DISTINCT TRIPHONE MODELING BY REFERENCE MODEL WEIGHTING

Dongpeng Chen and Brian Mak

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

ABSTRACT

State tying effectively strikes a balance between detailed modeling and robust parameter estimation for hidden Markov models (HMMs) in automatic speech recognition. However, triphone HMMs that are tied to the same state are not distinguishable in that state. Recently we proposed the idea of distinct acoustic modeling in which no states are tied. In our novel *clustered-based eigentriphone modeling* method, triphones (or states) are grouped into non-overlapping clusters, from each of which, an orthogonal eigenbasis is derived using weighted PCA. Then all member triphones (or states) of a cluster are projected as distinct points onto the space spanned by its eigenvectors.

In this paper, we propose a new simpler training method called *reference model weighting* (RMW) which removes the requirement of an orthogonal basis in eigentriphone, and directly uses a set of reference model vectors in a cluster as the basis. All member model vectors are then constrained to lie in the space spanned by these reference model vectors. The difference between eigentriphone modeling and reference model weighting is analogous to the difference between eigenvoice and reference speaker weighting in speaker adaptation. The new RMW method shows consistently better performance than eigentriphone and the baseline tied-state HMMs in WSJ0 word recognition and TIMIT phoneme recognition.

Index Terms— eigenvoice, eigentriphone, regularization, state tying, acoustic modeling

1. INTRODUCTION

Context-dependent (CD) acoustic modeling is crucial in automatic speech recognition (ASR). Usually it is impractical to collect enough training data for all CD acoustic units. It is observed that about 20% of distinct triphones account for around 80% of all triphone occurrences [1]. That is, the distribution of training data among CD units is very uneven, and the phenomenon is observed for both small and large corpora[1]. Naive maximum-likelihood (ML) estimation would yield poor recognition performance for those CD units with sparse training data. Parameter tying [2] is a common solution for the problem. It effectively reduces the model size and improves recognition speed at the same time. Among the various model parameters that have been tied successfully, state tying [3] is the most popular approach in CD acoustic modeling. Usually states are tied through a regression class tree using phonetic knowledge, and the depth of the tree naturally controls the degree of state tying. Another direction is the basis approach, in which bases or atoms are constructed and model parameters are derived from them as a linear combination of some basis vectors or functions, or by some transformation. Semi-continuous hidden Markov model (SCHMM)[4], subspace Gaussian mixture model (SGMM) [5], and the canonical state model [6] are some successful examples.

Regardless of which parameter tying method and/or basis approach is used, states are usually tied in an ASR system. A shortcoming of state tying is that it inevitably introduces "quantization errors": triphones tied to the same state are not distinguishable in that state. Recently we proposed the idea of *distinct acoustic modeling* in which no states are tied so that the quantization errors may be avoided. To robustly train the ensuing distinct triphone models which consist of no tied states, we also proposed the eigentriphone modeling (ETM) approach. Inspired by the eigenvoice method in speaker adaptation, an orthogonal eigenbasis is derived from the triphones of a base phone in model-based ETM. The eigenvectors are obtained by weighted principle component analysis (PCA) [7], and the eigenvectors are also called *eigentriphones*. Then all the triphone hidden Markov models (HMMs) of the base phone are projected as distinct points onto the space spanned by the eigentriphones, and the projections are estimated using a regularized form of the maximum likelihood eigen-decomposition (MLED) [8]. The model-based ETM is later generalized to cluster-based ETM [9]. In cluster-based ETM, triphones states are grouped into clusters, and eigentriphones are derived from each state cluster. By controlling the number and the size of clusters according to the amount of training data, cluster-based ETM allows more flexible control over the finer acoustic details to be modeled robustly.

ETM works well and outperforms conventional tied-state HMMs. It involves two major steps: weighted PCA and penalized MLED. When there are many Gaussian mixtures in a state, the dimension of state Gaussian mean supervectors can be very large and the computation of weighted PCA is slow. In this paper, we further simplify cluster-based ETM by removing the weighted PCA step, and directly use the state Gaussian mean supervectors as the bases for each state cluster. The new training method is called *reference model weighting* (RMW). RMW is inspired by *reference speaker weighting* adaptation method [10] and *cluster adaptive training* [11] just as ETM is inspired by eigenvoice.

