Speaker Adaptation

Speaker Adaptation

- Introduction: speaker-specific variation, modes of adaptation
  - Speaker normalization: VTL
  - Model-based adaptation: MAP
  - Model-based adaptation: MLLR
  - Model-based adaptation: Speaker space models

Thanks to Steve Renals, Univ. of Edinburgh for these slides

Speaker independent / dependent / adaptive

- **Speaker independent (SI)** systems have long been the focus for research in transcription, dialogue systems, etc.
- **Speaker dependent (SD)** systems can result in word error rates 2–3 times lower than SI systems (given the same amount of training data)
- **Speaker adaptive (SA)** systems... we would like
  - Error rates similar to SD systems
  - Building on an SI systems
  - Requiring only a small fraction of the speaker-specific training data used by an SD system

Thanks to Steve Renals, Univ. of Edinburgh for these slides

Speaker-specific variation

- **Acoustic model**
  - Speaking styles
  - Accents
  - Speech production anatomy (eg length of the vocal tract)
    - Also non-speaker variation, such as channel conditions (telephone, reverberant room, close talking mic) and application domain
  - Speaker adaptation of acoustic models aims to reduce the mismatch between test data and the models
- **Pronunciation model**: speaker-specific, consistent change in pronunciation
- **Language model**: user-specific documents (exploited in personal dictation systems)

Thanks to Steve Renals, Univ. of Edinburgh for these slides

Modes of Adaptation

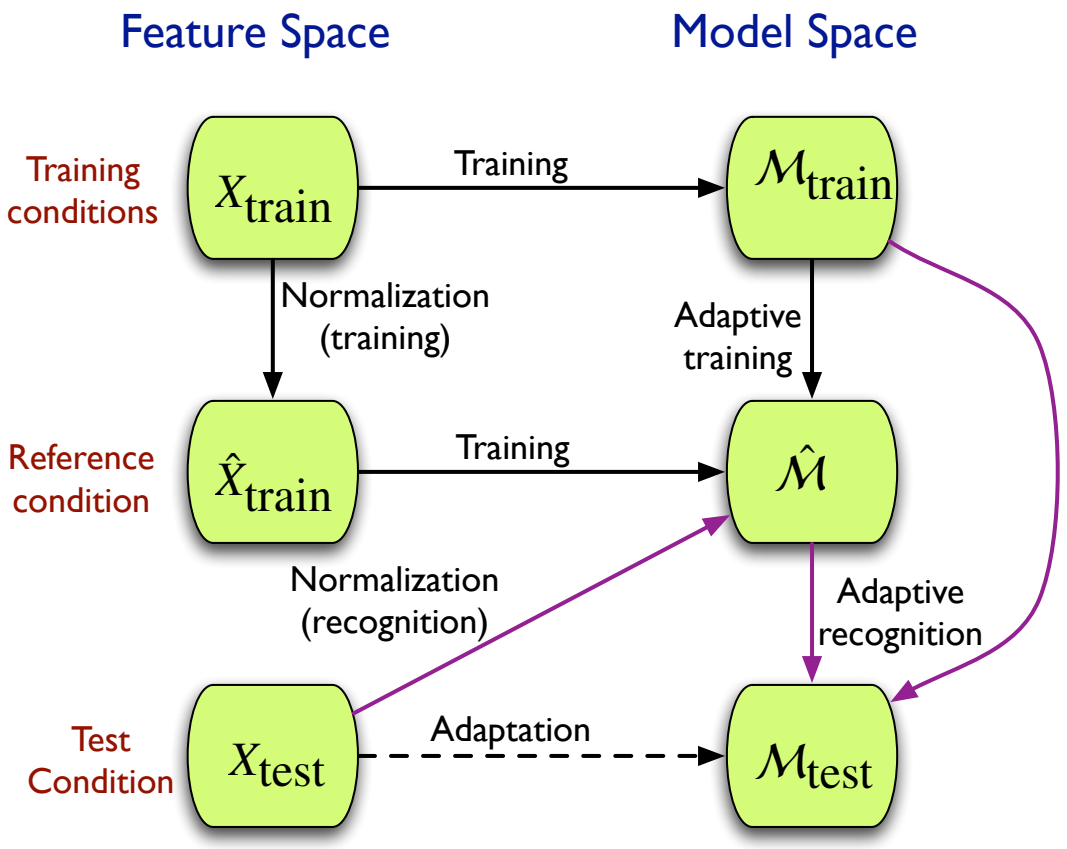
- **Supervised or unsupervised**
  - Supervised: the word level transcription of the adaptation data is known (and HMMs may be constructed)
  - Unsupervised: the transcription must be estimated (eg using recognition output)
- **Static or dynamic**
  - Static: All adaptation data is presented to the system in a block before the final system is estimated (eg as used in enrollment in a dictation system)
  - Dynamic: Adaptation data is incrementally available, and models must be adapted before all adaptation data is available (eg as used in a spoken dialogue system)

