Modeling Inter-cluster and Intra-cluster Discrimination Among Triphones

Tom Ko, Brian Mak and *Dongpeng Chen*

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
{tomko, mak, dpchen}@cse.ust.hk

Abstract

Discriminative training is a major contribution to the success of automatic speech recognition (ASR) in the last decade. However, since most ASR systems employ state tying which ties 'similar' states in a cluster, discriminative training may only improve inter-cluster discrimination, but states belonging to the same cluster obviously cannot be distinguished. Recently, the concept of distinct acoustic modeling was investigated by a new acoustic modeling method called *eigentriphone modeling*. In the new method, states are grouped, but not tied, into separate clusters, and the difference vectors between mean vectors of the member states and their cluster center vector are modeled by a basis approach using a set of eigenvectors which are also called eigentriphones. This paper investigates whether the inter-cluster discrimination achieved by discriminative training and intra-cluster discrimination obtained by eigentriphone modeling are additive. In a simple procedure that is applied to each state cluster, the discriminatively trained cluster center vector is integrated with the difference vectors trained by eigentriphone modeling to construct the final mean vectors of the distinct states in the cluster. Experimental evaluation on the WSJ0 5K task shows that the two techniques are indeed additive.

Index Terms— Eigentriphone, discriminative training, adaptation, regularization

1. Introduction

In context-dependent phone-based acoustic modeling, infrequent triphones need to be handled properly otherwise they will greatly affect the system performance due to the classification nature in speech recognition. Existing solutions to robust modeling of infrequent triphones can be roughly classified into three major categories: triphone-by-composition [1], parameter tying [2] and basis approach [3].

In triphone-by-composition methods, parameters of infrequent triphones are estimated through a composition of models of different order of context dependency. Model interpolation [4] and quasi-triphones [5] are typical examples of triphone-by-composition. Parameter tying methods mainly differ in their choice of acoustic units for tying. Example tying units are generalized triphones [6], state tying [7], shared distributions or senones [8], and tied subspace Gaussian distributions [9]. Among the various parameter tying methods, phonetic decision tree-based tying [10] is the most popular approach due to its proven effectiveness in balancing the trainability and resolution of the acoustic models. The key is that the infrequent triphones (and even unseen triphones) may share the same distribution with those frequent triphones in the same state cluster where the amount of training data is guaranteed. However, one potential problem is that the triphone states tied to the same cluster become identical to the recognizer, inducing quantization error in the state distributions and causing confusion between triphones during recognition.

As an alternative to the above two methods, basis approach tends to exploit the underlying relationship/factor between the context-dependent states. Examples of basis approach such as subspace Gaussian Mixture Model [11] or semi-continuous HMM [12] can be summarized by a general framework called the canonical state model [3]. It is assumed in CSM that every context-dependent state in a system can be transformed from some canonical states. These canonical states represent the underlying factor between the context-dependent states. In contrast to standard tying schemes, the model parameters are now 'related' with each other. In other words, a soft tying scheme is being used.

Recently, a new method for estimating parameters of triphone models called *eigentriphone modeling* [13, 14, 15] is proposed. In the most general form of eigentriphone modeling, models are first grouped into clusters, then an orthogonal basis is constructed from a set of well-trained reference models in the cluster. Each model in the cluster is now constrained to lie on the space spanned by the constructed basis, and is modeled as a linear combination of the eigenvectors of the basis. These eigenvectors, which are called *eigentriphones*, capture the most important context-dependent characteristics among the triphones. Since the number of eigentriphones is relatively small, even the infrequent models can be robustly trained using the new approach.

In [14, 15], the successful use of model-based eigentriphones and state-based eigentriphones were demonstrated. In both cases, an eigenbasis is derived for each monophone for modeling triphones in which no states are tied. Since the triphone models are distinct from each other, they are more discriminative as well. In the latest development of eigentriphone modeling [16], triphone states are grouped into clusters, from which eigentriphones are derived. The new method is called *cluster-based eigentriphone modeling*. From another perspective, eigentriphone modeling attempts to model intra-cluster discrimination — that is, discrimination among states belonging to the same state cluster — by modeling the difference vector between each (distinct) state in a cluster and its cluster center vector using a basis approach. In eigenvoice [17], the mean of reference speaker supervectors is chosen as the cluster center.

