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

Background

In context-dependent acoustic modeling, the number of modeling units grows exponentially while their training samples usually distribute unevenly.

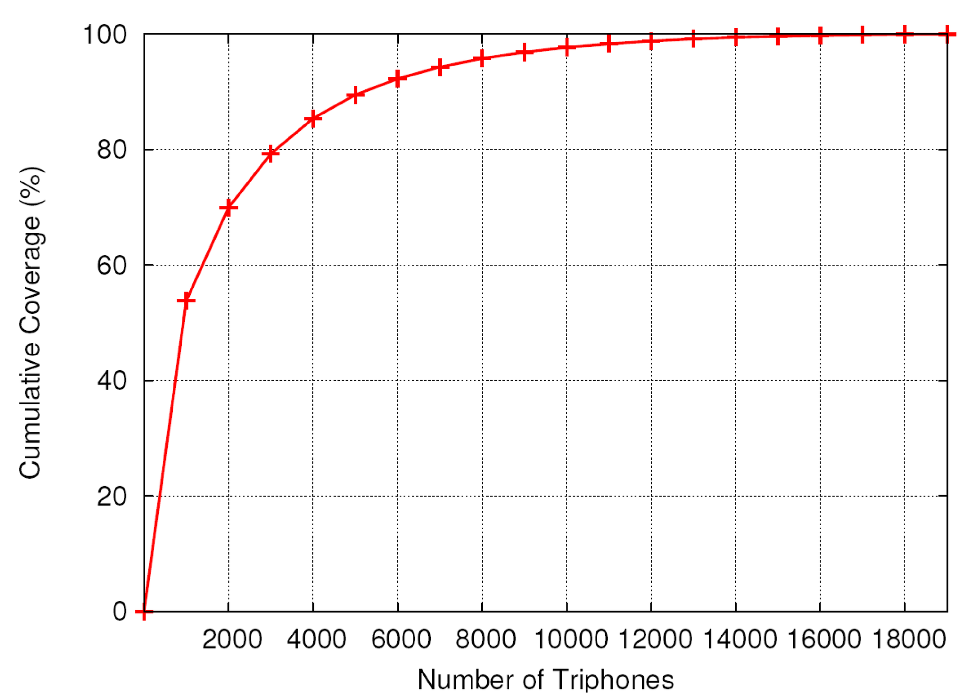

Fig. 1. Cumulative triphones coverage in the training set of WSJ0+1.

As a result, a number of modeling units might get so few training samples that they are poorly estimated. These poorly-trained models may affect the overall performance of the system.

Motivation

Parameter sharing (e.g. state tying) has been a common technique to tackle the problem of data sparsity, however it has the following drawbacks:

- **It may cause a potential drop of the overall discriminative power as some models (or some parts of models) are identical to the recognizer.**
- **Phonetic knowledge is often needed (e.g. phonetic decision tree), which may not be generalized easily for other acoustic units.**

Our Proposal: Eigentriphones

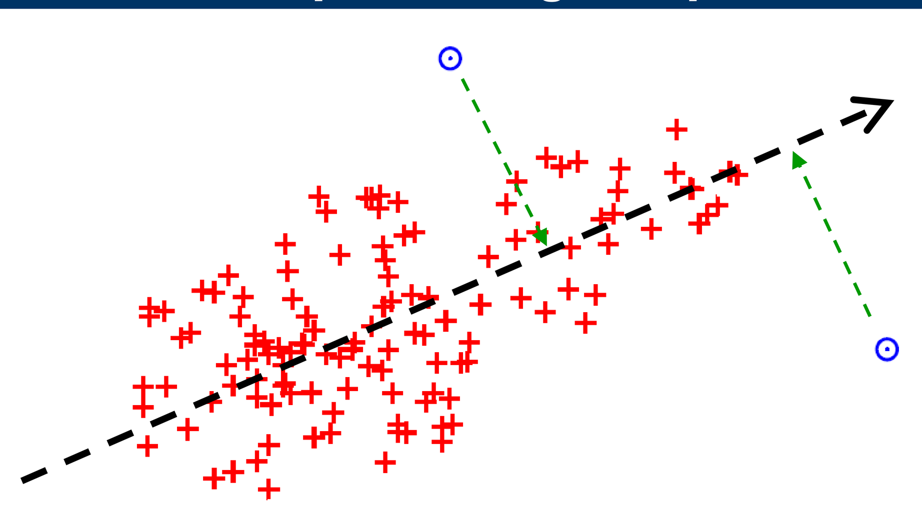

Fig. 2. An example showing the concept of eigentriphones.