The rest of this paper is organized as follows. In the next section, details of the cluster-based RMW acoustic modeling method will be described. Then in Section 3, the new method is evaluated empirically in WSJ0 word recognition and TIMIT phoneme recognition. Finally we will make the conclusions in Section 4.

2. CLUSTER-BASED REFERENCE MODEL WEIGHTING OVER STATE CLUSTERS

Although in general, the cluster-based reference model weighting (RMW) method may be applied over clusters of many acoustic units such as phones or triphones, in this paper, we follow the application of cluster-based eigentriphone modeling (ETM) method over state clusters in [9] and apply it over clusters of states.

[1]Except for artificially designed speech corpora, simply collecting more data does not solve the problem because more data only means more distinct triphones which will still be unevenly distributed.

2.1. State Clustering

While there are many ways to cluster states using various distance metrics, we follow the successful use of phonetic decision tree in state tying [3] and clustered-based ETM [9] to cluster states in our cluster-based RMW. In fact, the nodes in the same state-tying tree that is commonly used for building tied-state hidden Markov models (HMMs) are candidates of our state clusters. The optimal choice of nodes will be determined empirically using a development set of speech data for a given task[2].

2.2. Basic Procedure

The training procedure of cluster-based RMW is modified from that of cluster-based ETM as follows [9].

The selected nodes in the state tying tree are considered as tied states, and a conventional tied-state triphone HMM system is trained. Each triphone model is a 3-state left-to-right HMM, and each state is an M-component Gaussian mixture model (GMM). Then the selected nodes are treated as state clusters, and the following procedure is repeated for each state cluster q to find the reference states and to project the remaining member states as a linear combination of the reference states.

STEP 1 : Clone the tied-state GMM to all the member states in the state cluster q.

STEP 2 : Re-estimate only the Gaussian means of the cloned triphone states in STEP 1 for those triphones which have at least 3 training samples[3]. At the same time, collect the zeroth- and first-order statistics on the training data of each Gaussian component m of state j in state cluster q — that is, its soft occupation count, $\sum_t \gamma_{qjm}(t)$, and its mean vector, $\sum_t \gamma_{qjm}(t)\mathbf{x}_t$, where $\mathbf{x}_t$ is the acoustic vector at frame t. Furthermore, the soft occupation count for each state j may be computed by summing up the occupation counts of all its mixture components as $\sum_t \sum_m \gamma_{qjm}(t)$. We will call the resulting system the *untied-state HMM system*.

STEP 3 : Based on a threshold θ on the sample count, split the member states of the cluster Ω_q into two groups: the frequent state set Ω_q^F and the infrequent state set Ω_q^I.

STEP 4 : Stack up the M Gaussian means $\{\boldsymbol{\mu}_{qjm}, m = 1, \ldots, M\}$ of state j in the frequent state set Ω_q^F according to their order in the original tied-state GMM onto a Gaussian mean supervector $\mathbf{v}_{qj} \equiv [\boldsymbol{\mu}'_{qj1}, \boldsymbol{\mu}'_{qj2}, \cdots, \boldsymbol{\mu}'_{qjM}]'$. In addition, a Gaussian mean supervector $\mathbf{v}_{q0}$ is constructed similarly for the tied state which will be indexed by $j = 0$.

STEP 5 : Form the set of reference models, or more specifically, the reference state supervectors, Ω_q^R, using the mean supervectors from the tied state and the frequent states. That is, $\Omega_q^R \equiv \{\mathbf{v}_{qj} : j = 0 \cup j \in \Omega_q^F\}$.

