Eigentriphones for Context-dependent Acoustic Modeling

Tom Ko and Brian Mak

***Abstract*—Most automatic speech recognizers employ tied-state triphone hidden Markov models (HMM), in which the corresponding triphone states of the same base phone are tied. State tying is commonly performed with the use of a phonetic regression class tree which renders robust context-dependent modeling possible by carefully balancing the amount of training data with the degree of tying. However, tying inevitably introduces quantization error: triphones tied to the same state are not distinguishable in that state. Recently we proposed a new triphone modeling approach called *eigentriphone* modeling in which all triphone models are, in general, distinct. The idea is to create an eigenbasis for each base phone (or phone state) and all its triphones (or triphone states) are represented as distinct points in the space spanned by the basis. We have shown that triphone HMMs trained using model-based or state-based eigentriphones perform at least as well as conventional tied-state HMMs. In this paper, we further generalize the definition of eigentriphones over a cluster of acoustic units. Our experiments on TIMIT phone recognition and the Wall Street Journal 5K-vocabulary continuous speech recognition show that eigentriphones estimated from state clusters defined by the nodes in the same phonetic regression class tree used in state tying result in further performance gain.**

***Index Terms*—Eigentriphone, tied state, context dependency, regularization, weighted PCA.**

I. Introduction

A critical issue in context-dependent (CD) acoustic modeling is how to robustly estimate the model parameters of the rarely occurring acoustic units. For instance, it is found that the distribution of triphones in the HUB2 training set of the Wall Street Journal corpus [1] obeys the Pareto Principle (or the 80/20 Rule) [2]: roughly 80% of triphone occurrences in the corpus come from 20% of all the distinct triphones in the corpus [3]. Naive maximum-likelihood (ML) estimation of the hidden Markov model (HMM) parameters of these infrequent context-dependent acoustic units will produce poor triphone models, which will affect the overall performance of an automatic speech recognition (ASR) system. Past solutions for robust estimation of CD acoustic models may be roughly classified into three categories: triphone-by-composition [4], parameter tying [5], and a basis approach.

Model interpolation [6] and quasi-triphones [7] are typical examples of the triphone-by-composition method. In both examples, CD models are constructed by combining triphone models that may not be well trained with robustly trained acoustic models that capture weaker contextual information. For instance, in [6], a triphone state distribution is generated by a linear combination of its ML estimate and the state distributions from its corresponding left-context-dependent model, right-context-dependent model, and/or context-independent model using deleted interpolation. In [7], it is assumed that the left context of a phone influences mostly its beginning whereas its right context influences mostly its ending. Thus, a three-state triphone model is generated in such a way that the first and the last states are conditioned only on its left and right contexts respectively, whereas the middle state is context-independent. Recently, another example of triphone-by-composition called back-off acoustic modeling [8] was proposed. The new method combines the score of a triphone with scores from triphones that are estimated under broad phonetic class contexts of its left and right phones.

Parameter tying is another solution that is widely used in ASR systems because of its proven effectiveness in simultaneously reducing model size and enhancing recognition speed. Various HMM parameters have been tied successfully, for example, generalized triphones [9], tied states [10], shared distributions or senones [11], and tied subspace Gaussian distributions [12]. Among these parameter tying methods, state tying [10] is probably the most popular approach in context-dependent acoustic modeling. The degree of state tying — that is, the number of tied states — can be well managed by a (binary) regression class tree, using questions that are based on acoustics [13] or phonetic knowledge [14]. The use of a phonetic regression class tree offers the additional benefit of synthesizing unseen triphones in the test lexicon.

Recently, a basis approach is emerging. In the basis approach, one or more bases are constructed so that model parameters may be derived from the basis vectors or functions. For example, semi-continuous hidden Markov model (SCHMM) [15], [16] and subspace Gaussian mixture model (SGMM) [17] both employ a basis of Gaussians, whereas Bayesian sensing HMM [18] uses sets of state-dependent basis vectors. Similarly, in the canonical state model (CSM) [19] framework, there exists a finite set of canonical states from which every context-dependent state in an ASR system is transformed. The set of canonical states captures the relationship between the context-dependent states through some transformation functions. It has been shown that both SCHMM and SGMM can be derived from the CSM framework.

A common thread among all the three approaches is a set of elementary structures from which all the context-dependent models are built by linear interpolation, synthesis, or transformation. They are the models of various order of context dependency in the triphone-by-composition method; the common Gaussian pool in SCHMM or common subspace Gaussian pools in the subspace distribution clustering HMM in

All authors are with the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong. E-mail: {tomko, mak}@cse.ust.hk.

the parameter tying method; the canonical states in the CSM method. In the parameter tying and CSM methods, the use of the elementary structures helps factorize the whole set of acoustic models, resulting in a more compact and succinct representation of the models so that they may be estimated more robustly. However, parameter tying inevitably results in quantization error for the tied unit. As a consequence, models sharing the same tied unit are not distinguishable in that part of the models; in the extreme case when the respective parts of two acoustic models are tied, the two models become identical. In contrast, although the triphone-by-composition method is more complicated and needs to maintain and evaluate several sets of acoustic models, it has the advantage that, since the interpolation weights are usually different for each context-dependent model, context-dependent models created by the method are distinct from each other. Nevertheless, the three methods are complementary and they can be integrated together in a recognizer.

