Latent Correlation Analysis of HMM Parameters For Speech Recognition

Zhijian Ou, Jun Luo

ozj@tsinghua.edu.cn

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Objective: Exploit correlation between HMM parameters, eigenspace estimation

Method: Treat supervector as a latent variable under HMM
Estimate the hidden supervector's cov matrix directly from the frames using EM

Advantage: theoretically sound, dealing well with speaker-specific data sparseness

Why study parameter correlation?

correlation between different sounds

as a consequence of the slowly changing characteristics of some underlying factors (e.g. the speaker, speaking style, emotional state)

correlation between model parameters (Gaussian means)

Supervector

For speaker adaptation

- As a priori information about the inter-speaker variation
- Used to derive constraints for rapid speaker adaptation

Eigenvoice approach review

Attractive: based on the supervector's covariance matrix

clearly a good measure of parameter correlation

estimated simply as the sample covariance matrix from a set of training speaker supervectors

PCA: obtain the dominant eigenvectors, i.e. eigenvoices $e_1,\ldots,e_R$

Constraint

$$x = \mu + \sum_{r=1}^{R} w_r \cdot e_r$$

MAP eigenvoice adaptation: estimate weights $w_{1:R}$

$$\hat{w}_{1:R} = \arg\max_{w_{1:R}} \left[p\left(y_{1:T} \mid w_{1:R}\right) p_0\left(w_{1:R}\right) \right]$$

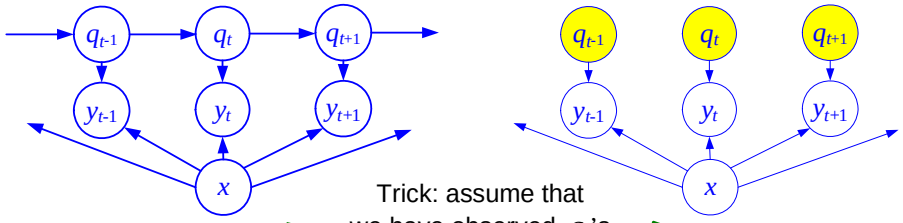
Assume: observable supervector

Assume: we have a set of well-trained speaker supervectors as its samples/observations

less sufficient speaker-specific data (may having unseen phones)**?** → resort to MLLR adaptation, EMAP, etc. to create the SD models

Latent Correlation Analysis of HMM Parameters

Bayesian network representation of the generative model of speech, incorporating the supervector variable x



Trick: assume that we have observed q_t's

$$p(x) = \mathcal{N}(x \mid \mu, \Sigma)$$

Essentially translated to do parameter estimation with $p(y_t \mid x, q_t = i) = \mathcal{N}(y_t \mid x_i, C_i)$

E-step: compute the posterior $p\left(x^{(n)} \middle| \bar{\theta}, y^{(n)}_{1:T_n}, q^{(n)}_{1:T_n}\right) \sim N\left(\mu^{(n)post}, \Sigma^{(n)post}\right)$

M-step: re-estimate $\hat{\mu} = \mu^{post}$ $\quad \hat{\Sigma} = \Sigma^{post} + \frac{1}{N}\sum_n \Sigma^{(n)post}$

Discussion

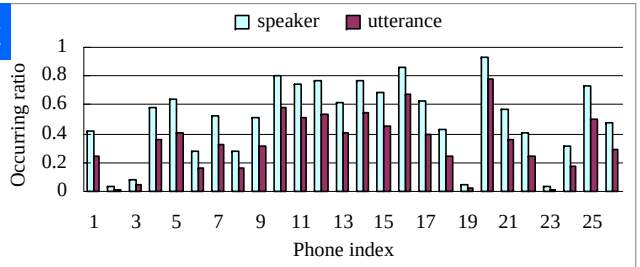
- Makes clear the deficiency of traditional methods
- Compare with Cluster Adaptive Training, etc.
  - Assume that we know the desirable number of basis vectors beforehand
  - The resulting basis vectors are not guaranteed to be orthogonal
  - No corresponding eigenvalues
- To conduct utterance-level correlation analysis, estimate utterance eigenvoices, and perform (unsupervised) utterance adaptation

Experiment Result

- OGI Numbers 30-word vocabulary
- Training: 6049 utterances spoken by 3059 speakers
- Test: 2061 utterances spoken by 1044 speakers
- 39-dim feature (12 MFCCs, Energy)+Δ+ΔΔ
- 26 monophone+sil+pause
- A speaker observes only 50.5% of the 26 phones
- An utterance observes fewer phones with 33.7%

EM+EV outperform MAP, MLLR and MLLR+EV consistently.

Utterance adaptation is useful.



Mixture num per state		1	2	4
Baseline		20.86	16.85	13.34
Speaker adaptation	MLLR	20.71	16.79	13.25
	MAP	20.75	16.83	13.32
	MLLR+EV	20.79	16.27	12.59
	EM+EV	**18.42**	**15.76**	**12.44**
Utterance adaptation	MLLR	20.71	16.80	13.29
	MAP	20.75	16.86	13.24
	MLLR+EV	20.81	16.62	13.20
	EM+EV	**18.31**	**15.20**	**11.97**