A Fully Automated Derivation of State-based Eigentriphones for Triphone Modeling with No Tied States using Regularization

Brian Mak Tom Ko

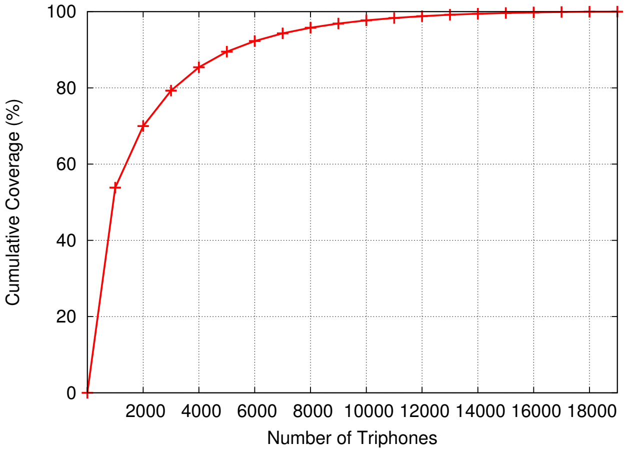
Department of Computer Science & Engineering
The Hong Kong University of Science and Technology
Hong Kong SAR, China

Interspeech-2011

Outline

1. Introduction
2. Motivation and Related Past Works
3. Eigentriphone acoustic modeling (ICASSP 2011)
4. Proposed improvements: Training with Regularized likelihood
5. Experimental evaluation
6. Summary

Vilfredo Pareto's 80/20 Principle



- WSJ0+WSJ1: 80% of samples are concentrated on the most frequent 20% of all seen triphones.
- How to train the infrequent triphones robustly?

Solution 1: Parameters Tying

Many HMM parameters may be tied:

- models: generalized triphones
- states: tied-state HMM
- Gaussians (mixtures) : TMHMM / SCHMM
- sub-vector Gaussians : SDCHMM
- means, covariances, weights

Solution 2: Model Interpolation

1. K. F. Lee on generalized triphones

$$\lambda \cdot \text{CD model} + (1 - \lambda) \cdot \text{CI model}$$

Solution 2: Model Interpolation

1. K. F. Lee on generalized triphones

$$\lambda \cdot \text{CD model} + (1-\lambda) \cdot \text{CI model}$$

2. Chang & Glass, "A back-off discriminative acoustic model for automatic speech recognition" (Interspeech 2009)

$$\lambda \cdot \text{CD model} + (1-\lambda) \cdot \text{broad-phonetic-class model}$$

Solution 2: Model Interpolation

1. K. F. Lee on generalized triphones

$$\lambda \cdot \text{CD model} + (1-\lambda) \cdot \text{CI model}$$

2. Chang & Glass, "A back-off discriminative acoustic model for automatic speech recognition" (Interspeech 2009)

$$\lambda \cdot \text{CD model} + (1-\lambda) \cdot \text{broad-phonetic-class model}$$

Side effect: acoustic score of each back-off CD model is distinct.

Solution 3: Basis Approach

1. Subspace Gaussian Mixture Model [Povey ..., ICASSP 2010]
   - A global basis for mixture i is used to derive the ith Gaussian mixture mean for each state j.

$$\mathbf{m}_{ji} = \mathbf{M}_i \mathbf{v}_j \quad (\text{state } j, \text{mixture } i)$$

Solution 3: Basis Approach

1. Subspace Gaussian Mixture Model [Povey ..., ICASSP 2010]
   - A global basis for mixture i is used to derive the ith Gaussian mixture mean for each state j.

$$\mathbf{m}_{ji} = \mathbf{M}_i \mathbf{v}_j \quad (\text{state } j, \text{mixture } i)$$

2. Bayesian Sensing HMM [Saon & Chien, ICASSP 2011]
   - A local basis for mixture i is used to derive the ith Gaussian mixture mean of state j.