In [3], [20], [21], we proposed a new context-dependent acoustic modeling method called *eigentriphone modeling*. In our method, triphone models of a base phone are factorized into a set of eigenvectors, which we call *eigentriphones*, and all triphone HMMs of that base phone are then projected onto the space spanned by the eigentriphones. Eigentriphones extract the most important context-dependent characteristics among the triphones so that the infrequent triphones can be robustly modeled in terms of these eigentriphones even with few training samples. Unlike the triphone-by-composition method, eigentriphone modeling is not required to maintain several sets of acoustic models of different orders of context dependency. Eigentriphone modeling may be used together with state tying, though we prefer not to so that the ensuing triphone models are distinct from each other. Although we call our method *eigentriphone modeling*, the method may be readily applied to creating other context-dependent acoustic units such as biphones or quinphones.

Besides summarizing the development of eigentriphones from our past works, this paper further generalizes the derivation of eigentriphones from *all* triphones of each base phone to *any* triphone or state clusters. We call the new derivation method *cluster-based eigentriphone modeling*. By changing the definition of triphone or state clusters, one may balance the amount of available training data with the resolution of eigentriphones. In particular, we propose to derive eigentriphones from the state clusters defined by the tied states in a phonetic regression class tree so that the quantization error due to conventional state tying is avoided, and the benefit of synthesizing unseen triphones by the phonetic regression tree can be incorporated into the eigentriphone modeling method.

The paper is organized as follows. In Section II, we will describe the model-based eigentriphone acoustic modeling approach. Then we will extend the model-based eigentriphone to state-based eigentriphone in Section III, and then to cluster-based eigentriphone in Section IV. It is followed by experimental evaluation in Section V and conclusions in Section VI.

TABLE I
COMPARISON BETWEEN EIGENVOICE AND MODEL-BASED EIGENTRIPHONE MODELING.

Item	Eigenvoice	Eigentriphone
number of bases	1	number of monophones
model to fallback	speaker-independent model	context-independent model
reference models	speaker-dependent models	all triphones of the base phone
adapted models	new speaker models	all triphones of the base phone

II. Model-based Eigentriphone

The eigentriphone acoustic modeling method belongs to the basis approach and is inspired by the eigenvoice adaptation [22]. The acoustic modeling of triphones with limited amount of training data may be thought of as an adaptation problem which is then solved by the eigenvoice approach. That is, all triphone models are first represented by some supervectors and they are assumed to lie in a low dimensional space[1] spanned by a set of eigenvectors. In other words, each triphone supervector is a linear combination of a small set of eigenvectors which are now called eigentriphones.

A. Eigenvoice vs. Eigentriphone

The eigenvoice adaptation and model-based eigentriphone acoustic modeling are very similar except that

- whereas there is only one set of eigenvectors in eigenvoice adaptation, each base phone requires a separate set of eigenvectors in eigentriphone modeling, and
- speaker-dependent models in eigenvoice are replaced by triphone models in eigentriphone modeling.

A comparison of the two methods is shown in Table I.

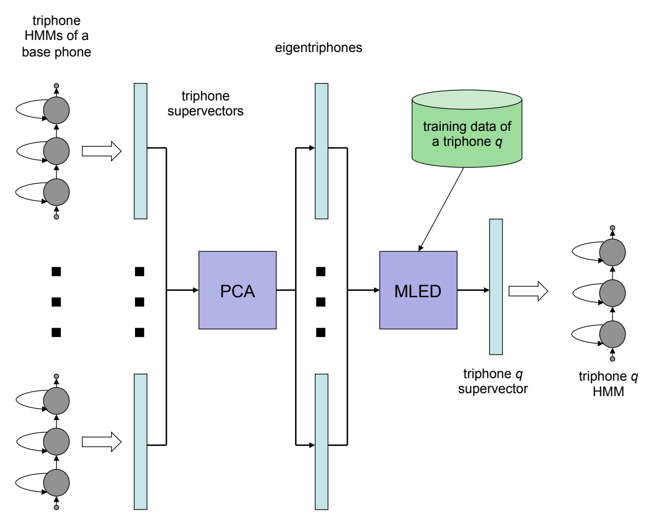

Fig. 1. The model-based eigentriphone acoustic modeling method. (PCA = principal component analysis; MLED = maximum-likelihood eigen-decomposition)

B. The Basic Procedure

Fig. 1 shows an overview of model-based eigentriphone acoustic modeling; it is similar to eigenvoice adaptation [22].

[1]The dimension of the space is low when compared with the dimension of the triphone supervectors.

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

STEP 1 : Monophone hidden Markov model (HMM) of base phone 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 of base phone i is then cloned to initialize *all* its N_i triphones in the training data. Note that (a) unlike common triphone cloning from a 1-mixture monophone HMM, in our eigentriphone procedure, triphones are cloned from an M-mixture monophone HMM, and (b) no state tying is performed.

STEP 3 : Re-estimate only the Gaussian means of triphones after cloning; their Gaussian covariances and mixture weights (which are copied from their base phone HMM) remain unchanged.