This work was supported by the Research Grants Council of the Hong Kong SAR under the grant numbers HKUST616513 and HKUST16206714.

In cluster-based eigentriphone modeling, it is empirically found that using the cluster mean supervector, which is estimated by maximum likelihood (ML) training, is better.

At the same time, discriminative training [18, 19, 20] has become the commonplace in acoustic modeling. However, since most automatic speech recognition systems employ state tying which ties 'similar' states together in a cluster, discriminative training may only improve inter-cluster discrimination, but states belonging to the same cluster obviously cannot be distinguished. In this paper, we would like to investigate whether the inter-cluster discrimination achieved by discriminative training and intra-cluster discrimination obtained by eigentriphone modeling are additive or complementary to each other. In a simple procedure that is applied to each state cluster, the discriminatively trained cluster center vector is integrated with the difference vectors trained by eigentriphone modeling to construct the final mean vectors of the distinct states in the cluster. Experimental evaluation on the WSJ0 5K task shows that the two techniques are indeed additive.

This paper is organized as follows. In Section 2, we first review the cluster-based eigentriphone acoustic modeling approach, and then describe how the procedures are modified using discriminatively trained cluster centers. That is followed by experimental evaluation in Section 3, and conclusions in Section 4.

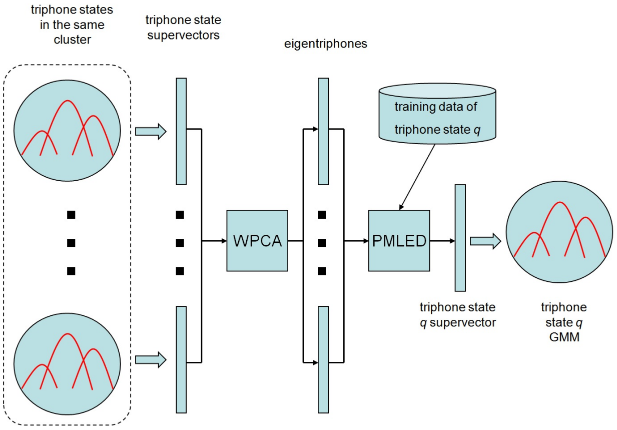
triphone states in the same cluster
triphone state supervectors
eigentriphones
training data of triphone state q
WPCA
PMLED
triphone state q supervector
triphone state q GMM

Figure 1: Overview of the cluster-based eigentriphone acoustic modeling method. (WPCA = weighted principal component analysis; PMLED = penalized maximum-likelihood eigendecomposition)

2. Cluster-based Eigentriphone Modeling

Fig. 1 shows an overview of the cluster-based eigentriphone acoustic modeling method. All triphone states are first represented by some supervectors and they are assumed to lie in a low dimensional space spanned by a set of eigenvectors. In other words, each triphone state supervector is a linear combination of a small set of eigenvectors which are now called eigentriphones.

2.1. Derivation of Cluster-based Eigentriphones

Cluster-based eigentriphone modeling consists of three major steps: (a) state clustering via a phonetic decision tree, (b) derivation of the eigenbasis, and (c) estimation of eigentriphone coefficients. The steps are summarized in further details below.

2.1.1. Conventional Tied-state Triphone HMM Training

We follow the steps in [21] to train a conventional tied-state triphone acoustic model. A tied-state acoustic model Λ^{ML} is obtained through maximum-likelihood (ML) training. Each tied state is represented by a J-component Gaussian mixture model (GMM) with diagonal covariance. For each tied state i in Λ^{ML}, create a state supervector $\mathbf{m}_i^{ML}$ by stacking up all Gaussian mean vectors in the state as below:

$$\mathbf{m}_i^{ML} = [\,\boldsymbol{\mu}_{i1}', \quad \boldsymbol{\mu}_{i2}', \quad \cdots, \quad \boldsymbol{\mu}_{iJ}']' \quad , \tag{1}$$

where $\boldsymbol{\mu}_{ij}$, $j = 1, 2, \ldots, J$ is the mean vector of the jth Gaussian component of the ith tied state.

Now the tied states are treated as state clusters.