Thanks to Steve Renals, Univ. of Edinburgh for these slides

Approaches to adaptation

- **Speaker Normalization**: Normalize the acoustic data to reduce mismatch with the acoustic models
  - Vocal Tract Length Normalization (VTLN)
- **Model based**: Adapt the parameters of the acoustic models to better match the observed data
  - Maximum a posteriori (MAP) adaptation of HMM/GMM parameters
  - Maximum likelihood linear regression (MLLR) of Gaussian parameters
  - Feature-based Maximum likelihood linear regression (fMLLR) of feature vectors
- **Speaker space:** Estimate multiple sets of acoustic models, characterizing new speakers in terms of these model sets
  - Cluster-adaptive training
  - Eigenvoices

Adaptation and normalization of acoustic models



Vocal Tract Length Normalization (VTLN)

- **Basic idea**: Normalize the acoustic data to take account of changes in vocal tract length
- Vocal tract length (VTL):
  - First larynx descent in first 2-3 years of life
  - VTL grows according to body size, and is sex-dependent
  - Puberty: second larynx descent for males
- VTL has large effect on the spectrum
  - Tube acoustic model: formant positions are inversely proportional to VTL
  - Observation: formant frequencies for women are 20% higher than for men (on average)
- **VTLN**: compensate for differences between speakers via a warping of the frequency axis

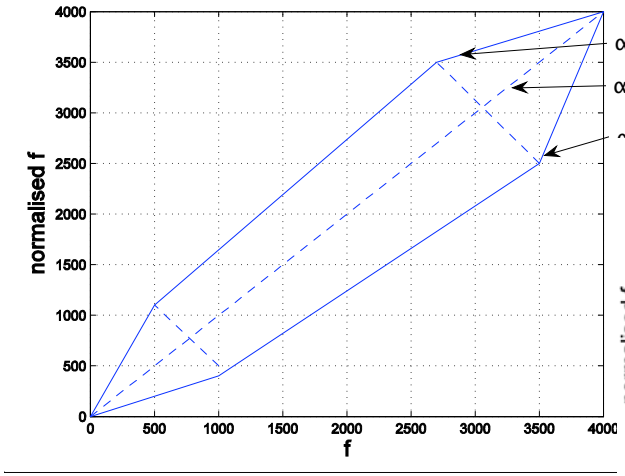
Approaches to VTLN

- Classify by frequency warping function
  - Piecewise linear
  - Power function
  - Bilinear transform

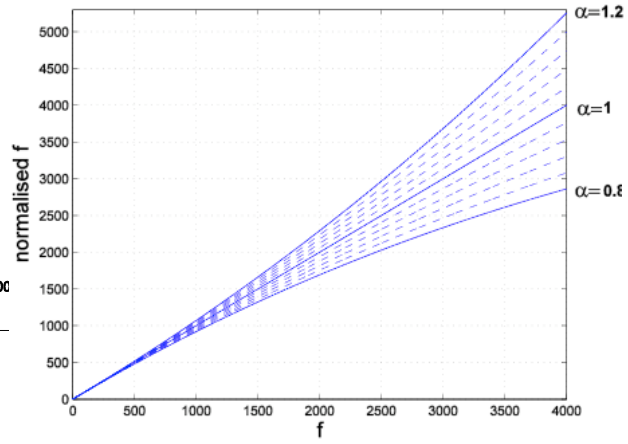
$$f \rightarrow \hat{f} = g_\alpha(f)$$

- Classify by estimation of warping factor α
  - Signal-based: estimated directly from the acoustic signal, through explicit estimation of formant positions
  - Model-based: maximize the likelihood of the observed data given acoustic models and a transcription. α is another parameter set so as to maximize the likelihood