STEP 4 : Create a triphone supervector $\mathbf{v}_{ip}$ for each triphone p of base phone i by stacking up all its Gaussian mean vectors from its three states as below.

$$\mathbf{v}_{ip} = \begin{array}{l} [\ \boldsymbol{\mu}_{ip11}, \quad \boldsymbol{\mu}_{ip12}, \quad \cdots, \quad \boldsymbol{\mu}_{ip1M}, \\ \ \ \boldsymbol{\mu}_{ip21}, \quad \boldsymbol{\mu}_{ip22}, \quad \cdots, \quad \boldsymbol{\mu}_{ip2M}, \\ \ \ \boldsymbol{\mu}_{ip31}, \quad \boldsymbol{\mu}_{ip32}, \quad \cdots, \quad \boldsymbol{\mu}_{ip3M}\ ] \end{array}, \tag{1}$$

where $\boldsymbol{\mu}_{ipjm}, j = 1, 2, 3$, and $m = 1, 2, \ldots, M$ is the mean vector of the mth Gaussian component at the jth state of triphone p. Similarly, a monophone supervector $\mathbf{m}_i$ is created from the monophone model of the base phone i.

STEP 5 : Collect all triphone supervectors $\mathbf{v}_{i1}$, $\mathbf{v}_{i2}$, ..., $\mathbf{v}_{iN_i}$ as well as the monophone supervector $\mathbf{m}_i$ of base phone i, and derive an eigenbasis from their correlation or covariance matrix using *principal component analysis* (PCA). The covariance matrix is computed as follows:

$$\frac{1}{N_i} \sum_p (\mathbf{v}_{ip} - \mathbf{m}_i)(\mathbf{v}_{ip} - \mathbf{m}_i)' \ . \tag{2}$$

Notice that the monophone supervector $\mathbf{m}_i$, instead of the mean of triphone supervectors, is used to "center" triphone supervectors so that the poor triphones may fall back to the monophone HMM in the worst case[2].

STEP 6 : 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 $K_i \leq N_i$) eigenvectors to represent the eigenspace of base phone i. These K_i eigenvectors are now called *eigentriphones* of phone i. In general, different base phones have a different number of eigentriphones, depending on the criterion used to decide the value of K_i.

STEP 7 : Now the supervector $\mathbf{v}_{ip}$ of any triphone p of base phone i is assumed to lie in the space spanned by the K_i eigentriphones. Thus, we have

$$\mathbf{v}_{ip} = \mathbf{m}_i + \sum_{k=1}^{K_i} w_{ipk} \mathbf{e}_{ik} \ , \tag{3}$$

where $\mathbf{w}_{ip} = [w_{ip1}, w_{ip2}, \ldots, w_{ipK_i}]$ is the eigentriphone coefficient vector of triphone p in the "triphone space" of base phone i.

STEP 8 : Estimate the eigentriphone coefficient vector $\mathbf{w}_{ip}$ of any triphone p by maximizing the likelihood $L(\mathbf{w}_{ip})$ of its training data:

$$\begin{aligned} L(\mathbf{w}_{ip}) &= \text{constant} - \\ &\sum_{j,m,t} \gamma_{ipjm}(t)(\mathbf{x}_t - \boldsymbol{\mu}_{ipjm}(\mathbf{w}_{ip}))' C_{ipjm}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_{ipjm}(\mathbf{w}_{ip})) \end{aligned} \tag{4}$$

where C_{ipjm} and $\gamma_{ipjm}(t)$ are the covariance and occupation probability of the mth Gaussian at the jth state of triphone p of base phone i given observation $\mathbf{x}_t$. The procedure is called *maximum-likelihood eigen-decomposition* (MLED) in [22]. Finally, the Gaussian mean of the mth mixture at the jth state of triphone p can be obtained from $\mathbf{v}_{ip}$ as

$$\boldsymbol{\mu}_{ipjm} = \mathbf{m}_{ijm} + \sum_{k=1}^{K_i} w_{ipk} \mathbf{e}_{ikjm} \ . \tag{5}$$

STEP 9 : If either the eigentriphone coefficients converge or the recognition accuracy of a development data set is maximized, go to STEP 10. Otherwise, re-align the training data using the model in STEP 8, re-estimate the Gaussian means and repeat STEP 4 – 9.

STEP 10 : After the eigentriphone "adaptation" of the Gaussian means, the Gaussian covariances and mixture weights of a triphone are re-estimated if its sample count is greater than the thresholds θ_v and θ_w respectively. Otherwise, they remain the same as those of the monophone model from which they are cloned.

The above basic procedure works but there are rooms to improve in at least two aspects:

- In STEP 5, all triphones are used to derive the eigenbasis and, thus, the eigentriphones. However, it is clear that due to uneven distribution of triphones in the training data, some triphones will be better trained than the others in STEP 3. Including the poorly trained triphone models in the subsequent PCA will affect the quality of the eigentriphones. One heuristic solution is to use only those triphones with sufficient training data, but how much is sufficient?
- In STEP 6, one has to make a hard decision on the dimension of the eigenspace (or, equivalently, the number of eigentriphones), K_i, to represent all the triphone models. A common practice is to pick a number of eigenvectors so that a certain percentage of the total variations is covered. May we not be forced to make a hard decision on the value of K_i?

In [23] and [24], we proposed using weighted PCA and regularization to solve the two problems respectively.

C. Improvement #1: Derivation Using Weighted PCA

The use of *weighted PCA* instead of the standard PCA has at least two advantages.

Firstly, to avoid making a hard decision on which triphones to use in the application of standard PCA. Instead, the use

[2]Empirically, we find that centering by the monophone supervector gives slightly better performance than if the mean of triphone supervectors is used.

of weighted PCA allows eigentriphones to be derived by taking *all* triphones into account. This is made possible by incorporating some notion of triphone reliability that is related to its training data sufficiency in the construction of the eigenbasis in weighted PCA.