2.1.2. Derivation of Cluster-based Eigentriphones

The following procedure is repeated for each state cluster i using its N_i triphone states that appear in the training corpus.

STEP 1: Untie the Gaussian means of all the triphone states in the cluster except the unseen triphone states. The means of the cluster GMM are then cloned to initialize *all* the untied triphone states. Note that the Gaussian variances and mixture weights of states in the cluster are still tied together.

STEP 2: Re-estimate only the Gaussian means of *all* triphone states after cloning; their Gaussian covariances and mixture weights remain unchanged as those of their cluster GMM.

STEP 3: Create a triphone state supervector $\mathbf{v}_{ip}$ for each triphone state p in cluster i by stacking up all its Gaussian mean vectors from its J-component GMM as in Eqn. 1.

STEP 4: Collect the state mean supervectors $\mathbf{v}_{i1}$, $\mathbf{v}_{i2}$, ..., $\mathbf{v}_{iN_i}$ as well as the ML-trained cluster center supervector $\mathbf{m}_i^{ML}$ of cluster i, and derive an eigenbasis from their correlation matrix using *weighted principal component analysis* (WPCA). The correlation matrix is computed as follows:

$$\frac{1}{F_i}\sum_p F_{ip}(\hat{\mathbf{v}}_{ip} - \mathbf{m}_i^{ML})(\hat{\mathbf{v}}_{ip} - \mathbf{m}_i^{ML})' \;, \tag{2}$$

where $\hat{\mathbf{v}}_{ip}$ are the standardized version of $\mathbf{v}_{ip}$ after it is normalized by its variances; F_{ip} is the frame count of the triphone state p in cluster i, and $F_i = \sum_p F_{ip}$. Note that we empirically find that using $\mathbf{m}_i^{ML}$ as the bias for correlation computation gives a better result than the arithmetic mean of the state supervectors $\{\hat{\mathbf{v}}_{ip}, p = 1, \ldots, N_i\}$.

STEP 5: Arrange the eigenvectors $\{\,\mathbf{e}_{ik},\ k = 1, 2, \ldots, N_i\,\}$ in descending order of their eigenvalues λ_{ik}, and pick the top K_i (where $K_i < N_i$) eigenvectors to represent the eigenbasis of cluster i. These K_i eigenvectors are now called *eigentriphones* of cluster i. Note that, in general, different clusters have a different number of eigentriphones.

2.1.3. Estimation of the Eigentriphone Coefficients

After the derivation of the eigentriphones, the supervector $\mathbf{v}_{ip}$ of any triphone state p in cluster i is assumed to lie in the space spanned by the K_i eigentriphones. Thus, we have

$$\mathbf{v}_{ip} = \underbrace{\mathbf{m}_i^{ML}}_{\text{cluster center}} + \underbrace{\mathbf{E}_i\mathbf{w}_{ip}}_{\text{difference vector}} \quad , \tag{3}$$

where $\mathbf{E}_i = [\mathbf{e}_{i1}, \ldots, \mathbf{e}_{iK_i}]$ is the matrix of the eigentriphones that is used to model the intra-cluster discrimination among the member states cluster i, and $\mathbf{w}_{ip} = [w_{ip1}, \ldots, w_{ipK_i}]'$ is the eigentriphone coefficient vector of triphone state p in the cluster. The second term $\mathbf{E}_i\mathbf{w}_{ip}$ models the difference vector between the cluster center and each distinct state in the cluster.

The eigentriphone coefficient vector $\mathbf{w}_{ip}$ is estimated by maximizing the objective function $Q(\mathbf{w}_{ip})$ in the *penalized maximum-likelihood eigen-decomposition* (PMLED) [14] as follows

$$Q(\mathbf{w}_{ip}) = L(\mathbf{w}_{ip}) - \beta \sum_{k=1}^{K_i} \frac{w_{ipk}^2}{\lambda_{ik}} \,, \qquad (4)$$

where $L(\mathbf{w}_{ip})$ is the log-likelihood of the training data; β is the regularization parameter; $\mathbf{w}_{ipk}$ is the coefficient for the kth eigentriphone.