$$\mathbf{m}_{ji} = \Phi_{ji} \mathbf{w}_t \quad (\text{state } j, \text{mixture } i, \text{time } t)$$

Solution 3: Basis Approach

1. Subspace Gaussian Mixture Model [Povey ..., ICASSP 2010]
   - A global basis for mixture i is used to derive the ith Gaussian mixture mean for each state j.

$$\mathbf{m}_{ji} = \mathbf{M}_i \mathbf{v}_j \quad (\text{state } j, \text{mixture } i)$$

2. Bayesian Sensing HMM [Saon & Chien, ICASSP 2011]
   - A local basis for mixture i is used to derive the ith Gaussian mixture mean of state j.

$$\mathbf{m}_{ji} = \Phi_{ji} \mathbf{w}_t \quad (\text{state } j, \text{mixture } i, \text{time } t)$$

3. Canonical State Model [Gales & Yu, Interspeech 2010]
   - A set of global (CI) canonical states:

$$s_g = \{..., \{c_g^{(m)}, \mathbf{m}_g^{(m)}, \Sigma_g^{(m)}\}, ...\}$$

   - A set of CD-state-dependent transforms:

$$\mathcal{T} = \{..., \{w_x^{(n)}, \theta_s^{(n)}\}, ...\}$$

   - CD state parameters are derived from some transformation of the canonical states parameters.

Proposed Solution: Eigentriphones (ICASSP 2011)

- No tied states!.
- Like the back-off acoustic model, the acoustic model scores are distinct.

Proposed Solution: Eigentriphones (ICASSP 2011)

- No tied states!.
- Like the back-off acoustic model, the acoustic model scores are distinct.
- A global basis for each base phoneme — eigentriphones.
- All triphones of a base phoneme are distinct points in the space of its eigentriphones.

Proposed Solution: Eigentriphones (ICASSP 2011)

- No tied states!.
- Like the back-off acoustic model, the acoustic model scores are distinct.
- A global basis for each base phoneme — eigentriphones.
- All triphones of a base phoneme are distinct points in the space of its eigentriphones.
- Acoustic modeling as an adaptation problem: derive the infrequent CD triphones using the Eigenvoice adaptation approach.

Eigentriphones vs. Eigenvoice

Item	Eigenvoice	Eigentriphone
No. of bases	1	39 (model-based)
		$3 \times 39 = 117$ (state-based)
Baseline model	SI model	CI model
Training models	SD models	frequent triphones models
Adaptation	new speaker; few data	infrequent triphones

For each base phoneme i:

1. Train a 3-state monophone HMM; M-mixture GMM states.

Model-based Eigentriphones Acoustic Modeling

For each base phoneme i:

1. Train a 3-state monophone HMM; M-mixture GMM states.
2. Initialize its triphones by cloning from the monophone HMM.

Model-based Eigentriphones Acoustic Modeling

For each base phoneme i:

1. Train a 3-state monophone HMM; M-mixture GMM states.
2. Initialize its triphones by cloning from the monophone HMM.
3. No state tying!

Model-based Eigentriphones Acoustic Modeling

For each base phoneme i:

1. Train a 3-state monophone HMM; M-mixture GMM states.
2. Initialize its triphones by cloning from the monophone HMM.
3. No state tying!
4. Two disjoint sets of triphones:

poor triphones: $\#\text{samples} < \theta_r = 200$

rich triphones: $\#\text{samples} \geq \theta_r = 200$

Model-based Eigentriphones Acoustic Modeling

For each base phoneme i:

1. Train a 3-state monophone HMM; M-mixture GMM states.
2. Initialize its triphones by cloning from the monophone HMM.
3. No state tying!
4. Two disjoint sets of triphones:

$$\text{poor triphones:} \quad \#\text{samples} < \theta_r = 200$$

$$\text{rich triphones:} \quad \#\text{samples} \geq \theta_r = 200$$

5. Re-estimate the Gaussian means only for the rich triphones.

Model-based Eigentriphones Acoustic Modeling

For each base phoneme i:

1. Train a 3-state monophone HMM; M-mixture GMM states.
2. Initialize its triphones by cloning from the monophone HMM.
3. No state tying!
4. Two disjoint sets of triphones:

$$\text{poor triphones:} \quad \#\text{samples} < \theta_r = 200$$

$$\text{rich triphones:} \quad \#\text{samples} \geq \theta_r = 200$$

5. Re-estimate the Gaussian means only for the rich triphones.
6. Gaussian covariances, mixture weights, transition probabilities are not reestimated.