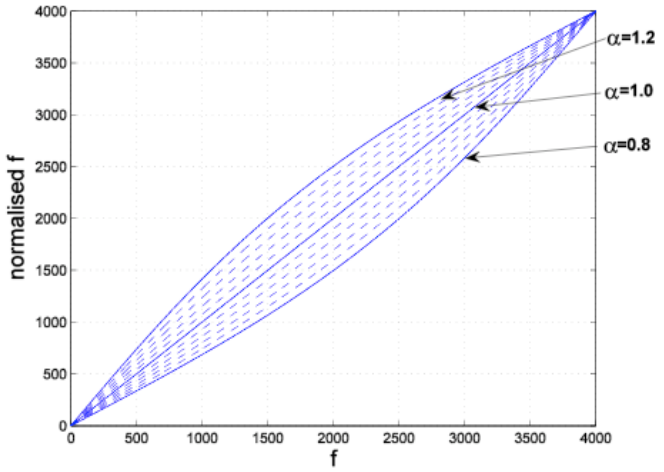
+ Warping functions



$$\hat{f} = \alpha f$$



$$\hat{f} = \alpha^{3f/8000} f$$



$$\hat{f} = f + \arctan \frac{(1-\alpha)\sin f}{1-(1-\alpha)\cos f}$$

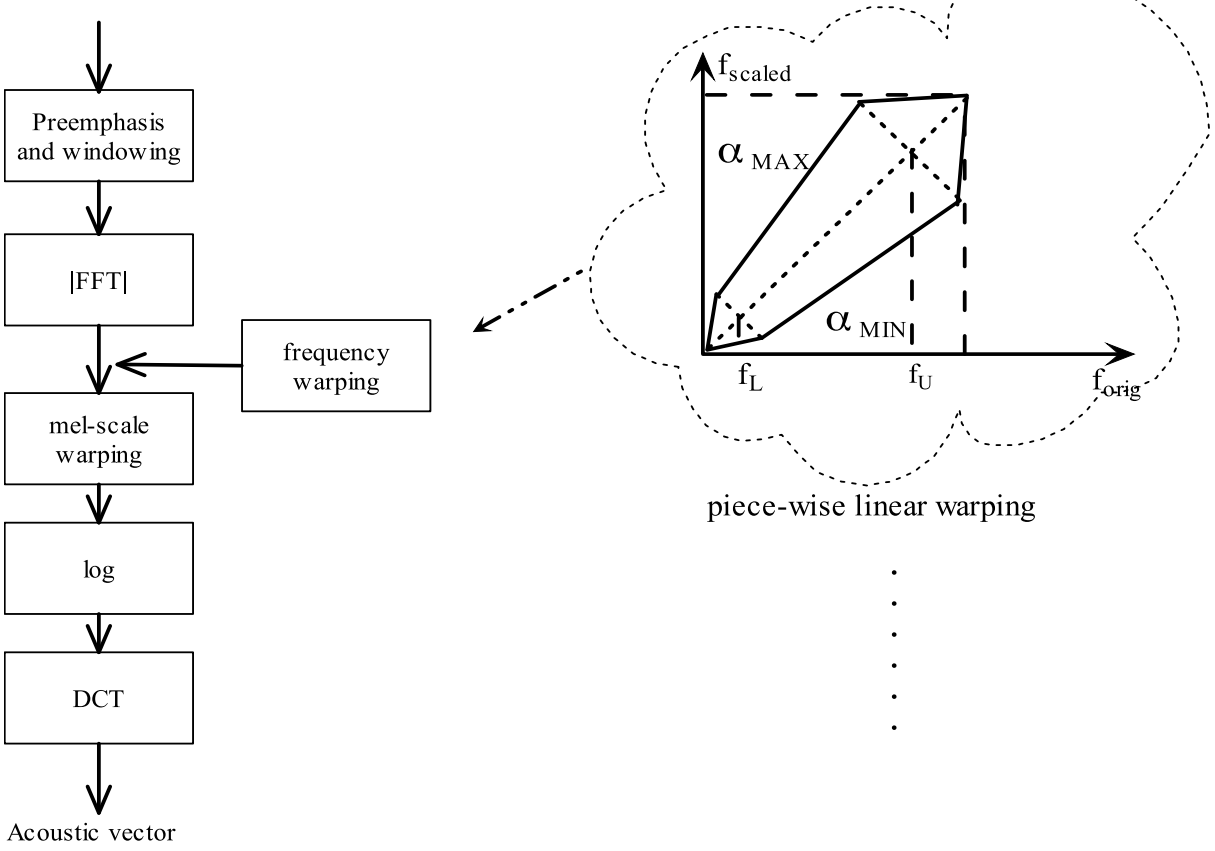
+ Approaches

- **Signal-based VTLN**
  - Basic idea: Estimate the warping factor from the signal without using the speech recognition models
  - Estimate warping factor α from formant positions: eg Eide and Gish (1996) used ratio of median position of 3rd formant for speaker s ($\bar{F}_{3,s}$) to the median for all speakers ($\bar{F}_3$):

$$\alpha_s = \frac{\bar{F}_{3,s}}{\bar{F}_3}$$

- **Model based VTLN**
  - Basic idea: Warp the acoustic features (for a speaker) to better fit the models — rather than warping the models to fit the features!
  - Estimate the warping factor α so as to maximise the likelihood of the acoustic models
  - After estimating the warp factors, normalize the acoustic data and re-estimate the models

+ Model-based VTLN



VTLN: WER (%) on conversational telephone speech

	Tot	Sub	Del	Ins	F	M
No adapt	37.2	24.2	8.8	4.2	36.7	37.6
Test only	36.4	23.6	8.5	4.3	36.1	36.7
1 pass	35.7	22.9	8.9	3.8	35.0	36.4
2 pass	35.0	22.5	8.8	3.7	34.2	35.8
3 pass	34.5	22.0	8.7	3.7	33.6	35.3
4 pass	34.2	22.0	8.6	3.6 ˜	33.3	35.1

- 7–10% relative decrease in WER is typical for VTLN
- VTLN removes the need for gender-dependent acoustic models

Model-based adaptation: The MAP family

- Basic idea: Use the SI models as a prior probability distribution over model parameters when estimating using speaker-specific data
  - Theoretically well-motivated approach to incorporating the knowledge inherent in the SI model parameters
- If the parameters of the models are denoted λ, then maximum likelihood (ML) training chooses them to maximize $p(X \mid \lambda)$
  - Maximum a posteriori (MAP) training maximizes:

    $$p(\lambda \mid X) \propto p(X \mid \lambda)p_0(\lambda)$$

  - $p_0(\lambda)$ is the prior distribution of the parameters
- The use of a prior distribution, based on the SI models, means that less data is required to estimate the speaker-specific models: we are not starting from complete ignorance

