Speaker Adaptation

Lecture 21

CS 753

Instructor: Preethi Jyothi

Speaker variations

- Major cause of variability in speech is the differences between speakers
  - Speaking styles, accents, gender, physiological differences, etc.
- Speaker independent (SI) systems: Treat speech from all different speakers as though it came from one and train acoustic models
- Speaker dependent (SD) systems: Train models on data from a single speaker
- Speaker adaptation (SA): Start with an SI system and adapt using a small amount of SD training data

Modes of speaker adaptation

- Batch/Incremental adaptation: User supplies adaptation speech beforehand vs. system makes use of speech collected as the user uses a system

- Supervised/Unsupervised adaptation: Knowing transcriptions for the adaptation speech vs. not knowing them

Types of speaker adaptation

- Training/Normalization: Modify only parameters of the models observed in the adaptation speech vs. find transformation for all models to reduce cross-speaker variation

- Feature/Model transformation: Modify the input feature vectors vs. modifying the model parameters.

Speaker adaptation

- Speaker adaptation techniques can be grouped into three families:

    1. Feature-based approaches

    2. Maximum a posterior (MAP) adaptation

    3. Linear transform-based adaptation

Speaker adaptation

- Speaker adaptation techniques can be grouped into three families:

    1. Feature-based approaches
    2. Maximum a posterior (MAP) adaptation
    3. Linear transform-based adaptation

Normalization

- Cepstral mean and variance normalization: Effectively reduce variations due to channel distortions