Model-based Eigentriphones Acoustic Modeling

For each base phoneme i:

1. Train a 3-state monophone HMM; M-mixture GMM states.
2. Initialize its triphones by cloning from the monophone HMM.
3. No state tying!
4. Two disjoint sets of triphones:

$$\text{poor triphones:} \quad \#\text{samples} < \theta_r = 200$$

$$\text{rich triphones:} \quad \#\text{samples} \geq \theta_r = 200$$

5. Re-estimate the Gaussian means only for the rich triphones.
6. Gaussian covariances, mixture weights, transition probabilities are not reestimated.
7. "Eigentriphone adaptation" for Gaussian means.
(See next page)

Model-based Eigentriphones Acoustic Modeling

For each base phoneme i:

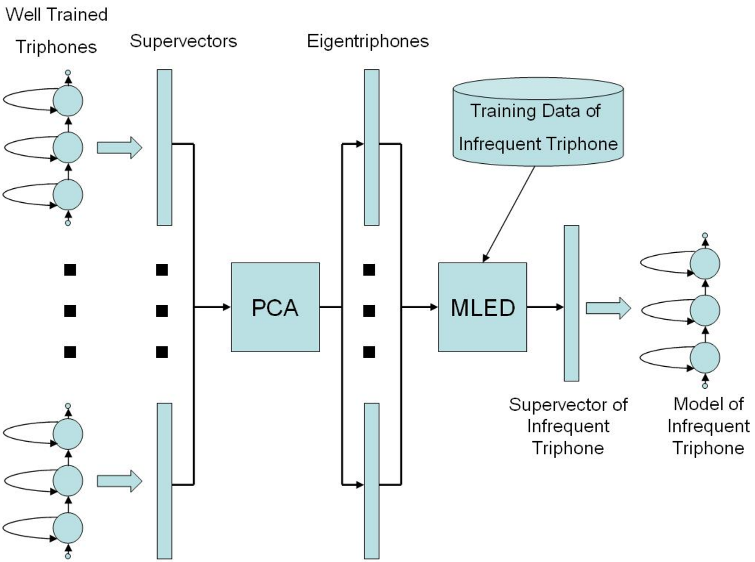
1. Train a 3-state monophone HMM; M-mixture GMM states.
2. Initialize its triphones by cloning from the monophone HMM.
3. No state tying!
4. Two disjoint sets of triphones:

$$\text{poor triphones:} \quad \#\text{samples} < \theta_r = 200$$

$$\text{rich triphones:} \quad \#\text{samples} \geq \theta_r = 200$$

5. Re-estimate the Gaussian means only for the rich triphones.
6. Gaussian covariances, mixture weights, transition probabilities are not reestimated.
7. "Eigentriphone adaptation" for Gaussian means.
   (See next page)
8. Re-estimate the other HMM parameters for the rich triphones.

Model-based Eigentriphones Acoustic Modeling ..



