A FULLY AUTOMATED DERIVATION OF STATE-BASED EIGENTRIPHONES FOR TRIPHONE MODELING WITH NO TIED STATES USING REGULARIZATION

Tom Ko and *Brian Mak*

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

ABSTRACT

Recently we proposed an alternative method called *eigentriphone* to solve the data insufficiency problem in triphone acoustic modeling without the need of state tying. The idea is to treat the acoustic modeling problem of infrequent triphones ("poor triphones") as an adaptation problem from the more frequent triphones ("rich triphones"): firstly, an eigenbasis is developed over the rich triphones that have sufficient training data and the eigenvectors are called *eigentriphones*; then the poor triphones are adapted in a fashion similar to eigenvoice adaptation. Since, in general, no states are tied in our method, all triphones (states) are distinct so that they can be more discriminative than tied-state triphones.

In our previous work, the number of eigentriphones was determined in advance with a set of development data. In this paper, we investigate simply using *all* of them with the help of regularization to naturally penalize the less important ones. In addition, the model-based eigenbasis is replaced by three state-based eigenbases. Experimental evaluation on the WSJ 5K task shows that triphone models trained using our new eigentriphone approach *without* state tying perform at least as well as the common tied-state triphone models.
Index Terms: Eigenvoice, adaptation, eigentriphone, regularization.

1. INTRODUCTION

In (context-dependent) triphone hidden Markov modeling (HMM), parameter sharing is generally applied to ensure that there are sufficient training data for the robust estimation of the shared parameters. In particular, state tying using a phonetic tree [2] is a commonplace, and good recognition performance is always reported. Nevertheless, one plausible problem with state tying is that the tied-state triphones may become less discriminative because a state of one triphone may be identical to a state of another triphone, causing confusion between the two triphones during recognition.

Recently, Chang and Glass proposed a back-off discriminative acoustic modeling method based on broad phonetic classes [3]. In their work, the acoustic score of a triphone is computed from an interpolation between the native triphone model that is based on single-phone contexts and triphone models that are based on broad-phonetic-class contexts. Although it can guarantee that every triphone has a distinct acoustic score, acoustic-phonetic knowledge is required to derive the broad phonetic classes. How to get the "optimal" (if there is one) broad phonetic classes for any modeling units (triphones, syllables, etc.) requires further investigation.

Last year, we proposed an alternative approach to context-dependent acoustic modeling called *eigentriphone* [5]: an eigenbasis is computed from the frequent triphones of each base phone, and acoustic modeling of infrequent triphones is treated as an adaptation problem using the established eigenbases. Our method is similar to [3] in that, since states are generally not tied in our eigentriphone method, all triphones in our method are distinct from each other. On the other hand, our method has the advantage that no phonetic knowledge is required, and the whole method is data-driven and can be fully automated from the derivation of the eigenbases to the determination of the number of eigentriphones and the parameter estimation of the final triphone models.

Our method is motivated by the eigenvoice adaptation method [4]. Our eigentriphones are analogous to the eigenvoices in the eigenvoice adaptation method, but whereas there is only one eigenbasis in eigenvoice adaptation, we have 39 eigenbases — one eigentriphone basis for each of the 39 base phones. Furthermore, the selected number of eigentriphones represents the dimension of the eigentriphone-space, which can be determined using a set of development data. In our previous work [5], all triphones of the same base phone i use the same number of eigentriphones K_i (though the value of K_i varies with different base phones) for computational simplicity. Presumably, the value of K_i should vary with different triphones of the same base phone depending on their amount of training samples. In this paper, we would like to avoid making a hard decision on the number of eigentriphones. Instead, we investigate the use of regularization to make a soft decision on the number of eigentriphones; the regularization term will naturally penalize the less important eigentriphones. In addition, besides other minor changes, the previous model-based eigentriphones are replaced by state-based eigentriphones since parameter sharing at the state level is usually more effective than parameter sharing at the model level [2].

