EIGENTRIPHONES: A BASIS FOR CONTEXT-DEPENDENT ACOUSTIC MODELING

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

In context-dependent acoustic modeling, it is important to strike a balance between detailed modeling and data sufficiency for robust estimation of model parameters. In the past, parameter sharing or tying is one of the most common techniques to solve the problem. In recent years, another technique which may be loosely and collectively called the subspace approach tries to express a phonetic or sub-phonetic unit in terms of a small set of canonical vectors or units. In this paper, we investigate the development of an eigenbasis over the triphones and model each triphone as a point in the basis. We call the eigenvectors in the basis *eigentriphones*. From another perspective, we investigate the use of the eigenvoice adaptation method as a general acoustic modeling method for training triphones — especially the less frequent triphones without tying their states so that all the triphones are really distinct from each other and thus may be more discriminative. Experimental evaluation on the 5K-vocabulary HUB2 recognition task shows that a triphone HMM system trained using only eigentriphones without state tying may achieve slightly better performance than the common tied-state triphones.

Index Terms— Eigenvoices, eigentriphones, context-dependent acoustic modeling, adaptation.

1. INTRODUCTION

It is well-known that for any reasonably complicated automatic speech recognition (ASR) task, it is crucial to use context-dependent (CD) acoustic units to model contextual acoustic variations. With an inventory of 40–60 context-independent (CI) phones, the number of CD phones grows exponentially with the extent of context that one considers. Take the triphones as an example; if we assume 40 CI monophones, theoretically, there may be a maximum total of $40^3 = 64,000$ triphones. Even though, in practice, most of these triphones do not appear, they may distribute very unevenly. For instance, Fig.1 depicts the triphones coverage in the HUB2 WSJ0/WSJ1 training corpus. There are 18,991 triphones, and only 3,510 of them have more than 200 samples. That is, about 80% of the training data are concentrated on (the most common) 20% of all the seen triphones. Thus a major challenge in CD modeling is to estimate the less frequent CD units reliably, otherwise the poorly-trained models may affect the overall performance of an ASR system.

Parameter sharing or tying has been a common technique to strike a balance between detailed context modeling and data insufficiency. The idea is to group the acoustic units of interest into disjoint classes so that members of the same class will share the same model parameters and thus their training data. The classes of acoustic units are usually determined automatically in some data-driven approach such as the decision tree clustering method. Various parameter tying units have been tried resulting in, for example, generalized triphones [1], tied states [2], shared distributions or senones [3], and tied subspace Gaussian distributions [4].

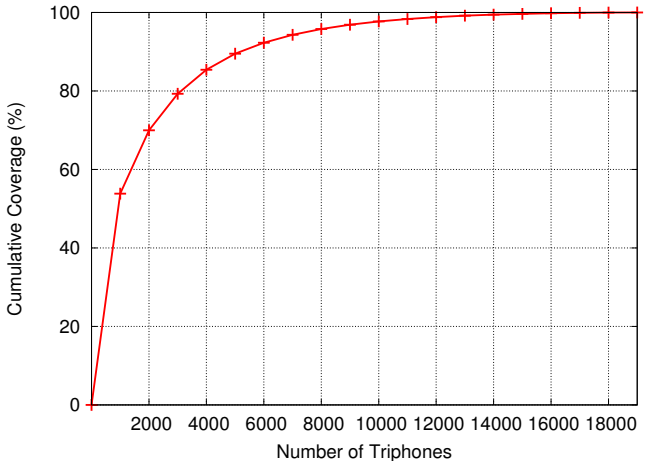

Fig. 1. Cumulative triphones coverage in the training set of HUB2

Another solution is model interpolation or smoothing, in which models of various modeling resolutions are interpolated to give a more reliable estimate of the poorly-trained CD model. For example, deleted interpolation is used in the Sphinx speech recognition system to improve the robustness of the generalized triphones [5]. Recently, the back-off acoustic modeling [6] is proposed to combine a triphone score with those from acoustic models that are defined on phonetic class contexts. The back-off method has the benefit that in contrast to the common tied-state triphones, the states in its triphones may not be tied so that each state is distinct with its own acoustic score.