STEP 6 : Take the set of reference state supervectors Ω_q^R as a basis, and assume that all infrequent state supervectors of cluster q lie in the vector space spanned by the basis. Let $\mathbf{B}_q = [\mathbf{v}_{q0}\, \mathbf{v}_{qj_1} \cdots \mathbf{v}_{qj_{K_q}}]$ be the matrix of the basis vectors, where $j_k \in \Omega_q^F$ and $K_q = |\Omega_q^F|$ is the number of the reference models in cluster q. The Gaussian mean supervector $\mathbf{v}_{qi}$ of each infrequent state $i \in \Omega_q^I$ is modeled as

$$\mathbf{v}_{qi} = \sum_{j \in \Omega_q^R} w_{qij}\mathbf{v}_{qj} = \mathbf{B}_q \mathbf{w}_{qi} \quad (1)$$

where $\mathbf{w}_{qi} = [1\ w_{qij_1} \cdots w_{qij_{K_q}}]'$ is the (interpolation) weight vector of the infrequent state i. Note that the weight for the tied-state mean supervector $\mathbf{v}_{q0}$ is fixed to 1; $\mathbf{v}_{q0}$ is treated as a bias for the estimation of $\mathbf{v}_{qi}$.

STEP 7 : Estimate the weight vector $\mathbf{w}_{qi}$ by maximizing the following log-likelihood $L(\mathbf{w}_{qi})$ of its training data after removing all the irrelevant terms:

$$-\sum_{t,m} \gamma_{qim}(t)(\mathbf{x}_t - \boldsymbol{\mu}_{qim})' \mathbf{C}_{qm}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_{qim}) \quad (2)$$

where $\mathbf{C}_{qm}$ is the covariance matrix of the mth Gaussian component of the original tied state that corresponds to state cluster q. Substitute Eq. (1) to Eq. (2) and take its first order derivative. Setting the derivative to zero, we have

$$\sum_{t,m} \gamma_{qim}(t)\mathbf{B}'_{qm}\mathbf{C}_{qm}^{-1}(\mathbf{x}_t - \mathbf{B}_{qm}\mathbf{w}_{qi}) = 0$$

$$\Rightarrow \quad \mathbf{w}_{qi} = \left[\sum_m \left(\sum_t \gamma_{qim}(t)\right) \mathbf{B}'_{qm}\mathbf{C}_{qm}^{-1}\mathbf{B}_{qm}\right]^{-1} \left[\sum_m \left(\sum_t \gamma_{qim}(t)\mathbf{x}_t\right) \mathbf{B}'_{qm}\mathbf{C}_{qm}^{-1}\right] \quad (3)$$

where $\mathbf{B}_{qm}$ is the sub-matrix of $\mathbf{B}_q$ when only the rows corresponding to the mth Gaussian component of the reference mean supervectors are considered.

During the estimation process, only the Gaussian means of the infrequent states are re-estimated. The other HMM parameters such as the Gaussian covariances, transition probabilities, and mixture weights are not updated; that is, they are the same as the baseline tied-state HMMs.

2.3. Improved Procedure Using Regularization

In STEP 3 of the basic procedure, one has to classify a state as frequent or infrequent based on a fixed threshold θ on its sample count. Although θ may be determined empirically, such hard decision does not take into account the wide distribution of sample counts among the states. In addition, it is more logical to put more weight to reference models/states that are better trained with more data.

Hence the RMW procedure is further enhanced by

- using *all* states as reference states, and the mean vectors of *all* of them are re-estimated using Eq. (1). Thus, the hard binary decision of frequent or infrequent states is avoided.
- penalizing the likelihood function with the addition of a regularization term that varies according to the occupation counts of the states: greater penalty for states with small counts and smaller penalty for states with large counts. The regularization term is necessary, otherwise the re-estimated model will degenerate to the untied-state HMM due to the maximum likelihood principle.

The following penalized log likelihood function was tried:

$$\hat{L}(\mathbf{w}_{qi}) = L(\mathbf{w}_{qi}) - \sum_{k \in \Omega_q} \frac{\lambda}{2\sum_{t,m} \gamma_{qim}(t)} ||w_{qik}||^2 \quad (4)$$

[2]Note that although the nodes selected for conventional tied-state HMMs and cluster-based RMW come from the same phonetic decision tree, they need not be the same for the two procedures.