This paper is organized as follows. In Section 2, we will first review our original model-based eigentriphone acoustic modeling approach, and then describe how the procedure is modified using regularization so that it is now fully automated. That is followed by experimental evaluation in Section 3 and conclusions in Section 4.

2. STATE-BASED EIGENTRIPHONES

We will first review the derivation procedure of model-based eigentriphones, and then describe the proposed improvements.

2.1. Derivation of Model-based Eigentriphones

Fig. 1 shows an overview of the model-based eigentriphone approach that we previously proposed in [5] for the estimation of the infrequent or "poor" triphones. The derivation procedure of eigentriphones is similar to that of eigenvoices in the eigenvoice adaptation method [4] except that (a) speaker-dependent models in eigenvoice

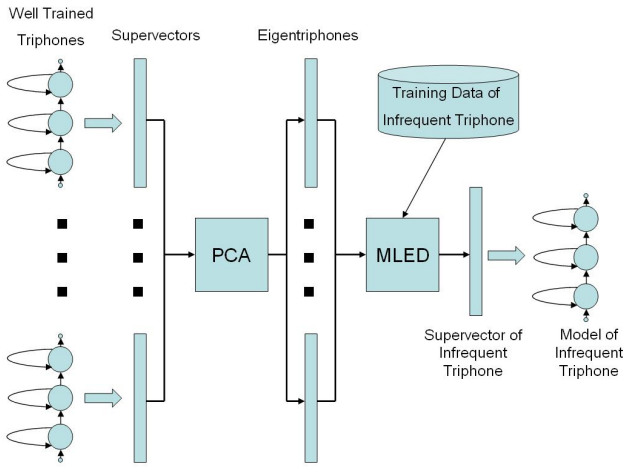

Fig. 1. The model-based eigentriphone adaptation approach.

are replaced by triphone models, and (b) whereas eigenvoice adaptation creates only one single set of eigenvoices for any speakers, a set of eigentriphones is created for each base phone (or monophone). Thus, since there are 39 base phones in our systems, 39 sets of eigentriphones have to be derived.

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

STEP 1: Monophone hidden Markov model (HMM) of base 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 triphones. *No state tying is performed for the triphones.*

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

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

STEP 4: Only Gaussian means of triphones in the rich set are re-estimated, Thus, *all* triphones of the same base phone will share the same set of Gaussian covariances, mixture weights, and transition probabilities which are copied from the base phone HMM.

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

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

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

STEP 6: Collect all rich triphone supervectors $\mathbf{v}_{i1}, \mathbf{v}_{i2}, \ldots, \mathbf{v}_{i|\Omega_i^R|}$ as well as the monophone supervector $\mathbf{m}_i$ of base phone i together, and derive an eigenbasis from their correlation matrix using *principal component analysis* (PCA).

STEP 7: Arrange the eigenvectors $\{\mathbf{e}_{ik}, k = 1, 2, \ldots, |\Omega_i^R|\}$ in descending order of their eigenvalues λ_{ik}, and select the top K_i eigenvectors so that they cover ϕ_v of the total variations. These K_i eigenvectors are called *eigentriphones* of phone i. In general, different base phones have a different number of eigentriphones.

STEP 8: Now the supervector $\mathbf{v}_{ip}$ of any poor triphone $p \in \Omega_i^P$ is assumed to lie in the eigenbasis spanned by the K_i eigentriphones. Thus, we have

$$\mathbf{v}_{ip} = \mathbf{m}_i + \sum_{k=1}^{K_i} w_{ipk}\mathbf{e}_{ik} \quad (2)$$

where $\mathbf{w}_{ip} = [w_{ip1}, w_{ip2}, \ldots, w_{ipK_i}]$ is the eigentriphone coefficients vector of triphone p in the "triphone space" of base phone i. Notice that the monophone supervector $\mathbf{m}_i$ is used instead of the mean of supervectors in Eqn.(2) so that the poor triphone model may fall back to the monophone HMM in the worst case.