Recently, another approach seems to be emerging, and it will be loosely and collectively called the subspace approach in this paper. In the subspace approach, a phonetic or sub-phonetic unit is expressed in terms of a small set of canonical vectors or units. Examples are semi-continuous hidden Markov model (SCHMM) [7], subspace Gaussian mixture model (SGMM) [8], and canonical state model (CSM) [9].

In this paper, we investigate the development of an eigenbasis over the triphones and model each triphone as a point in the basis. We call the eigenvectors in the basis *eigentriphones*. The eigentriphone method belongs to the recent subspace approach. Eigen-

This work was supported by the Research Grants Council of the Hong Kong SAR under the grant numbers HKUST617008 and HKUST617507.

triphone differs from the other methods as follows:

- the modeling unit in this paper is the whole triphone and not a sub-phonetic unit such as states in CSM. Nevertheless, the method is very flexible and may be applied to sub-phonetic units or bigger linguistic units such as syllables as well.
- state are generally not tied in our eigentriphone method so that, similar to the back-off acoustic modeling method and unlike the other subspace methods, all the triphones in our new method are distinct from each other. Thus, they can be more discriminative than those trained in other methods.
- no phonetic knowledge is required as the whole method is data-driven: the number of eigentriphones may be determined automatically.
- the eigentriphones can be found using formal eigen-decomposition without the application of any heuristic.
- unlike some state tying methods using a phonetic tree or back-off acoustic modeling using broad phonetic classes, since eigentriphones does not require any phonetic knowledge, one may easily modify it for other acoustic units such as syllables without having to re-derive all the phonetic questions which may not be as obvious as in the case of triphone state tying.

From another perspective, our new eigentriphones is motivated by eigenvoice adaptation. The estimation of triphones with insufficient training samples is treated as an adaptation problem, and the eigenvoice approach is used to solve the data insufficiency problem without sacrificing detailed context modeling.

This paper is organized as follows. In Section 2, we will describe our eigentriphone approach for training triphones with few training samples. That is followed by experimental evaluation in Section 3 and conclusions in Section 4.

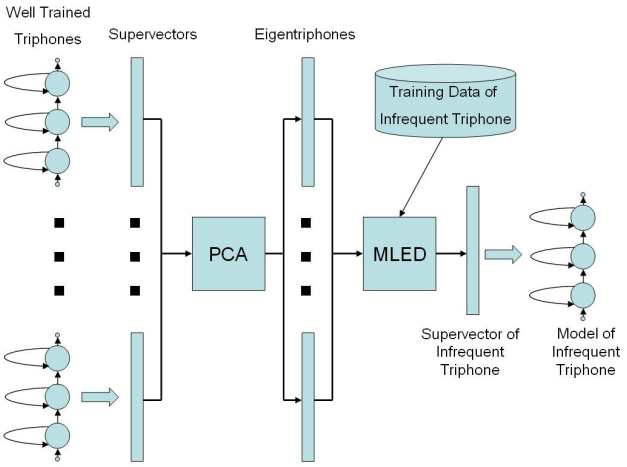

Fig. 2. An overview of the eigentriphone approach.

2. EIGENTRIPHONES

Fig. 2 shows an overview of the proposed eigentriphone approach for the estimation of "poor" triphones.

2.1. Derivation of the Eigenbasis for Triphones

The procedure is basically the same as the eigenvoice adaptation approach [10] except that speaker-dependent models are replaced by triphone HMMs, and that an eigenbasis is created for each base phone (or monophone). Thus, as there are 39 base phones in our experiments, 39 eigenbases have to be derived. The following procedure is repeated for each base phone over its triphones in the training data.

