means of triphones were re-estimated with Baum-Welch training.

The recognition results of these baselines are shown in Table 2.

4.3. Eigentriphone Acoustic Modeling

The eigentriphone model-based adaptation was carried using the baseline3 models according to the procedure described in Section 2. The dimension of each triphone supervectors is 3 (states) $\times 16$ (mixtures) $\times 39$ (MFCC) $= 1872$ parameters. The standard ML training was done using HTK [13]. There are altogether 16,713 triphones[3] used to deduce the 39 sets of eigentriphone. The regularization parameter β was set to 1.0, and the sample count thresholds for the re-estimation of covariances, mixture weights and transition probabilities were set to $\theta_v = \theta_w = \theta_t = 200$.

[2]20% of eigentriphones give the best results on the development data set.

[3]By default, triphones with no more than 2 samples are not updated by HTK. Thus, every reference triphone here has at least 3 samples.

Table 2. Recognition word accuracy (%) of various systems on the WSJ 5K task using bigram language model. (The figure with an $*$ is statistically and significantly better than Baseline1 result.)

Model	Description	Nov'93
Baseline1	eigentriphone modeling result using PCA (Interspeech 2011 [10])	92.44
Baseline2	tied-state triphones	91.97
Baseline3	no state tying; only Gaussian means of *all* triphones are trained	90.34
	+ eigentriphone "adaptation" using *weighted PCA* for the Gaussian means of *all* triphones; pruned to use 20% of eigentriphones[2]	91.43
	+ Copying Gaussian covariances from the tied-state triphones (baseline2)	92.44
	+ further re-estimation of Gaussian covariances, mixture weights, and transition probabilities when the respective re-estimation thresholds are met	92.67
	Final model if 40% of eigentriphones found by weighted PCA were used	92.88*

4.4. Results and Discussions

Table 2 shows the incremental improvements obtained from the new eigentriphone acoustic modeling procedure. As we can see, the new eigentriphone modeling procedure using weighted PCA produces triphone models better than our past effort using conventional PCA as in [10] by an absolute 0.23% even if only 20% of the total number of eigentriphones are used. The improvement goes up to an absolute 0.44% if 40% of eigentriphones are used, and the improvement is statistically significant when compared with the tied-state triphones (baseline2).

4.4.1. Effect of Weighted PCA and Eigentriphone Pruning

Different numbers of eigentriphones were tried for adapting the Gaussian means. The recognition performance of the ensuing models on the Nov'93 test set using bigram and trigram language model is plotted in Fig. 1 and Fig. 2 respectively. It is observed that the performance of models derived using weighted PCA is fairly constant when more than 20% of eigentriphones are used. On the other hand, the performance of models derived from conventional PCA degrades monotonically with the use of decreasing number of eigentriphones. The phenomenon may be explained by the eigenspectra obtained from the two PCA procedures as shown in Fig. 3 for the base phoneme [aa]. From Fig. 3, one can see that weighted PCA (with the proposed weighting function) effectively captures most of the variances in the data in much fewer leading eigentriphones than conventional PCA. This allows pruning the number of eigentriphones to about 20% of them to obtain a much more compact model.

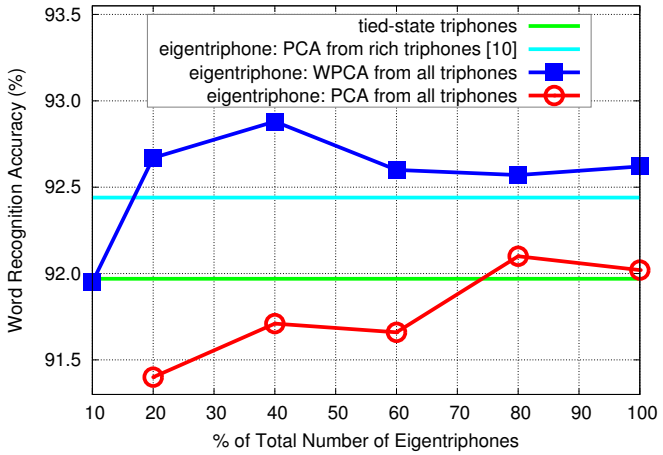

Fig. 1. Effect of weighted PCA and eigentriphone pruning using bigram LM.