Secondly, as we had shown in [21], the eigenvalue spectrum produced by weighted PCA rises more sharply than the spectrum given by standard PCA. It means that fewer leading eigentriphones produced by weighted PCA can capture more variations in the triphone supervectors. As a result, weighted PCA allows the use of fewer eigentriphones in eigentriphone acoustic modeling. This has an implication in the space requirement of eigentriphone acoustic modeling. Since each triphone model produced by eigentriphone modeling is distinct, all observed triphones — even those with few samples — in the database will be represented by a distinct HMM. Consequently, the model size resulted from eigentriphone modeling is much bigger than conventional tied-state HMMs. With the use of weighted PCA, fewer eigentriphones may be employed to model each triphone, and the model size can be reduced drastically by more than 50%.

In this paper, each triphone supervector is weighted by its sample count in the weighted PCA procedure. Thus, the covariance matrix in STEP 5 is replaced by

$$\frac{1}{N_i}\sum_p N_{ip}(\mathbf{v}_{ip}-\mathbf{m}_i)(\mathbf{v}_{ip}-\mathbf{m}_i)' \ , \tag{6}$$

where N_{ip} is the sample count of the triphone p of base phone i, and $N_i = \sum_p N_{ip}$.

D. Improvement #2: Soft Decision on the Number of Eigentriphones using Regularization

To avoid making a hard decision on the number of eigentriphones K_i to use, a new penalized log-likelihood function $Q(\mathbf{w}_{ip})$ is defined for the estimation of the eigentriphone coefficient vector $\mathbf{w}_{ip}$ using *all* eigentriphones of base phone i:

$$Q(\mathbf{w}_{ip}) = L(\mathbf{w}_{ip}) - \beta R(\mathbf{w}_{ip}) \ , \tag{7}$$

where β is the regularization parameter that controls the relative importance of the regularizer $R(\cdot)$ compared with the likelihood term $L(\cdot)$ of Eqn. (4). The regularizer should be chosen so that the more informative eigentriphones (with larger eigenvalues) are automatically emphasized and the less informative eigentriphones (with smaller eigenvalues) are automatically de-emphasized. In [24], the following regularizer is found effective

$$R(\mathbf{w}_{ip}) = \sum_{k=1}^{N_i}\frac{w_{ipk}^2}{\lambda_{ik}} \ . \tag{8}$$

The proposed regularizer represents a scaled Euclidean distance of the triphone from the base phone in the space spanned by the eigentriphones. It has the following properties:

- The squared coefficient of each eigentriphone, w_{ipk}, is inversely scaled by its eigenvalue so that a less informative eigentriphone will have less influence on the "adapted" triphone model.
- When there are a lot of training data, the likelihood term will dominate the objective function $Q(\mathbf{w}_{ip})$, and the "adapted" triphone model will converge to its conventional Baum-Welch training estimate.
- On the other hand, for a triphone with limited amount of training data, the penalty term will dominate and a smaller scaled Euclidean distance between the triphone and base phone is preferred. In other words, its "adapted" triphone model will fallback to its monophone model.

Thus, in effect, the regularizer of Eqn. (8) will provide a soft decision on the number of eigentriphones to use for each triphone (and not just for each base phone).

Differentiating the optimization function $Q(\mathbf{w}_{ip})$ of Eqn. (7) w.r.t. each eigentriphone coefficient w_{ipk}, and setting each derivative to zero, we have,

$$\sum_{n=1}^{N_i} A_{ipkn}w_{ipn} + \beta\frac{w_{ipk}}{\lambda_{ik}} = B_{ipk} \quad \forall k = 1,2,\cdots N_i \tag{9}$$

where

$$\begin{aligned} A_{ipkn} &= \sum_{j,m}\mathbf{e}'_{ikjm}C^{-1}_{ipjm}\mathbf{e}_{injm}\left(\sum_t \gamma_{ipjm}(t)\right) \\ B_{ipk} &= \sum_{j,m}\mathbf{e}'_{ikjm}C^{-1}_{ipjm}\left(\sum_t \gamma_{ipjm}(t)(\mathbf{x}_t-\mathbf{m}_{ijm})\right) \ . \end{aligned}$$

The eigentriphone coefficients may be easily found by solving the system of N_i linear equations represented by Eqn. (9), and the Gaussian means of the new model may be computed using Eqn. (5).

III. State-based Eigentriphone

In model-based eigentriphone acoustic modeling, high-dimensional triphone supervectors are constructed by concatenating Gaussian mean vectors from all the (three) states of each triphone HMM of a base phone. One may also apply the modeling framework to sub-phonetic units as well. In [24], *state-based eigentriphone* acoustic modeling was proposed in which an eigenbasis is developed for each state of each basis phone in a procedure similar to that of model-based eigentriphone modeling in Section II except that sample counts of triphone in Eqn.(6) are replaced by frame counts of triphone states. Compared with model-based eigentriphone acoustic modeling, state-based eigentriphone acoustic modeling produces three times more eigenbases, but its eigenvector dimension is only $1/3$ of the former.