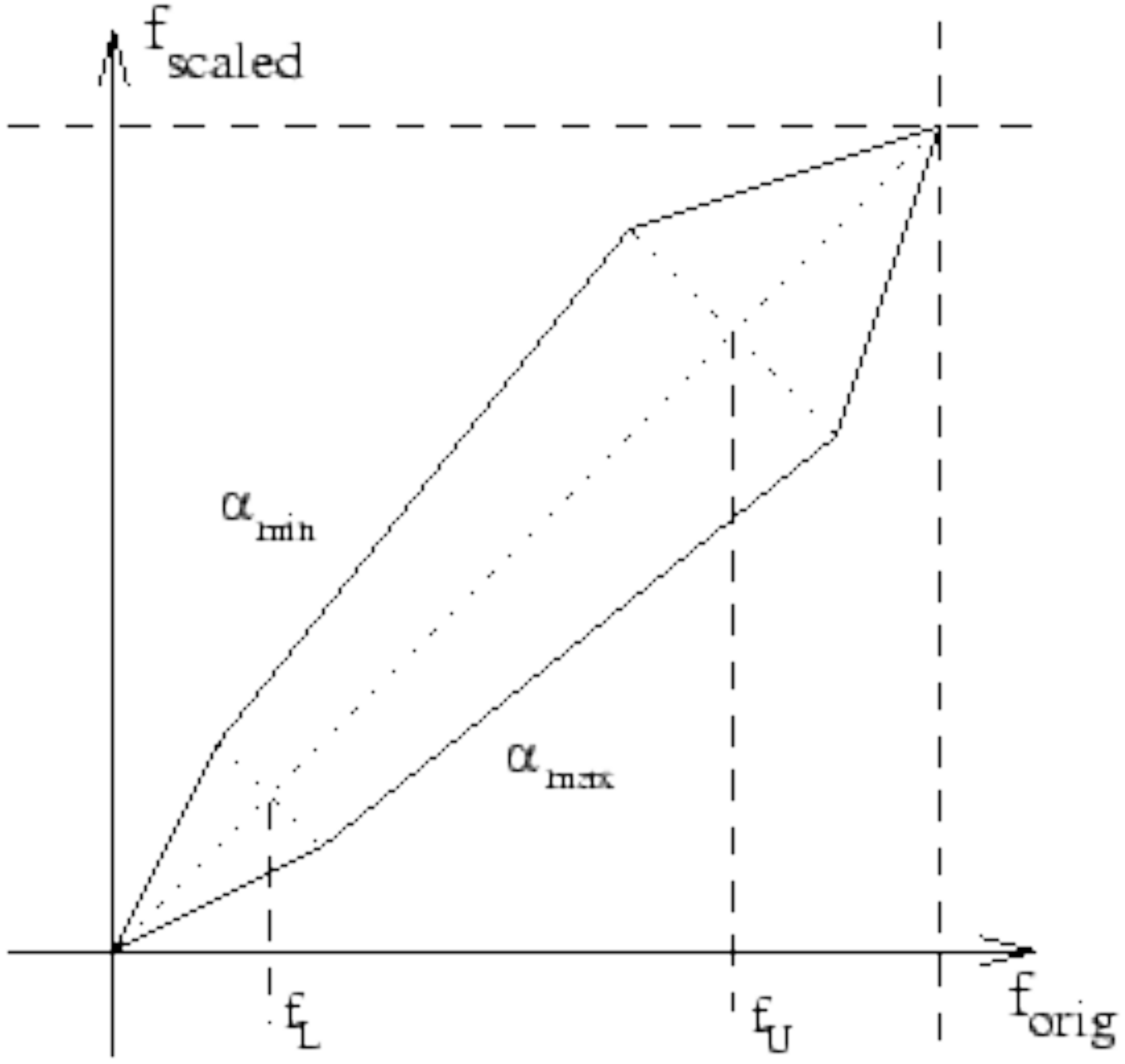
$$\mu_f = \frac{1}{T}\sum_t f_t$$

$$\sigma_f{}^2 = \frac{1}{T}\sum_t (f_t^2 - \mu_{f,t}^2)$$

$$\hat{f}_t = \frac{f_t - \mu_f}{\sigma_f}$$

- Mean subtracted from the cepstral features to nullify the channel characteristics

Vocal Tract Length Normalization (VTLN)



- VTLN is implemented by warping the frequency axis in the filterbank analysis

Image from: HTK Book, http://www1.icsi.berkeley.edu/Speech/docs/HTKBook3.2/node63_mn.html

Speaker adaptation

- Speaker adaptation techniques can be grouped into three families:

  1. Feature-based approaches

  2. Maximum a posterior (MAP) adaptation

  3. Linear transform-based adaptation

Maximum a posteriori adaptation

- Let λ characterise the parameters of an HMM and Pr(λ) be prior knowledge. For observed data X, the maximum a posterior (MAP) estimate is defined as:

$$\begin{aligned}\lambda^* &= \arg\max_{\lambda} Pr(\lambda|X) \\ &= \arg\max_{\lambda} Pr(X|\lambda) \cdot Pr(\lambda)\end{aligned}$$

- If Pr(λ) is uniform, then MAP estimate is the same as the maximum likelihood (ML) estimate

Recall: ML estimation of GMM parameters

ML estimate:

$$\mu_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m) x_t}{\sum_{t=1}^{T} \gamma_t(j,m)}$$

- where $\gamma_t(j, m)$ is the probability of occupying mixture component m of state j at time t

MAP estimation

ML estimate:

$$\mu_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m) x_t}{\sum_{t=1}^{T} \gamma_t(j,m)}$$

- where $\gamma_t(j, m)$ is the probability of occupying mixture component m of state j at time t

MAP estimate:

$$\hat{\mu}_{jm} = \frac{\tau \mu_{jm} + \sum_t \gamma_t(j,m) x_t}{\tau + \sum_t \gamma_t(j,m)}$$

- where μ_{jm} is prior mean chosen from previous EM iteration, τ controls the bias between prior and information from the adaptation data

MAP estimation

- MAP estimate is derived after 1) choosing a specific prior distribution for $\lambda = (c_1,\ldots,c_m, \mu_1,\ldots,\mu_m, \Sigma_1,\ldots,\Sigma_m)$ 2) updating model parameters using EM
- Property of MAP: Asymptotically converges to ML estimate as the amount of adaptation data increases
- Updates only those parameters which are observed in the adaptation data