STEP 9: Estimate the eigentriphone coefficient vector $\mathbf{w}_{ip}$ of the poor triphone p by maximizing the likelihood of its training data. Finally, the Gaussian mean of the mth mixture at the jth state of triphone p can be obtained from $\mathbf{v}_{ip}$ as

$$\boldsymbol{\mu}_{ipjm} = \mathbf{m}_{ijm} + \sum_{k=1}^{K_i} w_{ipk}\mathbf{e}_{ikjm} \ . \quad (3)$$

2.2. Investigation Issue #1: Soft Decision on the Number of Eigentriphones using Regularization

In our previous paper [5], the number of eigentriphones K_i is determined in advance using a separate development set. There are two drawbacks with the old scheme:

- In general, the value of K_i should also depend on the amount of available training data of each poor triphone p of base phone i. For triphones with more training data, a larger K_i will give them a better model.
- It is not clear how to set the threshold ϕ_v which determines the value of K_i. In [5], it is determined empirically using development data and the procedure is time-consuming.

Here, we attempt to make a soft decision on the value of K_i by using *all* eigentriphones with the help of regularization so that the less important eigentriphones are automatically de-emphasized. We define a new penalized log likelihood function $Q(\mathbf{w}_{ip})$ for the estimation of eigentriphone coefficient vector $\mathbf{w}_{ip}$ as follows:

$$Q(\mathbf{w}_{ip}) = L(\mathbf{w}_{ip}) - \beta R(\mathbf{w}_{ip}) \quad (4)$$

where β controls the relative importance of the regularization term $R(\cdot)$ compared with the likelihood term $L(\cdot)$. The log likelihood of the training data, is given by

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

where C_{ipjm} and $\gamma_{ipjm}(t)$ are the covariance and occupation probability of the mth Gaussian at the jth state of poor triphone p of base phone i given observation $\mathbf{x}_t$.

We investigate the following regularization term

$$R(\mathbf{w}_{ip}) = \sum_{k=1}^{|\Omega_i^R|} \frac{w_{ipk}^2}{\lambda_{ik}} \quad (5)$$

with the following considerations:

- The likelihood term should be more dominant when there are more training data. In particular, when a large amount of training data is available, the "adapted" triphone model should converge to its context-dependent estimate.
- Because of Eqn.(2), the adapted triphone model will converge to the monophone model for small amount of training data.
- Each eigentriphone coefficient, w_{ipk}, is scaled by the inverse of its corresponding eigenvalue so that a less informative eigentriphone will have less influence on the adapted model.

Differentiating the optimization function $Q(\mathbf{w}_{ip})$ of Eqn.(4) w.r.t. each eigentriphone coefficient, and setting each derivative to zero, we have,

$$\sum_{n=1}^{|\Omega_i^R|} A_{ipkn} w_{ipn} + \beta \frac{w_{ipk}}{\lambda_{ik}} = B_{ipk} \quad \forall k = 1, 2, \cdots |\Omega_i^R| \qquad (6)$$

where

$$A_{ipkn} = \sum_{j,m} \mathbf{e}'_{ikjm} C^{-1}_{ipjm} \mathbf{e}_{injm} \left(\sum_t \gamma_{ipjm}(t) \right)$$

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

The eigentriphone coefficients may be easily found by solving the system of $|\Omega_i^R|$ linear equations represented by Eqn.(6), and the Gaussian means of the new model may be computed using Eqn.(3). The training data may be re-aligned using the new model, obtaining a new set of occupation probabilities $\gamma_{ipjm}(t)$ so that the eigentriphone coefficients may be re-estimated. The procedure is repeated until the coefficients converge.