IV. Cluster-based Eigentriphone

Both model-based and state-based eigentriphone acoustic modeling methods discussed above derive eigenbases from all triphones of a base phone. In fact, the eigentriphone modeling framework is very flexible and can be applied to any group of phonetic or sub-phonetic units provided that they may be represented by supervectors of the same dimension. For example, if training data are really scarce, one may perhaps derive eigentriphones from broad phonetic classes (such as vowels, fricatives, etc.); on the other hand, when there are sufficient data, one may divide the triphones of a base phone into groups and derive eigentriphones from each triphone group. In this paper, we would like to investigate a more general framework of deriving eigentriphones from clusters of triphones or triphone states[3], and we call this *cluster-based eigentriphone* acoustic modeling. In particular, we will investigate eigentriphone modeling with general state clusters.

Common clustering algorithms such as k-means clustering, agglomerative hierarchical clustering, and decision tree, together with a well-defined distance metric or impurity function may be used to generate triphone or state clusters for cluster-based eigentriphone modeling. Instead of delving into various clustering algorithms, we resort to the use of phonetic decision tree for the purpose since it has been applied successfully to a few tasks such as state tying in ASR. In fact, we propose to use the triphone state clusters represented by the nodes in the same state-tying tree for deriving eigentriphones. There are several benefits for the choice:

- In a typical ASR system, there are 39 base phones and triphone models are 3-state HMMs. Thus, there will be 39 sets of model-based eigentriphones and $39 \times 3 = 117$ sets of state-based eigentriphones. On the other hand, there are many more tied states — usually hundreds to thousands — in an ASR system, which means that the use of the state clusters from tied-states will allow a higher resolution of eigentriphone modeling than the above model-based or state-based eigentriphone modeling. Moreover, the state-tying tree gives one the flexibility to decide the modeling resolution by going up or down the phonetic decision tree and choose the right nodes for cluster-based eigentriphone derivation[4].
- State-based eigentriphone modeling is a special case of cluster-based eigentriphone modeling in which each cluster consists of respective states from all triphones of a base phone. However, cluster-based eigentriphone modeling using tied-state clusters is computationally more efficient because the number of tied states is usually much greater than the number of monophone states so that there are fewer triphone state supervectors in each tied-state cluster to derive eigentriphones. From Eqn. (9), it is observed that the computation of eigentriphone coefficients involves solving a system of N_i linear equations with a computational complexity of $O(N_i^3)$. When there are fewer triphone states in a cluster, the computation of eigentriphone coefficients is faster.
- Most importantly, unseen triphones may also be synthesized using the same phonetic state-tying tree that defines state clusters for cluster-based eigentriphone modeling as in conventional tied-state triphone HMM systems.

The derivation of clustered-based eigentriphones from tied-state clusters is similar to that of state-based eigentriphones except that STEP 1–3 in the latter method are modified as follows: First, construct a conventional tied-state triphone HMM for each base phone in which each state is represented by an M-component GMM. Then, re-estimate only the Gaussian means of each triphone, and its state covariances and mixture weights are copied from the corresponding tied state.

[3]In general, triphones or triphone states in each cluster may not even come from the same base phone, though, in this paper, they do.

[4]Note that the nodes selected for conventional state tying need not be the same as the nodes selected for cluster-based eigentriphone modeling; the two processes simply use the same phonetic decision tree.

V. Experimental Evaluation

The newly proposed cluster-based eigentriphone modeling method was evaluated on two speech recognition tasks: phone recognition on TIMIT [25] and medium-vocabulary continuous speech recognition on Wall Street Journal (WSJ) [1] 5K task.

In both tasks, we will compare the performance of the following five acoustic modeling methods:

- baseline1: conventional Baum-Welch training of triphone HMMs with no state tying.
- baseline2: conventional Baum-Welch training of tied-state triphone HMMs.
- model-based eigentriphone modeling of triphone HMMs as described in II (and no states are tied).
- state-based eigentriphone modeling of triphone HMMs as described in III (and no states are tied).
- cluster-based eigentriphone modeling of triphone HMMs using tied-state clusters as described in IV (but no states are tied).

Cross-word triphones were employed in all experiments and were modeled as continuous-density hidden Markov models (CDHMMs). Each CDHMM was a 3-state strictly left-to-right HMM in which the state distributions were modeled by a mixture of 16 Gaussians with diagonal covariances. In addition, there were a 1-state short pause model and and a 3-state silence model whose middle state was tied to the short pause state. Feature vectors were standard 39-dimensional MFCC acoustic vectors, and they were extracted from the training speech data every 10ms over a window of 25ms. The HTK toolkit [26] was used for HMM training and decoding with a beam width of 350.

In all systems described below, the transition probabilities of triphone models of the same base phone were tied together to those of the monophone model. On the other hand, for conventional Baum-Welch HMM training, Gaussian means, covariances, and mixture weights of triphones were re-estimated after the triphones were cloned from the monophone models if their sample counts are greater than the following thresholds: θ_m = 30, θ_v = θ_w = 200 respectively. For eigentriphone modeling, the thresholds are θ_m = 3, θ_v = θ_w = 200 respectively[5]. The sample count threshold for Gaussian means is much lower for eigentriphone modeling because we would like to use as many triphones as possible for the derivation of eigentriphones, and weighted PCA using the proposed weights already takes into account the reliability of each Gaussian mean. Moreover, all eigentriphone modeling methods perform (weighted) PCA using correlation matrices.

Furthermore, all system parameters such as the regularization parameter β, grammar factor, insertion penalty, as well as the optimal number of tied states for conventional HMM training, and the optimal number of state clusters for

[5]By default, triphones with less than three samples are not updated by HTK.

cluster-based eigentriphone modeling were determined using the respective development data set.