"Adapt" triphones with few training samples from those with many samples.

Motivated by the eigenvoice adaptation method, we investigate the development of an eigenbasis over triphones and model each triphone as a point in the triphone-space.

The Eigentriphones

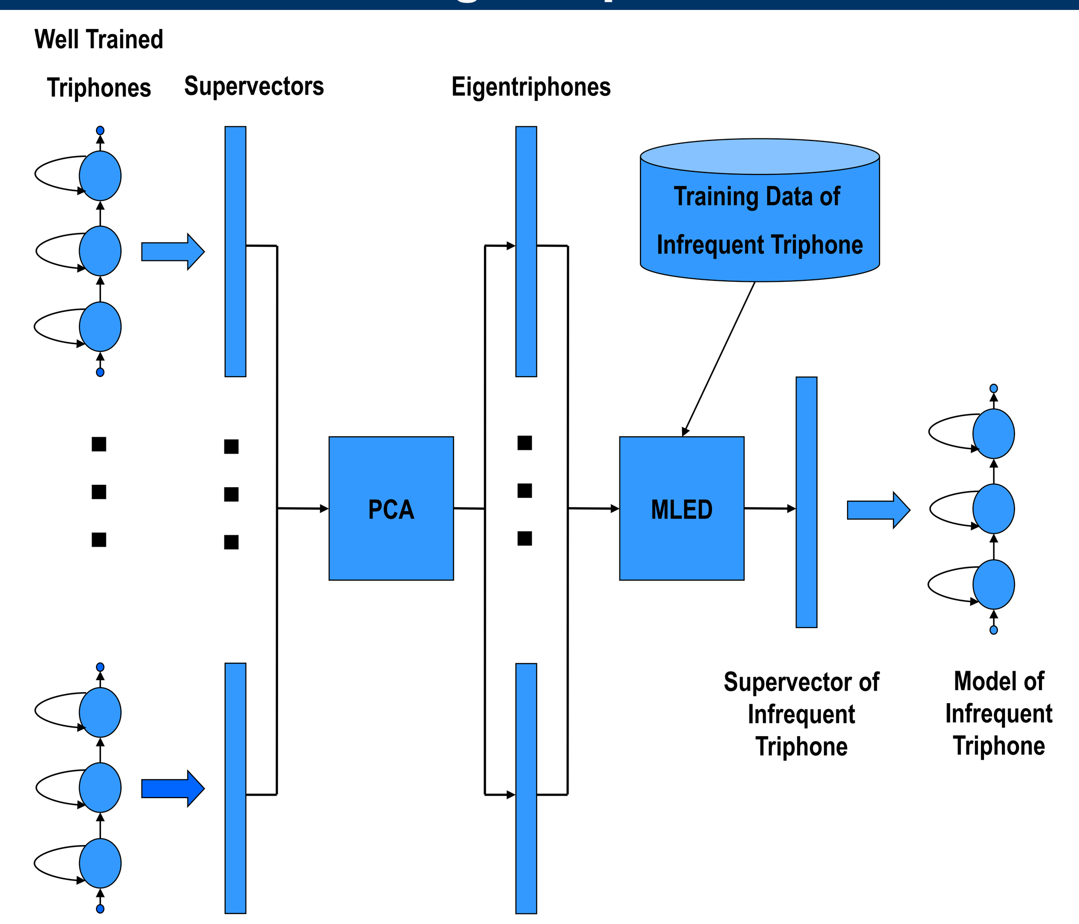

Fig. 3. An overview of the eigentriphone approach.

For each base phone, split all its triphones into 2 groups: the rich set and the poor set based on a sample count threshold θ_r.

Monophone HMMs are estimated with a mixture of Gaussians in each state. Then these monophones are cloned to initialize all their corresponding triphones.