Speaker adaptation

- Speaker adaptation techniques can be grouped into three families:

  1. Feature-based approaches

  2. Maximum a posterior (MAP) adaptation

  3. Linear transform-based adaptation

Linear transform-based adaptation

- Estimate a linear transform from the adaptation data to modify HMM parameters
- Estimate transformations for each HMM parameter? Would require very large amounts of training data.
    - Tie several HMM states and estimate one transform for all tied parameters
    - Could also estimate a single transform for all the model parameters
- Main approach: Maximum Likelihood Linear Regression (MLLR)

MLLR

- In MLLR, the mean of the m-th Gaussian mixture component μ_m is adapted in the following form:

$$\hat{\mu}_m = A\mu_m + b = W\xi_m$$

where $\hat{\mu}_m$ is the adapted mean, $W = [A, b]$ is the linear transform and ξ_m is the extended mean vector, $[\mu_m^T, 1]^T$

- W is estimated by maximising the likelihood of the adaptation data X:

$$W^* = \arg\max_W \{\log \Pr(X; \lambda, W)\}$$

- EM algorithm is used to derive this ML estimate

Regression classes

- So far, assumed that all Gaussian components are tied to a global transform
- Untie the global transform: Cluster Gaussian components into groups and each group is associated with a different transform
- E.g. group the components based on phonetic knowledge
  - Broad phone classes: silence, vowels, nasals, stops, etc.
  - Could build a decision tree to determine clusters of components

Speaker adaptation of NN-based models

- Approach analogous to MAP for GMMs: Can we update the weights of the network using adaptation speech data from a target speaker?
  - Limitation: Typically, too many parameters to update!
- Can we feed the network untransformed features and let the network figure out how to do speaker normalisation?
  - Along with untransformed features that capture content (e.g. MFCCs), also include features that characterise the speaker.
  - i-vectors are a popular representation which captures all relevant information about a speaker.

i-vectors

- Acoustic features from all the speakers (x_t) are seen as being generated from a Universal Background Model (UBM) which is a GMM with M diagonal co-variance matrices

$$x_t \sim \sum_{m=1}^{M} c_m \mathcal{N}(\mu_m, \Sigma_m)$$

- Let U_0 denote the UBM supervector which is the concatenation of μ_m for m = 1, … , M. Let U_s denote the mean supervector for a speaker s, which is the concatenation of speaker-adapted GMM means $\mu_m(s)$ for m = 1, … , M for the speaker s. The i-vector model is:

$$\mathrm{U}_s = \mathrm{U}_0 + \mathrm{V} \cdot v(s)$$

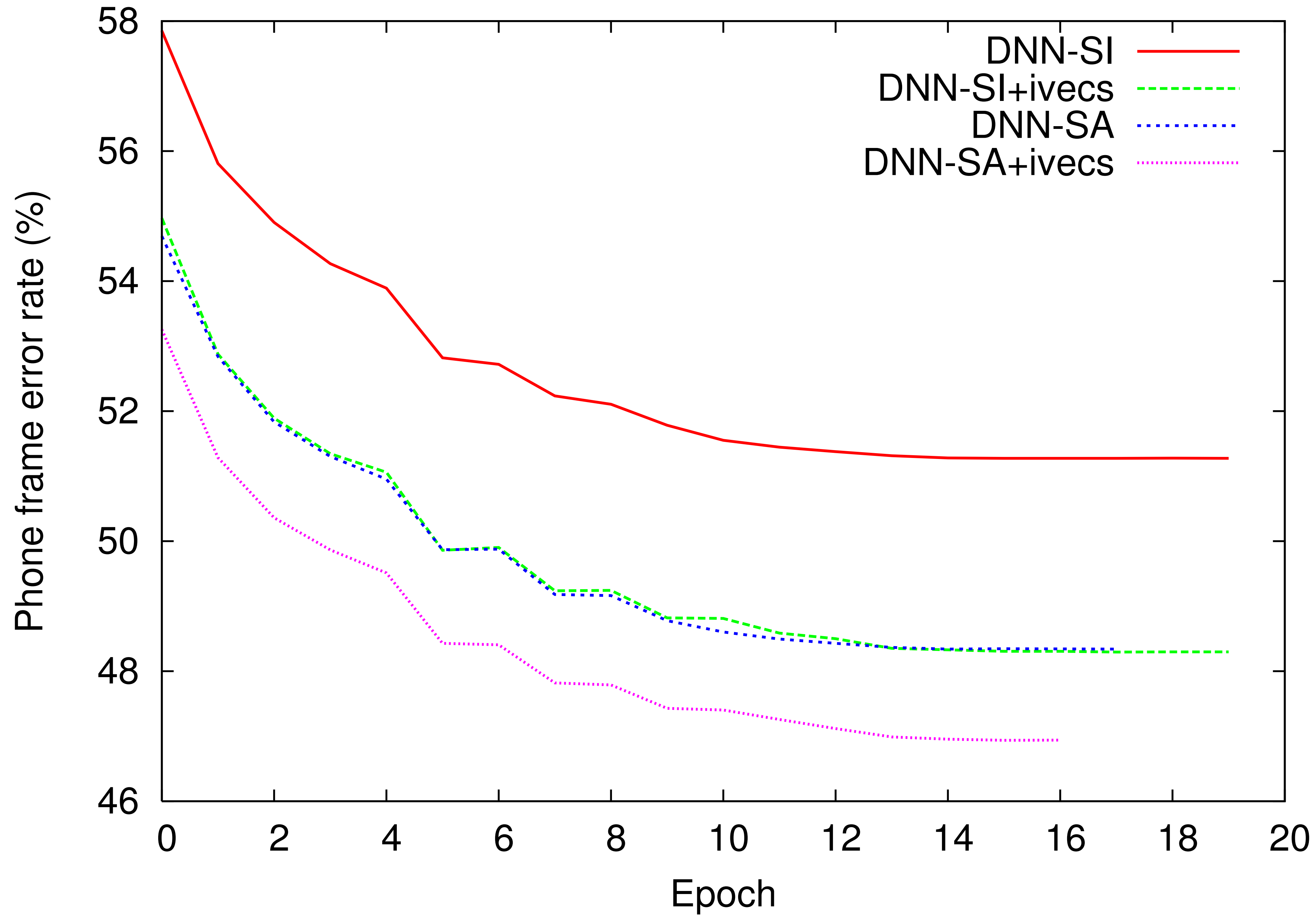
where V is the total variability matrix of dimensionality $\mathrm{M{\cdot}F} \times K$,
$v(s)$ is the ***i-vector*** of dimension K.

i-vectors

$$\mathrm{U}_s = \mathrm{U}_0 + \mathrm{V} \cdot v(s)$$

- Given adaptation data for a speaker s, how do we estimate V? How do we further estimate $v(s)$?
  - EM algorithm to the rescue.
- i-vectors are estimated by iterating between the estimation of the posterior distribution $P(v(s) \mid X(s))$ (where $X(s)$ denotes speech from speaker s) and update of the total variability matrix V.

ASR improvements with i-vectors



Model	Training	Hub5'00 SWB	RT'03 FSH	RT'03 SWB
DNN-SI	x-entropy	16.1%	18.9%	29.0%
DNN-SI	sequence	14.1%	16.9%	26.5%
DNN-SI+ivecs	x-entropy	13.9%	16.7%	25.8%
DNN-SI+ivecs	sequence	12.4%	15.0%	24.0%
DNN-SA	x-entropy	14.1%	16.6%	25.2%
DNN-SA	sequence	12.5%	15.1%	23.7%
DNN-SA+ivecs	x-entropy	13.2%	15.5%	23.7%
DNN-SA+ivecs	sequence	11.9%	14.1%	22.3%

Image from: Saon et al.,Speaker Adaptation of Neural Network Acoustic Models Using I-Vectors, ASRU 13