TABLE II
INFORMATION OF TIMIT DATA SETS.

Data Set	#Speakers	#Utterances	#Hours
training	462	3,696	3.14
core test	24	192	0.16
development	24	192	0.16

TABLE III
PHONE RECOGNITION ACCURACY (%) OF VARIOUS SYSTEMS ON TIMIT CORE TEST SET USING PHONE-TRIGRAM LANGUAGE MODEL. (THE FIGURE WITH AN $*$ IS STATISTICALLY AND SIGNIFICANTLY BETTER THAN BASELINE2 RESULT.)

Model	Accuracy
baseline1: conventional training (no state tying)	68.63
baseline2: conventional tied-state HMM training	71.95
model-based eigentriphone training model (no state tying)	71.27
state-based eigentriphone training model (no state tying)	71.03
cluster-based eigentriphone training model (587 state clusters)	**72.90**$*$

A. Phone Recognition on TIMIT

1) Speech Corpus and Experimental Setup: The standard NIST training set which consists of 3,696 utterances from 462 speakers was used to train the various models, whereas the standard core test set which consists of 192 utterances spoken by 24 speakers was used for evaluation. The development set is part of the complete test set, consisting of 192 utterances spoken by 24 speakers. Speakers in the training, development, and test set do not overlap. A summary of these data sets is shown in Table II.

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 phones [6] for error reporting. Moreover, the glottal stop [q] was ignored. At the end, there are altogether 15,546 cross-word triphone HMMs based on 48 base phones. Phone recognition was performed using Viterbi decoding with a trigram phone language model (LM) that was trained from the TIMIT training transcriptions using the SRILM language modeling toolkit [27]. The trigram LM has a perplexity of 14.39 on the core test set.

2) Acoustic Modeling: Five sets of triphone HMMs were built according to the five acoustic modeling methods mentioned in the beginning of this Section. For the conventional tied-state triphone HMM system (baseline2), there are a total of 587 tied states[6]. The dimension of triphone supervectors in model-based eigentriphone modeling is $3(\text{states}) \times 16(\text{mixtures}) \times 39(\text{MFCC}) = 1,872$. The dimension of triphone supervectors in state-based or cluster-based eigentriphone modeling is $16(\text{mixtures}) \times 39(\text{MFCC}) = 624$. The number of bases for the three methods is 44, 132, and 587, respectively[7]. In fact, cluster-based eigentriphone modeling was conducted using the clusters defined by the same 587 tied states in the baseline2 system.

3) Results and Discussion: Phone recognition results of the five systems are compared in Table III.

Though states are not tied in the three eigentriphone modeling methods, they outperform conventional HMM training without state tying by 3–4% absolute. In fact, their phone recognition performance is comparable to conventional tied-state HMM training, and cluster-based eigentriphone modeling actually outperforms the latter by 1% absolute.

Among the three eigentriphone modeling methods, the cluster-based method is the best, followed by the model-based method and then the state-based method. Both of the model-based method and state-based method estimate eigenbases from all triphones of a base phone, but the former method concatenates the three state supervectors of each triphone into one long triphone supervector for basis derivation. The better performance of the model-based method suggests that better eigenbases may be produced by making use of the correlation among the triphone HMM states. On the other hand, both of the state-based method and the current cluster-based method create eigenbases at the state level. The better performance of the cluster-based method must be attributed to the higher modeling resolution — 587 state clusters in the cluster-based method versus 132 state clusters in the state-based method — which more than compensates for the loss of state correlation as in the model-based method, and gives the best performance.

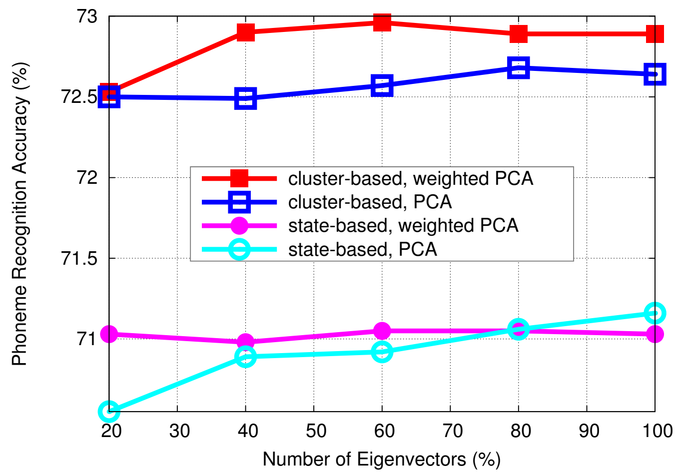

Fig. 2. Improvement of cluster-based eigentriphone modeling over state-based eigentriphone modeling on TIMIT phone recognition.

We further compared the performance of state-based eigentriphone modeling with cluster-based eigentriphone modeling when different forms of PCA was used, and when different proportions of eigentriphones were pruned. Eigentriphone pruning was done by first arranging the eigentriphones of each basis in descending order of their eigenvalues, and then retaining different number of leading eigentriphones for modeling the triphones. The result is shown in Fig. 2. Since the cluster-based method employs more state clusters than the state-based method (587 vs. 132), the former creates about four times more eigenbases than the latter. Equivalently, the number of eigentriphones in each eigenbasis produced by the former is only about $1/4$ of the latter on average. However, according

[6]The number of tied states was selected to maximize the phone recognition accuracy of the development set. It turns out the number is close to but not optimal on the core test set.