[3]We followed HTK's general practice and only re-estimated models that have at least 3 training samples.

where λ is the regularization parameter which has to be tuned empirically on a separate set of development data. The closed-form solution is given by

$$\mathbf{w}_{qi} = \left[\sum_m \left(\sum_t \gamma_{qim}(t)\right) \mathbf{B}'_{qm}\mathbf{C}^{-1}_{qm}\mathbf{B}_{qm} + \mathbf{R}\right]^{-1} \left[\sum_m \left(\sum_t \gamma_{qim}(t)\mathbf{x}_t\right) \mathbf{B}'_{qm}\mathbf{C}^{-1}_{qm}\right] \quad (5)$$

where $\mathbf{R}$ is a diagonal matrix, and

$$\mathbf{R} = \frac{\lambda}{\sum_{t,m}\gamma_{qim}(t)} \cdot \mathbf{I}_{|\Omega_q|\times|\Omega_q|} .$$

Table 1. Details of various WSJ0 and TIMIT data sets.

Data	WSJ0		TIMIT	
	#speaker	#utterance	#speaker	#utterance
train	83	7138	462	3696
test	8	330	24	192
dev	10	410	24	192

3. EXPERIMENTAL EVALUATIONS

In this section, cluster-based RMW was evaluated on two tasks: (1) continuous speech recognition on Wall Street Journal WSJ0 [12], and (2) phoneme recognition on TIMIT [13]. In both tasks, RMW is compared with conventional tied-state triphone hidden Markov modeling and eigentriphone acoustic modeling.

3.1. Experiment Setup

3.1.1. WSJ0 Continuous Speech Recognition

The standard WSJ0 SI-84 training set with 15 hours of speech was used for acoustic modeling. Evaluation was performed on the standard Nov92 5K non-verbalized test set, and the si_dt_05 data set was used as the development set for tuning system parameters such as the regularization parameter, decoding parameters, as well as for finding the optimal state-tying nodes and state clusters. Finally, a bigram language model (LM) with a perplexity of 147 was employed in this recognition task.

3.1.2. TIMIT Phoneme Recognition

Acoustic modeling used the standard TIMIT training set of about 3.14 hours of speech, and all systems were evaluated on the core test set with 192 utterances. Viterbi decoding used a trigram phone LM with a perplexity of 14.39 that was trained from the TIMIT training transcriptions using the SRILM language modeling toolkit. We followed the standard experimentation on TIMIT, and collapsed the original 61 phonetic labels in the corpus into a set of 48 phones for acoustic modeling. The latter were further collapsed into the standard set of 39 phonemes for error reporting. Moreover, the glottal stop [q] was ignored.

Details of the various data sets are listed in Table 1.

3.2. Acoustic Modeling

In all experiments, acoustic vectors were extracted at every 10ms over a window of 25ms. The traditional 39-dimensional MFCC vectors were used; they consist of 12 MFCCs and the normalized frame energy, and their 1st- and 2nd-order time derivatives. Conventional tied-state triphone HMM baselines were constructed for WSJ0 and TIMIT using the HTK toolkit. In both baseline systems, the number of tied states and the number of mixture per state were tuned using the respective development data set. Table 2 shows details of the two baseline systems.

Table 2. Details of the tied-state triphone HMM baselines.

Tied-state HMM Baseline	WSJ0	TIMIT
#tied states	1254	587
#mixtures/state	32	16
#physical triphones	15,337	15,547
#triphones (count ≥ 3)	11,747	7,902
Word/Phoneme Acc. on dev set	92.15%	75.08%
Word/Phoneme Acc. on test set	93.29%	72.04%

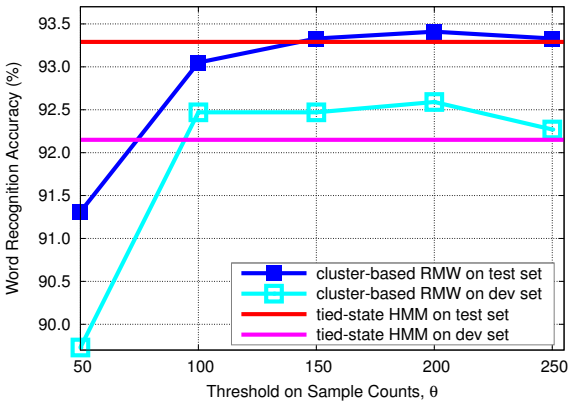

Fig. 1. WSJ0 recognition performance when the reference models are determined by thresholding on their sample counts.