STEP 1: For each base phone p, collect all its triphones from the training corpus and split them into 2 groups: the rich set R_p and the poor set P_p based on a threshold θ_r on the sample count. If the sample count of a triphone is greater than θ_r, it is put into the rich set, otherwise into the poor set.

STEP 2: Monophone hidden Markov models (HMM) are 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). Let's denote the GMM of the jth state (where $j = 1, 2, 3$) of base phone p as

$$p_{pj}(\mathbf{x}_t) = \sum_{m=1}^{M} c_{pjm}\mathcal{N}(\mathbf{x}_t; \mu_{pjm}, \Sigma_{pjm}) \; . \tag{1}$$

STEP 3: The monophone HMM of base phone p is cloned to initialize all its triphones. *No state tying is performed for the triphones.* The Gaussian means are then reestimated while the mixture weights and Gaussian covariances are kept unchanged. Thus, after the reestimation of the Gaussian means, the triphone q of the base phone p will have the following GMM:

$$p_{pqj}(\mathbf{x}_t) = \sum_{m=1}^{M} c_{pjm}\mathcal{N}(\mathbf{x}_t; \mu_{pqjm}, \Sigma_{pjm}) \; . \tag{2}$$

Notice that all triphones of the same base phone will share the same set of mixture weights and covariances and differ only in their Gaussian means.

STEP 4: For each triphone q in the rich set R_p of the base phone p, create a triphone supervector by stacking up all its Gaussian mean vectors in each of its 3 states as follows.

$$\mathbf{v}_{pq} = \begin{bmatrix} \mu_{pq11}, & \mu_{pq12}, & \cdots, & \mu_{pq1M}, \\ \mu_{pq21}, & \mu_{pq22}, & \cdots, & \mu_{pq2M}, \\ \cdots, & \cdots, & \cdots, & \cdots, \\ \mu_{pq31}, & \mu_{pq32}, & \cdots, & \mu_{pq3M} \end{bmatrix} \; . \tag{3}$$

STEP 5: Collect the triphone supervectors $\mathbf{v}_{p1}, \mathbf{v}_{p2}, \ldots, \mathbf{v}_{p|R_p|}$ of base phone p, and derive an eigenbasis from their correlation matrix by *principal component analysis* (PCA).

STEP 6: Arrange the eigenvectors $\mathbf{e}_{pk}, k = 1, 2, \ldots, |R_p|$ in descending order of their eigenvalues, and select the top K_p eigenvectors so that they together cover θ_v of the total variations. (In the current paper, $\theta_v = 80\%$.) These K_p eigenvectors are the *eigentriphones* of phone p. Notice that different base phones will, in general, have a different number of eigentriphones.

STEP 7: Now the supervector of any triphone of base phone p is assumed to lie in the basis spanned by its K_p eigentriphones. Thus, each triphone q' in the poor set P_p may be expressed as

$$\mathbf{v}_{pq'} = \mathbf{e}_{p0} + \sum_{k=1}^{K_p} w_{pq'k}\mathbf{e}_{pk} \tag{4}$$

where $\mathbf{e}_{p0}$ is the average of all the "rich" triphone supervectors of phone p.

STEP 8: The Gaussian means of the poor triphone q' may be derived from its supervector $\mathbf{v}_{pq'}$. On the other hand, its Gaussian covariances and mixture weights are simply copied from its corresponding monophone HMM.

STEP 9: The eigentriphone coefficients $\mathbf{w}_{pq'k}$ (where $k = 1, 2, \ldots, K_p$) may be estimated using the MLED algorithm as in eigenvoice [10] by maximizing the likelihood of its training data.

STEP 10: The estimation of the eigentriphone coefficients is repeated until the coefficients converge.

Table 1. Information of various WSJ data sets.

Data Set	#speakers	#utterances	vocab size
train (si_tr_s)	302	46,995	13,725
dev (si_dt_05)	10	496	1,842
eval (si_et_h2)	10	205	998

Table 2. Recognition performance of various system on the test set.