[7]Among the 48 phones that were selected for acoustic modeling, four phones are different variants of silence and closure, and they were modeled as monophone HMMs.

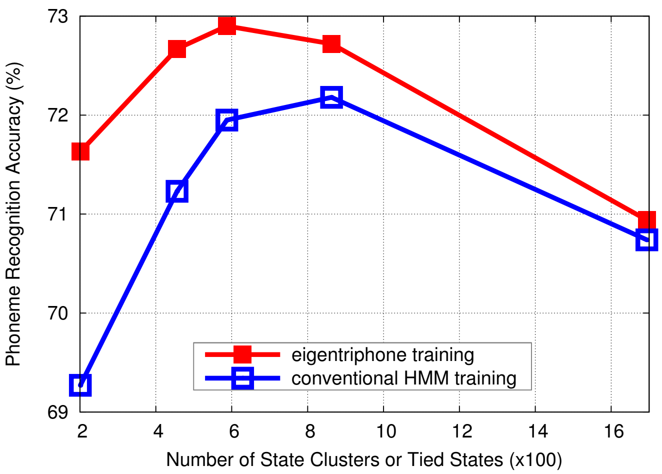

Fig. 3. Effect of the number of state clusters or tied states on cluster-based eigentriphone modeling and conventional tied-state HMM training on TIMIT phone recognition with phone-trigram language model.

to Fig. 2, one may still prune 60% of the eigentriphones in both methods without any performance loss[8]. The figure also shows that weighted PCA is more effective than standard PCA in deriving the eigentriphones in both methods.

Table III only shows the best results of various systems under the optimal settings determined by the development set. The effect of the number of state clusters on cluster-based eigentriphone modeling was also studied and compared with the effect of using different number of tied states on conventional HMM training as shown in Fig. 3. The results show that for the same number of state clusters (or tied states), cluster-based eigentriphone modeling always performs better than conventional tied-state HMM training, and the optimal number of state clusters is similar for both acoustic modeling methods. The difference between the two curves in the figure represents the amount of quantization error that is recovered by the current cluster-based eigentriphone modeling method.

There is a price to pay for the better performance of eigentriphone modeling. Since the triphone models produced by eigentriphone modeling are distinct, their Gaussian means are different from each other. Thus, their model size is much bigger than the models produced by conventional tied-state HMM training. For example, the conventional tied-state HMMs (baseline2) in Table III only require 1.4MB to store their Gaussian mean vectors[9], but cluster-based eigentriphone training requires 26.56MB with the use of the leading 40% eigenvectors. Nevertheless, we believe this modest increase in memory requirement by the eigentriphone modeling method will not pose a problem in most applications given the low price of today's memory chips.

[8]Note that 40% of eigentriphones in the cluster-based eigentriphone modeling method contain fewer eigentriphones than 40% of eigentriphones in the state-base eigentriphone modeling method. Specifically, the former is about $1/4$ of the latter.

[9]The memory requirement of transition probabilities and Gaussian variances are not considered here as cluster-based eigentriphone modeling copies them from the conventional tied-state HMMs. Thus, the memory requirements of these quantities for both training methods are the same.

B. Experiment on WSJ

1) Speech Corpus 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'92 and Nov'93 5K non-verbalized test set were used for evaluation using the standard 5K-vocabulary trigram language model (LM) 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 IV.

2) Acoustic Modeling: There were 18,777 cross-word triphones based on 39 base phones. For the conventional tied-state system (baseline2), the best performance was obtained with 7,374 tied states. The dimension of triphone supervectors in model-based, state-based, and cluster-based eigentriphone modeling are the same as those in the TIMIT experiment, namely 1872, 624, and 624, respectively; the number of bases for the three methods is 39, 117, and 7,374 respectively.

TABLE IV
INFORMATION OF WSJ DATA SETS.

Data Set	#Speakers	#Utterances	Vocab Size	OOV	LM Perplexity
SI284	283	37,413	13,646	11.95%	—
si_dt_05.odd	10	248	1,260	0	—
Nov'92	8	330	1,270	0	56.94
Nov'93	10	215	1,004	0.29%	61.82

TABLE V
WORD RECOGNITION ACCURACY (%) OF VARIOUS SYSTEMS ON THE WSJ 5K TASK USING TRIGRAM LANGUAGE MODEL.

Model	Nov'92	Nov'93
baseline1: conventional training; no state tying	95.61	94.05
baseline2: conventional tied-state HMM training	**96.32**	94.21
model-based eigentriphone training model	96.26	94.52
state-based eigentriphone training model	95.87	94.15
cluster-based eigentriphone training model	**96.32**	**94.54**

3) Results and Discussion: The word recognition results of various systems[10] are shown in Table V.

Comparing the performance of baseline1 and baseline2, we once again observe the effectiveness of state tying in triphone acoustic modeling. However, eigentriphone modeling can be an alternative: all the three variants of the method give comparable, if not better, recognition performance on WSJ. The state-based method is again the weakest among the three eigentriphone modeling methods, the model-based method and the cluster-based method have similar performance with the latter being slightly better. On the Nov'92 test set, the cluster-based eigentriphone modeling method has the same word recognition accuracy as the conventional tied-state HMM training method, but on the Nov'93 test set, the former actually

[10]Some of the results are different from those already reported in our past conference papers due to minor changes in the training procedures such as the number of Baum-Welch iterations.