3.3. Experiment 1: Reference Models Determined by Thresholding on the Occupation Counts

The basic procedure of Section 2.2 was applied to train cluster-based RMW models by varying the sample count threshold θ, and the effect on RMW's performance is shown in Fig. 1 together with the tied-state HMM baseline on both the test and development data sets. We see that some, though insignificantly small, improvement may be achieved when θ is appropriately set. However, the performance is actually worse when θ gets below 150. Note that when $\theta = 0$, the model becomes the untied-state HMM. The result is expected: when θ is small, many of the reference models are actually poorly trained, and one cannot expect good models coming out of bad reference models.

3.4. Experiment 2: Using all Models as Reference Models and Regularization

WSJ0 recognition was repeated with the improvement suggested in Section 2.3 by taking all states in a state cluster as the set of reference models, and applying the regularization of Eq. (4). The effect of the regularization parameter λ on the performance of RMW is shown in Fig. 2. Since the soft occupation count $\sum_{t,m}\gamma_{qjm}(t)$ is in terms of frame (and each frame is 10ms), its value is quite large. When $\lambda = 0$, the system again is equivalent to an untied-state HMM system. From Fig. 2, one can see that λ has to be sufficiently large to mitigate the negative impact of the poor reference models on RMW training. On the other hand, once λ is set above 200,000, the resulting models perform equally well until λ becomes too large (above 2,400,000 in this example)

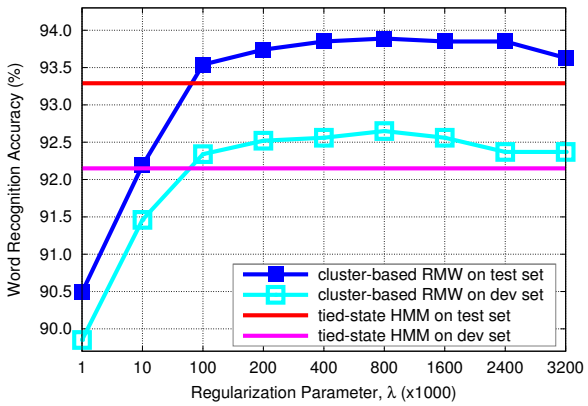

Fig. 2. Effect of regularization on WSJ0 recognition performance when all states in a cluster are adopted as the reference models.

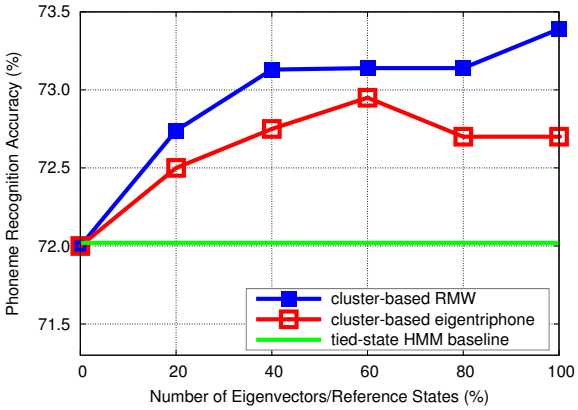

Fig. 3. Comparison between RMW and ETM when different proportions of reference states or eigentriphones are used.

and the effective number of reference models becomes too small. Most importantly, when λ is properly tuned, the resulting models significantly outperform the tied-state HMM baseline system.

3.5. Experiment 3: RMW vs. Eigentriphone with Reduced Reference Set

Cluster-based RMW simplifies cluster-based ETM by removing the construction of an orthogonal basis using weighted PCA for each state cluster. We implemented the latest cluster-based ETM in [9] and compared the two methods on TIMIT phoneme recognition. Since it has been shown that ETM works well with reduced numbers of eigenvectors[4] [7], during the comparison, we also tried to reduce the set of reference models in RMW correspondingly. Whereas ETM reduces the number of eigenvectors for modeling according to their eigenvalues, RMW reduces the number of reference models/states according to their occupation counts. Note that reducing the reference models in the regularized RMW is different from the basic RMW because (1) the former uses a penalized likelihood training criterion, and (2) all states, including the reference states (and not only infrequent states), are re-estimated in regularized RMW.