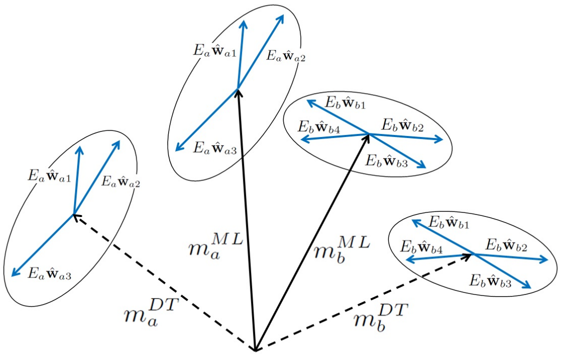
$E_a\hat{\mathbf{w}}_{a1}$
$E_a\hat{\mathbf{w}}_{a2}$
$E_a\hat{\mathbf{w}}_{a3}$
$E_b\hat{\mathbf{w}}_{b1}$
$E_b\hat{\mathbf{w}}_{b4}$
$E_b\hat{\mathbf{w}}_{b2}$
$E_b\hat{\mathbf{w}}_{b3}$
$E_a\hat{\mathbf{w}}_{a1}$
$E_a\hat{\mathbf{w}}_{a2}$
$E_a\hat{\mathbf{w}}_{a3}$
m_a^{ML}
m_b^{ML}
$E_b\hat{\mathbf{w}}_{b1}$
$E_b\hat{\mathbf{w}}_{b4}$
$E_b\hat{\mathbf{w}}_{b2}$
$E_b\hat{\mathbf{w}}_{b3}$
m_a^{DT}
m_b^{DT}

Figure 2: An illustration of the inter-cluster and intra-cluster discriminations provided by discriminative training and eigentriphone modeling respectively. m_a^{ML} and m_b^{ML} are the centers of clusters a and b obtained through ML training; m_a^{DT} and m_b^{DT} are the centers of clusters a and b obtained through discriminative training.

2.2. Investigation Issue: Cluster-based eigentriphones with Discriminatively Trained Bias

Compared with a conventional tied-state system, the discrimination among triphone states within the same state cluster — or the intra-cluster disrimination —- is now modeled by an addition of the difference vectors $\mathbf{E}_i\mathbf{w}_{ip}$ from the cluster centers in the acoustic space (Fig. 2). Meanwhile, the discrimination between state clusters is given by the ML-trained cluster centers $\mathbf{m}_i^{ML}$. The inter-state-cluster discrimination can be readily enhanced by discriminative training (DT). Let us denote the corresponding DT-trained cluster centers by $\mathbf{m}_i^{DT}$. Thus, we have the following two pieces of additional discrimination information:

additional *inter-cluster* discrimination:	$\mathbf{m}_i^{DT} - \mathbf{m}_i^{ML}$
intra-cluster discrimination:	$\mathbf{E}_i\mathbf{w}_{ip}$

This paper investigates integrating these two pieces of complementary discrimination information and models the supervector of the distinct triphone state p of cluster i as follows

$$\mathbf{v}_{ip} = \mathbf{m}_i^{ML} + \underbrace{\left(\mathbf{m}_i^{DT} - \mathbf{m}_i^{ML}\right)}_{\text{additional inter-cluster discrimination}} + \underbrace{\mathbf{E}_i\mathbf{w}_{ip}}_{\text{intra-cluster discrimination}} \,. \qquad (5)$$

Table 1: Information of various WSJ data sets.

Data Set	#Speakers	#Utterances	Vocab Size
SI-84	83	7,138	8,911
SI-284	283	37,413	13,646
dev. set	10	410	1,591
Nov'92	8	330	1,270

3. Experimental Evaluation

3.1. Speech Corpora and Experimental Setup

Two sets of experiments were conducted on the Wall Street Journal (WSJ) continuous speech recognition: one using the smaller SI-84 WSJ0 training set and another one using the larger SI-284 WSJ0+1 training set. The SI-84 training set consists of 15 hours of 7,138 WSJ0 read utterances from 83 speakers. The SI-284 training set is a superset of the SI-84 training set, consisting of all the WSJ0 utterances plus an addition of 30,275 WSJ1 utterances from 200 speakers for a total of about 70 hours of read speech. All the training data were endpointed. The standard Nov'92 5K non-verbalized test set was used for evaluation while the 1992 WSJ 5K development data set was used for tuning the system parameters. These data sets are summarized in Table 1. The language models in the experiments were the standard 5K-vocabulary bigram and trigram that came along with the WSJ corpus which have a perplexity of 147 and 57 respectively.