2.3. Investigation Issue #2: State-based Eigentriphones

The eigentriphone adaptation framework described so far uses the whole triphone model to construct a supervector, and we will call the resulting eigenvectors as *model-based eigentriphones*. Actually the construction unit can be very flexible. In this paper, we would like to investigate the performance of *state-based eigentriphones* which are obtained by creating three separate eigenbases, one from each state. Compared with model-based eigentriphones, state-based eigentriphones have 3 times more eigenbases, but the dimension of each state-based eigentriphone is 1/3 of a model-based eigentriphone.

2.4. Other Improvements

Re-estimation thresholds θ_v^R, θ_w^R, and θ_t^R are further defined to control the re-estimation of triphone covariances, mixture weights, and transition probabilities respectively based on its sample count. That is, the quantity will only be re-estimated if the sample count of the triphone exceeds the respective threshold, otherwise, its value is simply copied from its monophone model.

Table 1. Information of various WSJ data sets.

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

3. EXPERIMENTAL EVALUATION

3.1. Speech Corpora and Experimental Setup

The standard SI-284 Wall Street Journal (WSJ) training set was used for training the speaker-independent models. It consists of 7,138 WSJ0 utterances from 83 speakers, and 30,275 WSJ1 utterances from 200 speakers. Thus, there are a total of about 70 hours of read speech in 37,413 training utterances from 283 speakers. All the training data were endpointed. Both of the standard Nov'92 and Nov'93 5K non-verbalized test sets were used for evaluation using the standard 5K-vocabulary bigram and trigram that came along with the WSJ corpus. The data set, si_dt_05.odd, was used for tuning all system parameters. It contains alternate sentences from the 1993 WSJ 5K Hub development data set but sentences with OOV words were removed. These data sets are summarized in Table 1.

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

The sample count thresholds for categorizing poor and rich triphones, namely, θ_m^R and θ_m^P, were set to 30 and 200 respectively. As a result, the two sets overlapped; there were 8,896 triphones in the rich set and 15,884 triphones in the poor set. The sample count thresholds for the re-estimation of the covariances, mixture weights, and transition probabilities were set to 200, 30, 200 respectively. The regularization parameter β was set to 15. These values were all determined empirically from maximizing the recognition accuracy on the development data set.

3.2. Baseline Systems

Three baseline systems were trained for comparison.

1. Baseline1: A conventional tied-state triphone system. A total of 6,481 tied states were derived using a phonetic decision tree. The number of tied states was selected to maximize the recognition accuracy on the development set.
2. Baseline2: A triphone system with no tied states. Monophone HMMs were first trained and then cloned to initialize the corresponding triphones. Then the Gaussian means and covariances, mixture weights, and transition probabilities of a triphone were *only* re-estimated if its sample count exceeds the corresponding thresholds, θ_m^R, θ_v^R, θ_w^R, and θ_t^R respectively.
3. Baseline3: Similar to Baseline2 except that after monophones cloning, *only* Gaussian means of those triphones with sample count greater than the threshold θ_m^R were estimated; the remaining HMM parameters (covariances, mixture weights, and transition probabilities) still keep their corresponding monophone values. Thus, all triphones of the same base phone differ only in their Gaussian means.

3.3. The Eigentriphone Model

The state-based eigentriphone adaptation was carried using the Baseline3 models according to the procedure described in Section 2. The dimension of each triphone state supervector is 16 (mixtures) $\times 39$ (MFCCs) $= 624$. Afterwards, the remaining HMM parameters of each triphone: Gaussian covariances, mixture weights, and transition probabilities were further re-estimated if its sample count exceeds the respective re-estimation thresholds.

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

Model	Description	Nov'92	Nov'93
Baseline1	tied-state triphones	94.56	91.40
Baseline2	no state tying; rich triphones are re-estimated; poor triphones are clones of monophones	93.98	91.79
Baseline3	no state tying; only Gaussian means of rich triphones are re-estimated	93.50	90.83
	+ state-based eigentriphone "adaptation" of means for poor triphones	93.78	91.37
	+ re-estimation of Gaussian variances, mixture weights, and transition probabilities when the respective re-estimation thresholds are met	94.53	92.44*

Table 3. Recognition word accuracy (%) of various systems on the WSJ 5K task using trigram language model.