For each triphone in the rich set, create a supervector by stacking up all its Gaussian mean vectors from all of its states.

For each base phone, collect its triphone supervectors in the rich set and derive an eigenbasis from their correlation matrix by PCA.

Arrange the eigenvectors in the descending order of their eigenvalues, and select the top K eigenvectors so that they cover θ_v of the total variations.

Then the supervector v of each poor triphone is expressed as a linear combination of the eigentriphones e_k:

$$v = e_0 + \sum_{k=1}^{K} w_k e_k$$

where e_0 is the average of the rich triphone supervectors.

The eigentriphone coefficients w_k (where k = 1, ..., K) of each poor triphone are estimated using the MLED algorithm by maximizing the likelihood of its training data.

The *Gaussian means* of the poor triphone are derived from its supervector while its *Gaussian covariances* and *mixture weights* are copied from the corresponding monophone.

Experimental Setup & Results

Category	Setup
Training Set	46,995 utterances from WSJ0+1 short-term training data
Development Set	496 utterances from WSJ1 5K development set
Test Set	205 utterances from WSJ1 Nov'93 5K test set
#Seen Triphones	18,991
#Gaussian / state	16
#State / phone	3
Language Model	Bigram
Dictionary	CMU dictionary
Feature Vector	Standard 39-d MFCC
Sample Count Threshold θ_r	200
#Triphones in Rich Set	3,510
Variation Coverage Threshold θ_v	80%

Model	Word Acc.
Baseline 1: tied-state triphones	**91.45%**
Baseline 2: no state tying; only Gaussian means of all triphones re-estimated; other parameters are copied from monophones	**89.99%**
+ eigentriphone "adaptation" of Gaussian means for the poor set	**91.09%**
+ further training of Gaussian covariances and mixture weights for the rich set	**91.58%**

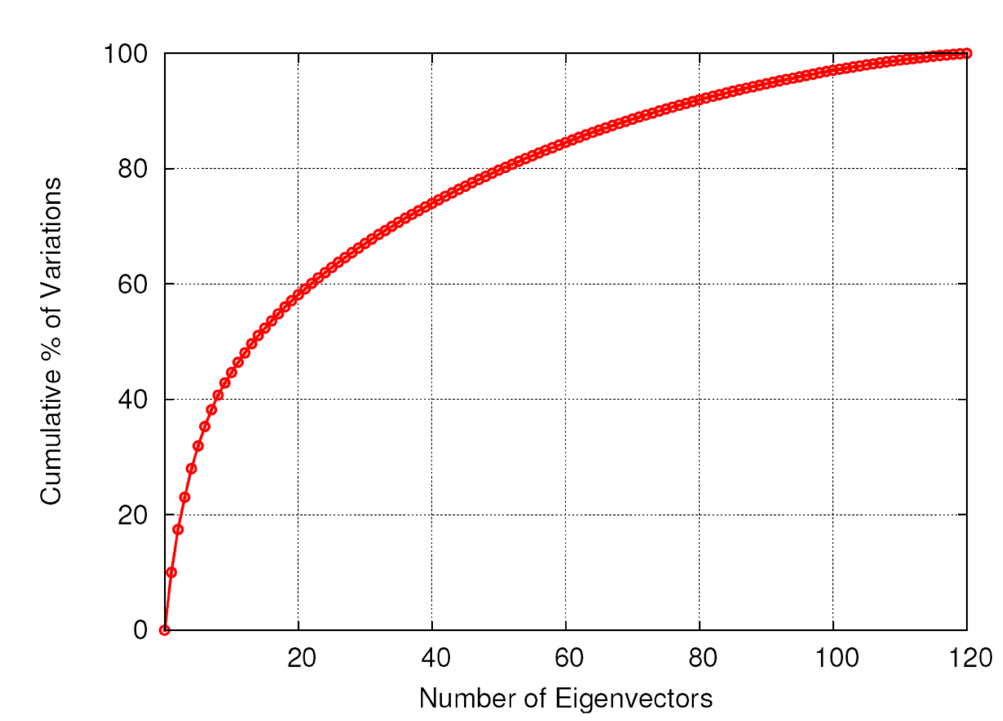

Fig. 4. Variation coverage by the eigentriphones derived from the rich set of the base phone [er].