The comparison on TIMIT phoneme recognition using different proportions of reference states or eigentriphones is shown in Fig. 3. From the figure, it is observed that cluster-based ETM reaches its optimal performance when 60% of eigenvectors are kept, while cluster-based RMW gets its best performance when all states are used as reference states. However, RMW always performs better than the optimal ETM result if at least 40% of reference states are used. When the same proportion of reference states or eigentriphones is used, RMW always performs better than ETM. We believe that the regularization using occupation counts in RMW is more effective than the regularization using the eigenvalues of the selected eigentriphones in ETM.

3.6. Summary

The recognition performances of the tied-state HMM baseline, cluster-based eigentriphone modeling (ETM), and cluster-based reference model weighting method (RMW) on WSJ0 and TIMIT are summarized in Table 3. The summary shows that

- by comparing the performance of tied-state HMM and untied-state HMM, we see that state tying successfully solves the robust parameter estimation problem of triphones that have little amount of data.
- cluster-based ETM and RMW further improves the performance of tied-state HMM by reducing the quantization error in each tied state. All the recognition accuracy improvements are statistically significant.
- cluster-based RMW is even more effective than ETM, and yet its training procedure is computationally simpler and faster. At the end, RMW reduces the word error rate (WER) in WSJ0 by 0.92% absolute, and the phone error rate (PER) in TIMIT by 1.35% absolute.

[4]The main reason for reducing the number of eigenvectors to use in ETM is to reduce the final model size trained by ETM.

Table 3. Summary of results. Results with '*' are significantly better than the tied-state HMM baseline system.

Model	WSJ0	TIMIT
untied-state HMM	89.84%	68.83%
tied-state HMM	93.29%	72.04%
eigentriphones (ETM)	93.89%*	72.90%*
reference model weighting (RMW)	**94.13%***	**73.39%***
relative WER/PER reduction	12.5%	4.8%

4. CONCLUSIONS AND RELATION TO PRIOR WORK

Although state tying [3] effectively solves the robust parameter estimation problem of acoustic units that have little amount of training data, it inevitably introduces quantization errors among the states that are tied together. Eigentriphone acoustic modeling [1, 14, 7, 9] successfully eliminates the quantization errors by untying the states in tied-state HMM, and then representing the untied states as distinct points in the space spanned by an eigenbasis for each state cluster. From another perspective, one may treat eigentriphone modeling of an untied state as the eigenvoice adaptation [8] of the untied state from its tied state. Thus, one may apply other adaptation methods as well. In this paper, inspired by the advantage of reference speaker weighting [10] and cluster adaptive training [11] over eigenvoice in speaker adaptation, we propose reference model weighting (RMW). Specifically, RMW removes the need of constructing an orthogonal basis for each state cluster in eigentriphone modeling. The use of a regularizer that directly relates to the occupation counts of each reference state seems to account for its better performance over eigentriphone modeling.

5. REFERENCES

[1] Tom Ko and Brian Mak, "Eigentriphones: A basis for context-dependent acoustic modeling," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 4892–4895.

[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, vol. 1, pp. 520–523.

[3] 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, vol. 3, pp. 2203–2206.

[4] X. Huang and M. A. Jack, "Semi-continuous hidden Markov models for speech signals," *Computer Speech and Language*, vol. 3, no. 3, pp. 239–251, July 1989.

[5] Daniel Povey et al., "Subspace Gaussian mixture models for speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010, pp. 4330–4333.

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

[7] Tom Ko and Brian Mak, "Derivation of eigentriphones by weighted principal component analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, March 2012, pp. 4097–4100.

[8] 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, no. 4, pp. 695–707, Nov 2000.

[9] Tom Ko and Brian Mak, "Eigentriphones for context-dependent acoustic modeling," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 6, pp. 1285–1294, June 2013.

[10] T. J. Hazen, "A comparison of novel techniques for rapid speaker adaptation," *Speech Communications*, vol. 31, pp. 15–33, May 2000.

[11] M. F. J. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, July 2000.

[12] D. B. Paul and J. M. Baker, "The design of the Wall Street Journal-based CSR corpus," in *Proceedings of the DARPA Speech and Natural Language Workshop*, Feb. 1992.

[13] V Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, August 1990.

[14] Tom Ko and Brian Mak, "A fully automated derivation of state-based eigentriphones for triphone modeling with no tied states using regularization," in *Proceedings of Interspeech*, Florence, Italy, August 2011, pp. 781–784.