Local estmation vs. Structural MAP

- The main drawback to MAP adaptation is that it is local
  - Only the parameters belonging to Gaussians of observed states will be adapted
  - Large vocabulary speech recognition systems have about 105 Gaussians: most will not be adapted
- Structural MAP (SMAP) approaches have been introduced to share Gaussians
  - Share Gaussians by organizing them in a tree, whose root contains all the Gaussians
  - At each node in the tree compute mean offset and diagonal variance scaling term
  - For each node, its parent is used as a prior distribution
- The MLLR family of adaptation approaches addresses this by assuming that transformations for a specific speaker are systematic across Gaussians, states and models

MAP adaptation uses

- MAP adaptation is very useful for domain adaptation:
  - Example: MAP adapting a conversational telephone speech system (100s of hours of data) to multiparty meetings (10s of hours of data) works well with MAP
- As the amount of training data increases, so the MAP estimate converges to the ML estimate

The Linear Transform family

- Basic idea Rather than directly adapting the model parameters, estimate a transform which may be applied the Gaussian means and covariances
- Linear transform applied to parameters of a set of Gaussians: adaptation transform parameters are shared across Gaussians
- This addresses the locality problem arising in MAP adaptation, since each adaptation data point can affect many of (or even all) the Gaussians in the system
- There are relatively few adaptation parameters, so estimation is robust

MLLR: Maximum Likelihood Linear Regression

- MLLR is the best known linear transform approach to speaker adaptation
- Affine transform of mean parameters

$$\hat{\mu} = A\mu + b$$

- If the observation vectors are d -dimension, then **A** is a d × d matrix and **b** is d-dimension vector
- If we define **W** = [**bA**] and $\eta = [1 \mu T]T$, then we can write:

$$\hat{\mu} = W\eta$$

- In MLLR, W is estimated so as to maximize the likelihood of the adaptation data
- A single transform W can be shared across a set of Gaussian components (even all of them!)