Analysis of Eigentriphone Coefficients

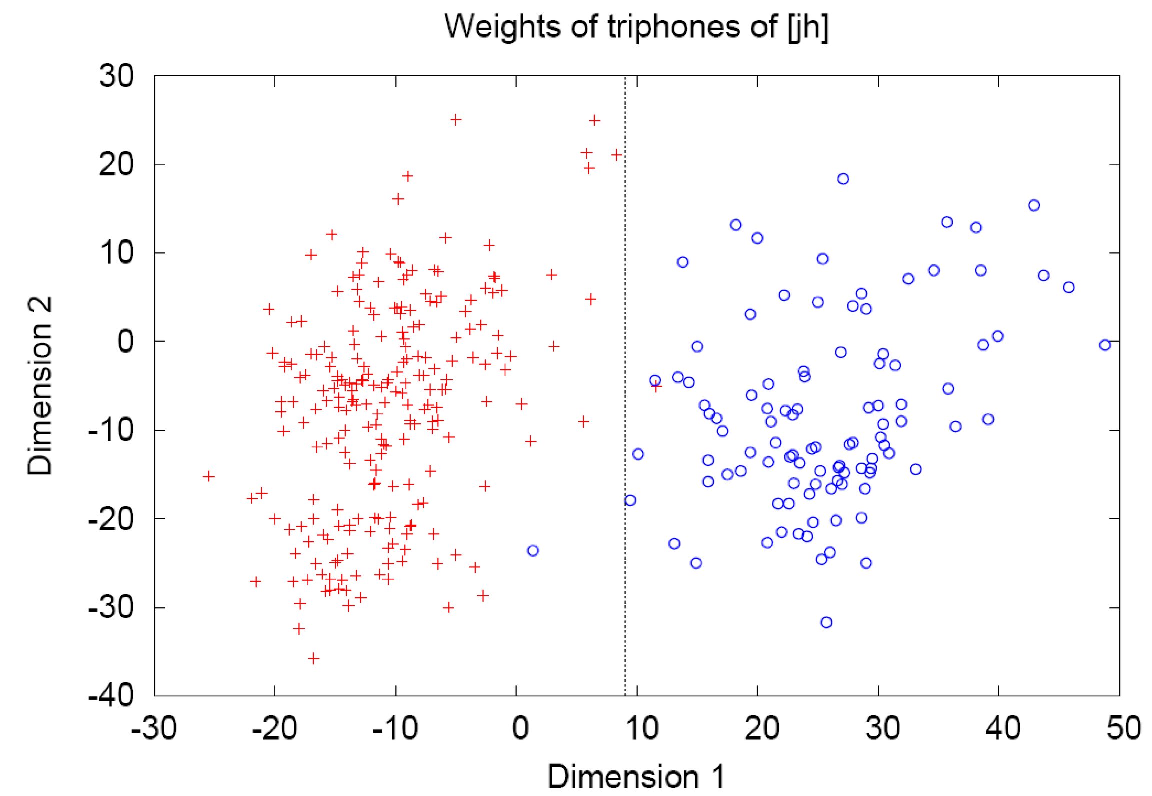

Fig. 5. The first 2 eigentriphone coefficients of all the triphones of [jh].

- 102 triphones lie to the right of a dotted vertical line; all of them except one have a consonant as their right context.
- 226 triphones lie to the left of the same dotted line; all except one have a vowel as their right context.

Conclusion & Future Work

- In this work, the Gaussian means of the infrequent triphones were "adapted" using the proposed eigentriphone framework. Experimental results show that our method performs slightly better than tied-state triphones.
- In the future, we would like to extend our method to Gaussian variances and mixture weights.
- Right now, all triphones of the same base phone use the same number of eigentriphones. An automatic way of deciding this number for each triphone depending on its amount of training samples is being investigated.
- The adaptation perspective of our new acoustic modeling method suggests that other adaptation algorithms could be investigated as well.
- The effect of discriminative training under the new eigentriphone framework will be investigated.
- The whole method is data-driven: no phonetic knowledge is required and the method can be generalized easily for other modeling units like syllables, and words.