Gaussian mean of a poor triphone in the eigentriphone space is:

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

where

- i : base phoneme index
- p : triphone index
- $\mathbf{v}_{ip}$: supervector for the Gaussian means of p
- $\mathbf{e}_{ik}$: kth largest eigenvector in the basis of phoneme i
- w_{ipk} : kth weight of triphone p
- $\mathbf{m}_i$: supervector for the Gaussian means of monophone i

#Eigenvectors for Different Variations Coverage

Base Phone	100%	80%	60%	40%	20%
t	535	146	51	11	2
d	468	150	58	13	3
s	451	107	32	8	2
n	446	124	41	8	2
ah	434	100	26	7	2
er	411	127	46	10	2
l	390	120	41	7	1
z	382	101	33	9	3
iy	379	100	32	7	2
k	365	95	28	7	2

Improvement 1: Regularized Optimization

Before : Objective = max log likelihood of training data

$$Q(\mathbf{w}_{ip}) = L(\mathbf{w}_{ip})$$

Improvement 1: Regularized Optimization

Before : Objective = max log likelihood of training data

$$Q(\mathbf{w}_{ip}) = L(\mathbf{w}_{ip})$$

Now : Objective = max penalized log likelihood of training data

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

- Penalty function:

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

Improvement 1: Regularized Optimization

Before : Objective = max log likelihood of training data

$$Q(\mathbf{w}_{ip}) = L(\mathbf{w}_{ip})$$

Now : Objective = max penalized log likelihood of training data

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

- Penalty function:

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

- Now each triphone of a base phoneme may use a different number of eigentriphones.
- In general, favor small weights.
- For triphone with few data, back-off to the monophone means.
- Try to de-emphasize eigentriphones with small eigenvalues.

The poor triphones set and the rich triphones set now overlaps.

poor triphones: $\#\text{samples} \leq \theta_m^P = 200$

rich triphones: $\#\text{samples} \geq \theta_m^R = 30$

Improvement 3: State-based Eigentriphones

An eigenspace for each of the 3 states of the triphones

$\Rightarrow 3 \times 39 = 117$ bases

Improvement 4: More Detailed Control on Parameter Estimation Thresholds

θ_m^P	poor triphone threshold	200
θ_m^R	rich triphone threshold mean reestimation threshold	30
θ_v^R	variance reestimation threshold	200
θ_w^R	mixture weight reestimation threshold	30
θ_t^R	transition reestimation threshold	200

Evaluation on 5K WSJ

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

- Feature Extraction
  - 10ms frames; 25ms window
  - standard 39-dimensional MFCC acoustic vectors.
- Acoustic Models
  - 18,777 cross-word triphones CDHMM derived from 39 base phonemes; 6,481 tied states
  - left-to-right 3-state HMMs; 16 Gaussian components / state
- Language Model: bigram, PP $= \sim 110$; trigram, PP $= \sim 60$.
- $\beta = 15$ (empirically determined)

Bigram Result: Tied-state Triphones vs. Eigentriphones

Model	Description	Nov'92
baseline1	tied-state triphones	94.56%

Bigram Result: Tied-state Triphones vs. Eigentriphones

Model	Description	Nov'92
baseline1	tied-state triphones	94.56%
baseline2	no state tying; HMM parameters are re-estimated according to the thresholds: $\theta_m^R, \theta_v^R, \theta_w^R, \theta_t^R$	93.98%

Bigram Result: Tied-state Triphones vs. Eigentriphones

Model	Description	Nov'92
baseline1	tied-state triphones	94.56%
baseline2	no state tying; HMM parameters are re-estimated according to the thresholds: $\theta_m^R, \theta_v^R, \theta_w^R, \theta_t^R$	93.98%
baseline3	no state tying; only Gaussian means of rich triphones are re-estimated	93.50%

Bigram Result: Tied-state Triphones vs. Eigentriphones

Model	Description	Nov'92
baseline1	tied-state triphones	94.56%
baseline2	no state tying; HMM parameters are re-estimated according to the thresholds: $\theta_m^R, \theta_v^R, \theta_w^R, \theta_t^R$	93.98%
baseline3	no state tying; only Gaussian means of rich triphones are re-estimated	93.50%
	+ state-based eigentriphone "adaptation" of means for poor triphones	93.78%