Model	Word Acc.
Baseline1: tied-state triphones	91.45%
Baseline2: no state tying; rich triphones reestimated; poor triphones are clones of monophones	89.72%
Baseline3: no state tying; only Gaussian means of all triphones reestimated	89.99%
+ eigentriphone "adaptation" for the poor set	91.09%
+ further training for the rich set	91.58%

3. EXPERIMENTAL EVALUATION

3.1. Speech Corpora and Experimental Setup

The standard SI-284 Wall Street Journal (WSJ) training set plus additional WSJ adaptation data and short-term training data was used for training the speaker-independent model. It consists of 8,720 WSJ0 utterances from 101 WSJ0 speakers and 38,275 WSJ1 utterances from 201 WSJ1 speakers. Thus, there is a total of about 44 hours of read speech in 46,995 training utterances from 302 speakers.

The standard Nov'93 5K non-verbalized Hub2 test set si_et_h2 was used for evaluation using the standard 5K-vocabulary bigram that came along with the WSJ corpus. The WSJ1 5K development set si_dt_05 was used for tuning the system parameters. Notice that utterances containing OOV words were removed from both the development and evaluation test sets. A summary of these data sets is shown in Table 1.

There were altogether 18,991 cross-word triphones based on 39 base phonemes. Each triphone model was a strictly left-to-right 3-state continuous-density hidden Markov model (CDHMM), with a Gaussian mixture density of at most $M = 16$ components per state. In addition, there were a 1-state short pause model and a 3-state silence model. The traditional 39-dimensional MFCC vectors were extracted at every 10ms over a window of 25ms. Recognition was performed using the HTK toolkit [11] with a beam search threshold of 250.

The sample count threshold for rich triphones θ_r was set to 200, and the variation coverage threshold θ_v was set to 80%.

3.2. Baseline Systems

Three baseline systems were trained for comparison.

- Baseline1: A conventional tied-state triphone system. There were totally 5,864 tied states which were derived from a phonetic decision tree.
- Baseline2: A triphone system with no tied states. The corresponding monophone system was first trained and then cloned to initialize the triphones. Then only the triphones in the rich sets were reestimated. All model parameters — transition probabilities, Gaussian means and covariances, and mixture weights — of the rich sets were reestimated. The triphones in the poor set were all tied to their monophone model.
- Baseline3: Same as the second baseline except that now only the Gaussian means of *all* triphones were estimated and they all shared the same monophone transition probabilities, mixture weights, and Gaussian covariances.

The recognition word accuracies of these three baselines are shown in Table 2.

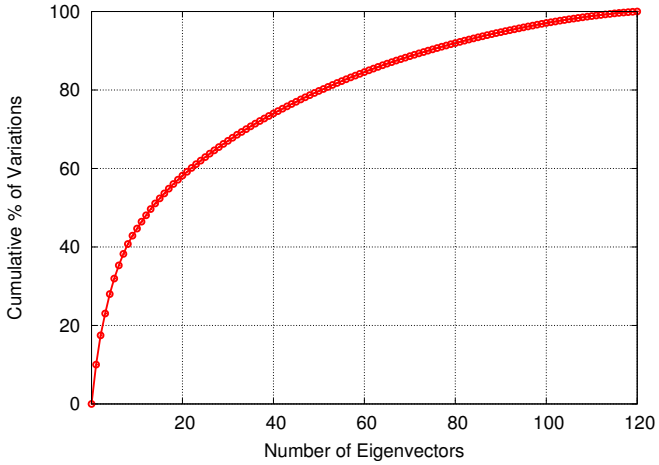

Fig. 3. Variation coverage by the eigentriphones derived from the rich set of the base phone [er].