There were altogether 15,061 cross-word triphones in WSJ0 training set and 18,777 cross-word triphones in WSJ0+1 training set 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 $J = 16$ 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. The HTK toolkit [21] was used for maximum likelihood HMM estimation and discriminative training as well as speech decoding.

3.2. Acoustic Modeling

The performance (in term of word accuracy) of the following four acoustic modeling methods are compared on the WSJ 5K recognition tasks:

- baseline1: conventional ML training of tied-state triphone HMMs.
- baseline2: minimum-phone-error (MPE) discriminative training of tied-state triphone HMMs resulted from baseline1.
- cluster-based eigentriphone modeling of triphone HMMs applied after baseline1 (and no states are tied).
- cluster-based eigentriphone modeling of triphone HMMs applied after baseline1 with discriminatively trained cluster centers extracted from the models of baseline2 (again no states are tied).

The SI-84 tied-state baselines consist of 1,277 tied-states and the SI-284 tied-state baselines consist of 7,374 tied-states. For simplicity, the cluster-based eigentriphone modeling was conducted using the clusters defined by the tied states in the baseline systems. In general, the optimal choice of state clusters

Table 2: Recognition word accuracy (%) of various systems on the WSJ Nov'92 5K evaluation set using bigram or trigram LM.

Train	Model Description	Bigram	Trigram
SI-84	Baseline1: ML-trained tied-state triphones	93.09	95.46
	Baseline2: MPE-trained tied-state triphones	93.46 (+0.37)	95.78 (+0.32)
	Cluster-based eigentriphone modeling	93.89 (+0.80)	95.74 (+0.28)
	Cluster-based eigentriphone modeling with discriminatively trained cluster centers	93.98 (+0.89)	95.95 (+0.49)
SI-284	Baseline1: ML-trained tied-state triphones	94.25	96.32
	Baseline2: MPE-trained tied-state triphones	94.28 (+0.03)	96.54 (+0.22)
	Cluster-based eigentriphone modeling	94.30 (+0.05)	96.53 (+0.21)
	Cluster-based eigentriphone modeling with discriminatively trained cluster centers	94.64 (+0.39)	96.73 (+0.41)

for eigentriphone modeling can be different from the tied states chosen by conventional tied-state HMM even though they come from the same state tying tree. The dimension of triphone state supervectors is 16 (mixtures) x 39 (MFCC) = 624. For each state cluster, all seen triphone states were used to derive the eigentriphones, then the top 20% of eigentriphones were used in PMLED. The regularization parameter β in PMLED was set to 10.

Table 3: Relative amount of triphones in the Nov'92 test set that are considered infrequent in the SI-84 or SI-284 training set for different definition of infrequency.

Sample Count Below	SI-84	SI-284
10	5.37	0.82
20	11.2	1.75
30	15.7	2.55
40	19.5	3.55
50	23.5	4.53

3.3. Results and Discussion

Word recognition results of various systems are compared in Table 2. First of all, we can see that the previously proposed cluster-based eigentriphone modeling performs at least as well as the discriminatively trained tied-state triphones. With a trigram LM, both of them give an absolute 0.2% – 0.3% (or, a relative 5.4% – 6.6%) reduction in the word error rates (WERs) when compared with conventional ML training of tied-state triphones. This suggests that the exploitation of intra-cluster discrimination between member states of a state cluster (obtained by eigentriphone modeling) may be as important as the additional inter-cluster discrimination obtained by discriminative training. Moreover, the performance gain by eigentriphone modeling alone is more prominent with the smaller training set of SI-84 than the larger training set SI-284. This shows that eigentriphone modeling is particular effective with sparse training data.

The last row for each of the tasks shows the recognition performance of the models after integrating the two approaches, and it gives the best performance among the four modeling methods: absolute reduction of WER by 0.89% in WSJ0 and 0.39% in WSJ0+1 when bigram LM was used, and 0.49% in WSJ0 and 0.41% in WSJ0+1 when trigram LM was used. Thus, it seems that eigentriphone modeling and discriminative training are complementary to each other, and the improvement given by each of them are additive.