+ MLLR

- Regression Classes
  - The number of transforms may obtained automatically
  - A set of Gaussian components that share a transform is called a **regression class**
  - Obtain the regression classes by constructing a regression class tree
  - Each node in the tree represents a regression class sharing a transform
- Estimating the Transforms
  - The linear transformation matrix W is obtained by finding its setting which optimizes the log likelihood
  - Log likelihood $$L = \sum_r \sum_n \gamma_r(n) \log \left(K_r \exp \left(-\frac{1}{2} (\mathbf{x}_n - \mathbf{W}\boldsymbol{\eta}_r)^T \boldsymbol{\Sigma}_r^{-1} (\mathbf{x}_n - \mathbf{W}\boldsymbol{\eta}_r) \right) \right)$$
  - where r ranges over the components belonging to the regression class

MLLR in practice

- Mean-only MLLR results in 10–15% relative reduction in WER
- Provides improvement in addition to VTLN (another 5–10% relative reduction in WER, after VTLN)
- Few regression classes and well-estimated transforms work best in practice
- Robust adaptation available with about 1 minute of speech; performance similar to SD models available with 30 minutes of adaptation data
- Such linear transforms can account for any systematic (linear) variation from the speaker independent models, for example those caused by channel effects.

Constrained MLLR (cMLLR)

aka feature-space MLLR (fMLLR)

- Basic idea use the same linear transform for both mean and covariance

$$\hat{\mu} = \mathbf{A}\mu + \mathbf{b}$$

$$\hat{\mathbf{\Sigma}} = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T$$

- No closed form solution but can be solved iteratively
- Log likelihood for cMLLR

$$L = \mathcal{N}(\mathbf{A}\mathbf{x}_n + \mathbf{b}; \mu, \mathbf{\Sigma}) + \log(|\mathbf{A}|)$$

  - Equivalent to applying the linear transform to the data!

- Iterative solution amenable to online/dynamic adaptation, by using just one iteration for each increment
- Similar improvement in accuracy to standard MLLR

Speaker-adaptive training (SAT)

- Basic idea Rather than SI seed (canonical) models, construct models designed for adaptation
  - Estimate parameters of canonical models by training MLLR mean transforms for each training speaker
  - Train using the MLLR transform for each speaker; interleave Gaussian parameter estimation and MLLR transform estimation
- SAT results in much higher training likelihoods, and improved recognition results ˜
- But: increased training complexity and storage requirements
- SAT using cMLLR, corresponds to a type of speaker normalization at training time

+ Speaker Space Methods

- Gender-dependent models: sets of HMMs for male and for female speakers
- Speaker clustering: sets of HMMs for different speaker clusters Drawbacks:
- Hard division of speakers into groups Fragments training data
- Weighted speaker cluster approaches which use an interpolated model to represent the current speaker
- Cluster-adaptive training Eignevoices

Cluster-adaptive training

- Basic idea: Represent a speaker as a weighted sum of speaker cluster models
- Different cluster models have shared variances and mixture weights, but separate means
- For a new speaker, mean is defined as $\boldsymbol{\mu} = \sum_c \lambda_c \boldsymbol{\mu}_c$
  - Given the canonical models, only the λc mixing parameters need to be estimated for each speaker
  - Given sets of weights for individual speakers, means of the clusters may be updated
- CAT can reduce WER in large vocabulary tasks by about 4–8% relative

+ Summary

- One of the most intensive areas of speech recognition research since the early 1990s
- Substantial progress, resulting in significant, additive, consistent reductions in word error rate
- Close mathematical links between different approaches Linear transforms at the heart of many approaches