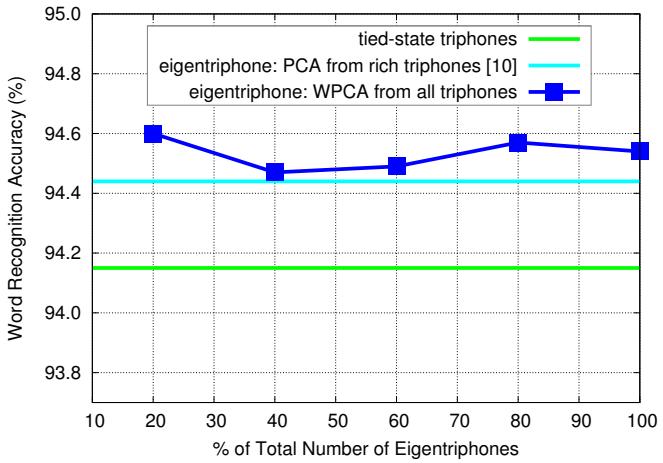

Fig. 2. Effect of weighted PCA and eigentriphone pruning using trigram LM.

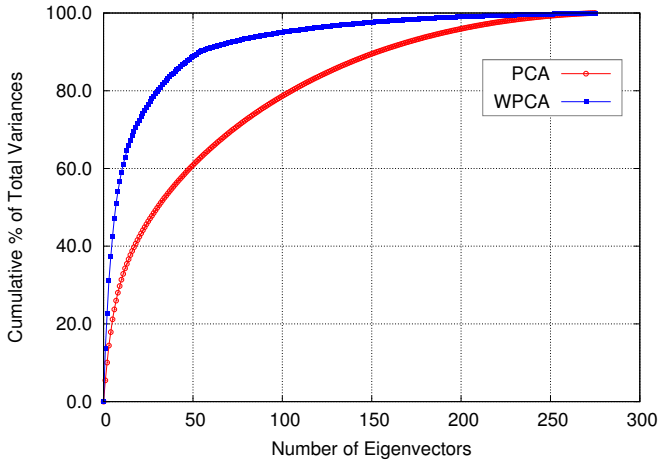

Fig. 3. Eigenspectra obtained from weighted PCA and conventional PCA for the base phoneme [aa].

5. CONCLUSIONS AND FUTURE WORK

In this paper, the eigentriphone acoustic modeling procedure is improved by using weighted PCA in deriving the eigenvectors (or eigentriphones). Each reference triphone supervector is assigned a weight that is proportional to its amount of training data so that the reliability of the triphone supervectors is taken into account in the PCA procedure. As a result, a few leading eigentriphones are sufficient to represent all the triphones, rendering the final triphone models much compact than before: for example, for the WSJ task in Section 4, the number of model parameters drop from 50.4 millions (when 100% of eigentriphones are used) to 17.6 million (when 20% of eigentriphones are used), which is about two times that of conventional tied-state triphone models[4].

Besides weighted PCA, we would like to investigate other dimension reduction procedures such as LDA or ICA.

6. REFERENCES

[1] K. F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Trans. on ASSP*, vol. 38, no. 4, pp. 599–609, April 1990.

[2] S. J. Young and P. C. Woodland, "The use of state tying in continuous speech recognition," in *Proc. of Eurospeech*, vol. 3, 1993, pp. 2203–2206.

[3] M. Hwang, "Shared distribution hidden Markov models for speech recognition," *IEEE Trans. on SAP*, vol. 1, no. 4, pp. 414–420, October 1993.

[4] E. Bocchieri and B. Mak, "Subspace distribution clustering hidden Markov model," *IEEE Trans. on SAP*, vol. 9, no. 3, pp. 264–275, March 2001.

[5] H.-A. Chang and J. R. Glass, "A back-off discriminative acoustic model for automatic speech recognizer," in *Proc. of Interspeech*, 2009, pp. 232–235.

[6] D. Povey *et al.*, "Subspace Gaussian mixture models for speech recognition," in *Proc. of ICASSP*, 2010, pp. 4330–4333.

[7] G. Saon and J.-T. Chien, "Bayesian sensing hidden Markov models," *IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[8] M. J. F. Gales and K. Yu, "Canonical state models for automatic speech recognition," in *Proc. of Interspeech*, 2010, pp. 58–61.

[9] T. Ko and B. Mak, "Eigentriphones: A basis for context-dependent acoustic modeling," in *Proc. of ICASSP*, Prague, Czech Republic, May 2011, pp. 4892–4895.

[10] ——, "A fully automated derivation of state-based eigentriphones for triphone modeling with no tied states using regularization," in *Proc. of Interspeech*, Florence, Italy, Aug 2011, pp. 781–784.

[11] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer-Verlag New York, Inc., 2002.

[12] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. on SAP*, vol. 8, no. 4, pp. 695–707, Nov 2000.

[13] S. Young *et al.*, *The HTK Book (Version 3.4)*. University of Cambridge, 2006.

[4]Note that since no states are tied, each triphone derived by our eigentriphone acoustic modeling procedure is distinctive (though some variances may be shared), the number of model parameters is expected to be larger.