As the strength of the eigentriphone modeling method is its ability to construct distinct models robustly for the infrequent triphones, we hypothesize that the performance gain in a task will depend on how often those triphones that are infrequent in the training set appear in the test set. Thus, we count the relative amount of infrequent triphones in the two training sets that appear in the test set for different definitions of infrequency, and summarize the findings in Table 3. It can be seen that many more triphones in the Nov'92 test set appear infrequently in WSJ0 than in WSJ0+1. This is expected as the training set of WSJ0+1 is about 4 times bigger than the training set of WSJ0, and the latter is actually a subset of the former. Thus, the benefit of eigentriphone modeling is more pronounced in the WSJ0 task than in the WSJ0+1 task.

4. Conclusions and Relation to Prior Work

This paper successfully shows that the cluster-based eigentriphone modeling [13, 14, 15, 16] can be further improved by replacing the ML-trained cluster centers by the discriminatively trained centers. Standard discriminative training [18, 19, 20] of tied-state triphones aims at maximizing the inter-cluster discrimination among tied states, whereas the cluster-based eigentriphone modeling eliminates the quantization errors in tied states by untying the states belonging to the same tied state. Besides untying states, eigentriphone modeling further models each distinct member state of each state cluster (formerly a tied state) by a difference vector from the cluster center, thus effectively achieving additional discrimination among the member states. The two approaches are integrated together in this paper so that both inter- and intra-cluster discriminations are modeled in the new cluster-based eigentriphone modeling algorithm that uses discriminatively trained cluster centers.

Experimental evaluation on WSJ 5K task shows that the new algorithm may combine the gains achieved by each of discriminative training and eigentriphone modeling, and gives the best recognition performance.

5. References

[1] Owens M. Ji Ming, O'Boyle P. and Smith F. J., "A Bayesian approach for building triphone models for continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 678–684, 1999.

[2] S. Takahashi and S. Sagayama, "Four-level tied-structure for efficient representation of acoustic modeling," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1995.

[3] M. J. F. Gales and K. Yu, "Canonical state models for automatic speech recognition," in *Proceedings of Interspeech*, 2010.

[4] K. F. Lee, *The Development of the SPHINX System*, Kluwer Academic Publishers, 1989.

[5] A. Ljolje, "High accuracy phone recognition using context clustering and quasitriphonic models," *Computer Speech and Language*, vol. 8, pp. 129–151, 1994.

[6] K. F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 38, pp. 599–609, 1990.

[7] S. J. Young and P. C. Woodland, "The use of state tying in continuous speech recognition," in *Proceedings of the European Conference on Speech Communication and Technology*, 1993.

[8] M. Y. Hwang and X. D. Huang, "Shared-distribution hidden Markov model for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, pp. 414–420, 1993.

[9] E. Bocchieri and B. Mak, "Subspace distribution clustering hidden Markov model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 264–275, 2001.

[10] J. J. Odell S. J. Young and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the Workshop on Human Language Technology*, 1994.

[11] D. Povey et al., "Subspace Gaussian mixture models for speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[12] X. D. Huang and M. A. Jack, "Semi-continuous hidden Markov models for speech signals," *Computer Speech and Language*, vol. 3, pp. 239–251, 1989.

[13] T. Ko and B. Mak, "Eigentriphones: A basis for context-dependent acoustic modeling," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2011.

[14] T. Ko and B. Mak, "A fully automated derivation of state-based eigentriphones for triphone modeling with no tied states using regularization," in *Proceedings of Interspeech*, 2011.

[15] T. Ko and B. Mak, "Derivation of eigentriphones by weighted principal component analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[16] T. Ko and B. Mak, "Eigentriphones for context-dependent acoustic modeling," *IEEE Transactions on Audio, Speech and Language Processing*, submitted.

[17] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 695–707, 2000.

[18] B. H. Juang and S. Katagiri, "Discriminative training for minimum error classification," *IEEE Transaction on Signal Processing*, vol. 40, no. 12, pp. 3043–3054, Dec 1992.

[19] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 25–47, Jan 2002.

[20] D. Povey and P.C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, may 2002, vol. 1, pp. I–105 –I–108.

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