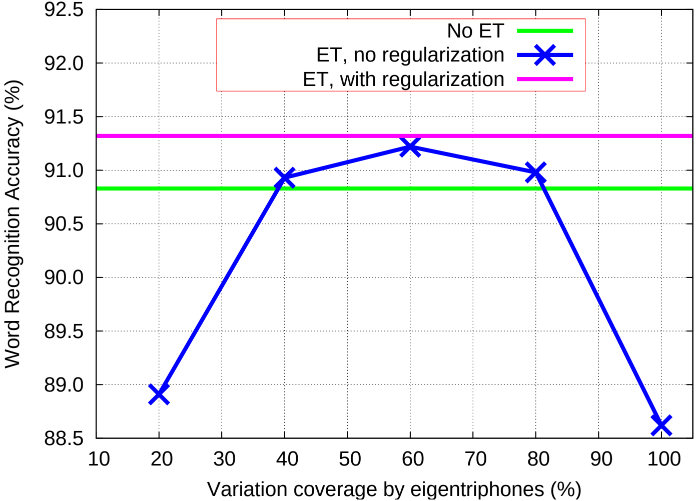
Bigram Result: Tied-state Triphones vs. Eigentriphones

Model	Description	Nov'92
baseline1	tied-state triphones	94.56%
baseline2	no state tying; HMM parameters are re-estimated according to the thresholds: $\theta^R_m, \theta^R_v, \theta^R_w, \theta^R_t$	93.98%
baseline3	no state tying; only Gaussian means of rich triphones are re-estimated	93.50%
	+ state-based eigentriphone "adaptation" of means for poor triphones	93.78%
	+ remaining HMM parameters are re-estimated according to the thresholds: $\theta^R_v, \theta^R_w, \theta^R_t$	94.53%

Trigram Result: State-based vs. Model-based Eigentriphones

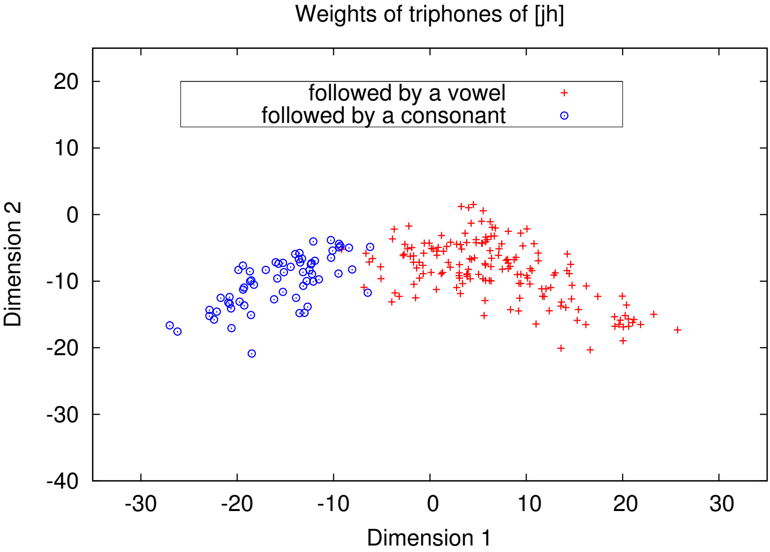
System	Nov'92	Nov'93
tied-state triphone system	96.45%	93.89%
state-based eigentriphone system	96.41%	94.47%
model-based eigentriphone system	96.47%	94.44%

Result: Effect of Regularization (bigram, nov'93)



- Note: Just after "eigentriphone adaptation" of the Gaussian meas; no further re-estimation of other HMM parameters.

Analysis: Triphones of [jh]



Summary and Conclusions

- The expanded set of rich triphones give better results.
- The use of regularization improves performance by avoiding a hard decision on the number of eigentriphones (eigenvectors) for each triphone of the same base phoneme.
- Model-based eigentriphones are preferred over state-based eigentriphones for simplicity since both give similar performance.
- Tied states are not necessary.
- Triphones trained using the eigentriphone approach are mostly distinct.