System	Nov'92	Nov'93
tied-state triphone system	96.45	93.89
state-based eigentriphone system	96.41	94.47
model-based eigentriphone system	96.47	94.44

3.4. Results and Discussions

The recognition performance of the various systems using bigram and trigram language models are shown in Table 2 and Table 3 respectively. It is observed that

- the proposed state-based eigentriphone adaptation approach for the estimation of Gaussian means of the poor triphones with subsequent re-estimation of their remaining HMM parameters is effective. Each step gives incremental improvement over the Baseline3 system.
- Triphones without state-tying but trained using our proposed state-based eigentriphone approach outperform tied-state triphones on the Nov'93 test set, and perform slightly worse on the Nov'92 test set. However, most of the differences in their recognition accuracies are not statistically significant.
- Table 3 also compares the performance of state-based and model-based eigentriphones; they performed equally well. Since model-based eigentriphones result in fewer weights, they are preferred over the state-based eigentriphones.

3.5. Effect of Regularization for the Soft Decision on the Number of Eigentriphones

The performance of the proposed regularization method that makes a soft decision on the number of eigentriphones is compared with that of using a hard decision so that a fixed percentage, ϕ_v, of the total variations are covered. The comparison result on the Nov'93 test set is shown in Fig 2. It can be seen that when a fixed number of eigentriphones are to be used, the number should not be too large or too small: when it is too large, the triphones with small amount of training data cannot be robustly estimated; when it is too small, the triphones with large amount of training data will be under-trained. The proposed regularization term helps avoid making the hard decision, and the resulting system performs at least as well as a system that uses the best hard number of eigentriphones.

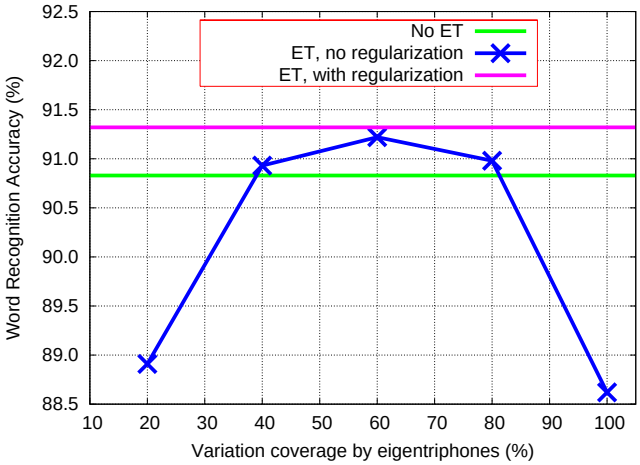

Fig. 2. Effect of soft/hard decision on the number of eigentriphones.

4. CONCLUSIONS AND FUTURE WORK

We successfully avoid making a hard decision on the number of eigentriphones to use in the eigentriphone adaptation framework for distinctive acoustic modeling with no state-tying. This is achieved by adding an appropriate regularization term in the objective estimation function of the eigentriphone coefficients so that eigentriphones with smaller eigenvalues are de-emphasized. The resulting recognition systems perform at least as well as tied-state triphone systems.

In the current work, only Gaussian means of the poor triphones were "adapted". We would like to investigate the adaptation of other HMM parameters, such as Gaussian variance and mixture weights in the future. We will also look at discriminative training of triphones trained using eigentriphone method. This can be done at two levels: (a) discriminative eigentriphones may be derived using discriminative component analysis methods such as LDA; (b) discriminative training methods such as MCE or MMI training.

5. ACKNOWLEDGEMENTS

This work was supported by the Research Grants Council of the Hong Kong SAR under the grant numbers HKUST617008.

6. REFERENCES

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

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

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

[4] 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.

[5] Tom Ko and B. Mak, "Eigentriphones: A basis for context-dependent acoustic modeling," in *Proc. of ICASSP*, 2011.