3.3. The Eigentriphone Model

Eigenbasis was derived from the triphones in the baseline3 models according to the procedure described in Section 2. Since all the triphones in this baseline have the same mixture weights and Gaussian covariances, one may create the triphone supervectors by stacking up the Gaussian mean vectors of a state GMM in a consistent order for all the triphones of the same base phone. The dimension of these triphone supervectors is 3 (states) $\times$16 (mixtures) $\times$39 (MFCC) $= 1872$ parameters. PCA was performed for the rich triphones of each base phone and the number of eigentriphones was determined by the variation coverage threshold $\theta_v = 80\%$. Fig. 3 shows the cumulative variation coverage by the eigenvectors of the rich triphones of the phone [er]. In the case of [er], $\theta_v = 80\%$ requires selecting 50 eigentriphones.

After the eigentriphone coefficients were estimated ("adaptation") for the poor triphones, the Gaussian covariances, mixture weights, and transition probabilities of the rich triphones may further be reestimated.

The recognition performance of the eigentriphone systems are also shown in Table 2.

3.4. Results and Discussions

From Table 2 one can see that

- comparing the 3 baselines, it is clear that state tying using a phonetic decision tree is effective and may boost the recognition accuracy by 1.5% absolute.
- the proposed eigentriphone approach is also effective in training the poor triphones, resulting in a system whose recognition performance is even better than a conventional tied-state triphone system by 0.13% absolute.

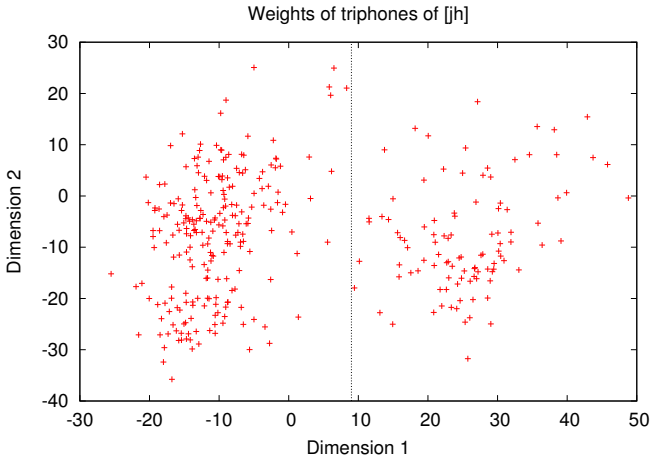

Fig. 4. The first 2 eigentriphone coefficients of all the triphones of the base phone [jh].

3.5. Analysis of Eigentriphone Coefficients

We took a quick look at the first two eigentriphone coefficients of all the triphones of each base phone. One of the most interesting result is shown in Fig. 4 for the base phone [jh].

There are 328 points ([jh] triphones) on the plot. A line with the first coefficient being 9 is drawn on the plot as well. It is interesting to see that for the 102 triphones lying to the right of this line, all of them except one have a consonant as their right context; whereas, for the 226 triphones lying to the left of the line, again all except one of them have a vowel as their right context.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we successfully derive a basis for acoustic modeling. In the case of triphone modeling, the result is a set of eigentriphones from which the less frequent triphones may be estimated. From another perspective, we try to view the data insufficient problem in context-dependent acoustic modeling as an adaptation problem, and existing adaptation techniques are applied for acoustic modeling. Experimental results show that triphones with no tied states trained in this way performs slightly better than tied-state triphones.

In this work, only the Gaussian means of the poor triphones were "adapted" and all the remaining model parameters were copied from their corresponding context-independent models. In the future, We would like to extend the method to include other model parameters. Moreover, the adaptation perspective of our new acoustic modeling method may suggest that other adaptation algorithms be investigated as well.

Finally, a basis consisting of eigentriphones of acoustic model renders a compact representation of acoustic models: one may simply store the eigentriphones for each base phone, and then each triphone model will be represented by a small set of eigentriphone coefficients.

5. REFERENCES

[1] K. F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 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 *Proceedings of the European Conference on Speech Communication and Technology*, 1993, vol. 3, pp. 2203–2206.

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

[4] E. Bocchieri and Brian Mak, "Subspace distribution clustering hidden Markov model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 264–275, March 2001.

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

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

[7] 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.

[8] 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.

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

[10] 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.

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