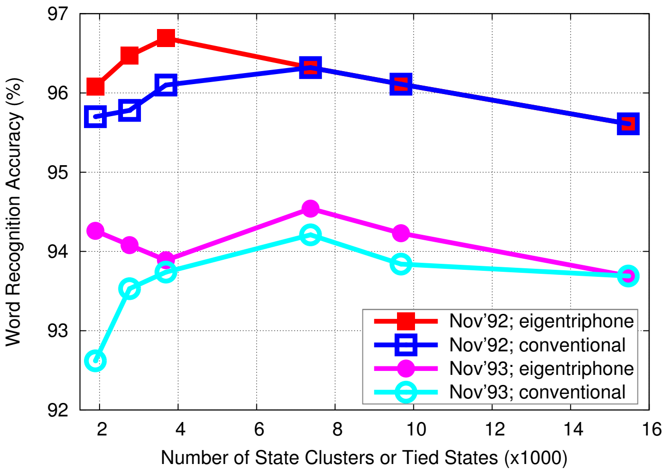

Fig. 4. WSJ recognition performance of cluster-based eigentriphone modeling and conventional tied-state HMM training with varying number of state clusters or tied states.

reduces the word error rate of the latter by 5.7%.

The performance of the cluster-based eigentriphone modeling method and conventional tied-state HMM training method over varying number of state clusters or tied states were also studied. The results are shown in Fig. 4. It can be seen that cluster-based eigentriphone modeling always perform better than conventional tied-state HMM training for the same number of state clusters or tied states. Note that on the Nov'92 test set, cluster-based eigentriphone modeling may achieve a better result of 96.69% word accuracy by using 3,690 state clusters. The worse result of the method in Table V was obtained with 7,374 state clusters which were found to be optimal on the development set.

VI. Conclusions

State tying is a commonplace in the construction of triphone hidden Markov models (HMM) for speech recognition. State tying effectively balances data among frequently and rarely occurring triphones to achieve robust estimation of their HMMs. However, it also introduces quantization errors among the tied states; that is, the tied states are not distinguishable. This paper presents another solution called eigentriphone modeling to the robust estimation of rarely occurring triphones without requiring state tying so that all trained triphones are generally distinct from each other. Three variants of the method are investigated, namely the model-based, state-based, and cluster-based eigentriphone modeling. The three variants differ in the modeling unit (triphones or triphone states) and resolution. With no surprise, empirically we find that the more general cluster-based eigentriphone modeling method using state clusters produced by the common phonetic state tying tree gives the best performance in both TIMIT phone recognition and WSJ word recognition. The use of state tying tree to define state clusters also allows us to synthesize unseen triphones using the same tree.

Although we call our method eigentriphone modeling, it can be applied to other phonetic units such as quinphones as well. Cluster-based eigentriphone modeling is also very flexible; in this paper, we only investigate its use on state clusters. In the future, we would like to investigate cluster-based eigentriphone modeling on other kinds of clusters such as clusters of triphones.

References

[1] 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.

[2] V. Pareto and A. S. Schwier, *Manual of Political Economy*. Augustus M. Kelley, Publishers, Jun. 1971.

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

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

[5] S. Takahashi and S. Sagayama, "Four-level tied-structure for efficient representation of acoustic modeling," in *Proc. of ICASSP*, 1995, pp. 520–523.

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

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

[8] H.-A. Chang and J. R. Glass, "A back-off discriminative acoustic model for automatic speech recognition," in *Proc. of Interspeech*, 2009.

[9] K. F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Trans. on SAP*, vol. 38, pp. 599–609, 1990.

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

[11] M. Y. Hwang and X. D. Huang, "Shared-distribution hidden Markov model for speech recognition," *IEEE Trans. on SAP*, vol. 1, pp. 414–420, 1993.

[12] E. Bocchieri and B. Mak, "Subspace distribution clustering hidden Markov model," *IEEE Trans. on SAP*, vol. 9, pp. 264–275, 2001.

[13] S. J. Young, "The general use of tying in phoneme-based HMM speech recognisers," in *Proc. of ICASSP*, 1992.

[14] J. J. O. 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.

[15] 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.

[16] J. R. Bellegarda and D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition," *IEEE Trans. on ASSP*, vol. 38, no. 12, pp. 2033–2045, December 1990.

[17] D. P. et al., "Subspace Gaussian mixture models for speech recognition," in *Proc. of ICASSP*, 2010.

[18] G. Saon and J.-T. Chien, "Bayesian sensing hidden Markov models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 43–54, Jan. 2012.

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

[20] T. Ko and B. Mak, "A fully automated derivation of state-based eigentriphones for triphone modeling with no tied states using regularization," in *Proc. of Interspeech*, 2011, pp. 781–784.

[21] ——, "Derivation of eigentriphones by weighted principal component analysis," in *Proc. of ICASSP*, 2012, pp. 4097–4100.

[22] 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.

[23] T. Ko and B. Mak, "Derivation of eigentriphones by weighted principal component analysis," in *Proc. of ICASSP*, 2012.

[24] ——, "A fully automated derivation of state-based eigentriphones for triphone modeling with no tied states using regularization," in *Proc. of Interspeech*, 2011.

[25] 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.

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

[27] A. Stolcke, "SRILM an extensible language modeling toolkit," in *Proc. of ICSLP*, 2002.