Speaker Adaptation

Steve Renals

Automatic Speech Recognition— ASR Lecture 11
January-March 2012

Overview

Speaker Adaptation

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

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)
- A Speaker adaptive (SA) system... we would like
  - Error rates similar to SD systems
  - Building on an SI system
  - Requiring only a small fraction of the speaker-specific training data used by an SD system

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)

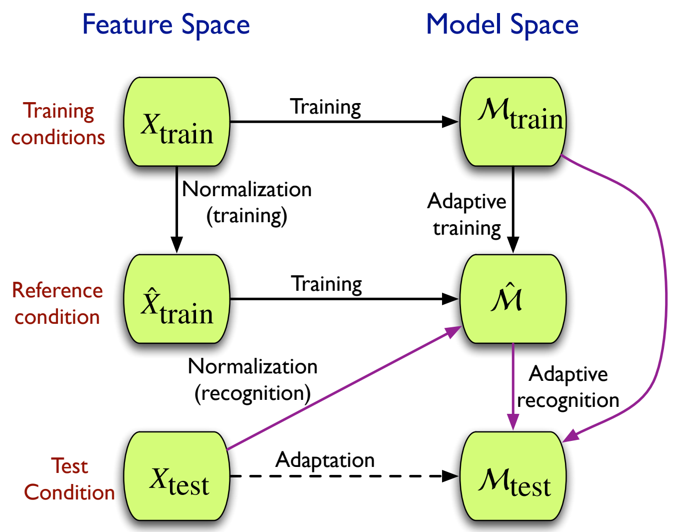
Modes of adaptation

- 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)

Approaches to adaptation

- 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
- Speaker Normalization: Normalize the acoustic data to reduce mismatch with the acoustic models
  - Vocal Tract Length Normalization (VTLN)
- Speaker space: Estimate multiple sets of acoustic models, characterizing new speakers in terms of these model sets
  - Cluster-adpative training
  - Eigenvoices

Adaptation and normalization of 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 $\boldsymbol{\lambda}$, then maximum likelihood (ML) training chooses them to maximize $p(\mathbf{X} \mid \boldsymbol{\lambda})$
- Maximum a posteriori (MAP) training maximizes:

$$p(\boldsymbol{\lambda} \mid \mathbf{X}) \propto p(\mathbf{X} \mid \boldsymbol{\lambda}) p_0(\boldsymbol{\lambda})$$

  $p_0(\boldsymbol{\lambda})$ 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

Refresher: ML estimation of GMM/HMM

- The mean of the mth Gaussian component of the jth state is estimated using a weighted average

$$\boldsymbol{\mu}_{mj} = \frac{\sum_n \gamma_{jm}(n)\mathbf{x}_n}{\sum_n \gamma_{jm}(n)}$$

- Where $\sum_n \gamma_{jm}(n)$ is the component occupation probability
- The covariance of the Gaussian component is given by:

$$\boldsymbol{\Sigma}_{mj} = \frac{\sum_n \gamma_{jm}(n)(\mathbf{x}_n - \boldsymbol{\mu}_{jm})(\mathbf{x}_n - \boldsymbol{\mu}_{jm})^T}{\sum_n \gamma_{jm}(n)}$$

MAP estimation

- What is $p_0(\boldsymbol{\lambda})$?
- Conjugate prior: the prior distribution has the same form as the posterior. There is no simple conjugate prior for GMMs, but an intuitively understandable approach may be employed.
- If the prior mean is $\boldsymbol{\mu}_0$, then the MAP estimate for the adapted mean $\hat{\boldsymbol{\mu}}$ of Gaussian is given by:

$$\hat{\boldsymbol{\mu}} = \frac{\tau \boldsymbol{\mu}_0 + \sum_n \gamma(n)\mathbf{x}_n}{\tau + \sum_n \gamma(n)}$$

  - τ is a *hyperparameter* that controls the balance between the ML estimate of the mean, its prior value. Typically τ is in the range 2–20
  - $\mathbf{x}_n$ is the adaptation vector at time n
  - $\gamma(n)$ the probability of this Gaussian at this time
- As the amount of training data increases, so the MAP estimate converges to the ML estimate

Local estimation

- Basic idea 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 10^5 Gaussians: most will not be adapted
  - Structural MAP (SMAP) approaches have been introduced to share Gaussians
  - 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 is very useful for domain adaptation:
  - Example: adapting a conversational telephone speech system (100s of hours of data) to multiparty meetings (10s of hours of data) works well with MAP

SMAP: Structural MAP

- Basic idea share Gaussians by organising 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
- This has been shown to speed adaptation compared with standard MAP, while converging to the same solution as standard MAP in the large data limit

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
- Maximum Likelihood Linear Regression (MLLR) is the best known linear transform approach to speaker adaptation

Interim Summary

- Speaker-specific variation
- Adaptation: supervised/unsupervised, static/dynamic
- Model-based adaptation: MAP
- Introduction to model-based adaptation
- Next lecture: MLLR, adaptive training, speaker space models and vocal tract length normalisation

Speaker Adaptation 2

Steve Renals

Automatic Speech Recognition— ASR Lecture 12
7 March 2011

Overview

Speaker Adaptation

- Introduction: speaker-specific variation, modes of adaptation
- Model-based adaptation: MAP and MLLR
- Adaptive training
- Model-based adaptation: Speaker space models
- Speaker normalization: VTLN

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
- Maximum Likelihood Linear Regression (MLLR) is the best known linear transform approach to speaker adaptation

MLLR: Maximum Likelihood Linear Regression

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

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

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

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

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

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
- For an adaptation set, work down the tree until arriving at the most specific set of nodes for which there is sufficient data
- Regression class tree constructed in a similar way to state clustering tree
- In practice the number of regression may be very small: one per context-independent phone class, one per broad class, or even just two (speech/non-speech)

Estimating the transforms

- The linear transformation matrix W is obtained by finding its setting which optimizes the log likelihood
- Mean adaptation: 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
- Differentiating L and setting to 0 results in an equation for $\mathbf{W}$: there is no closed form solution if $\boldsymbol{\Sigma}$ is full covariance; can be solved if $\boldsymbol{\Sigma}$ is diagonal (but requires a matrix inversion)
- Variance adaptation is also possible
- See Gales and Woodland (1996), Gales (1998) for details

MLLR in practice

- Mean-only MLLR results in 10–15% relative reduction in WER
- 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)

- 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

$$L = \mathcal{N}(\mathbf{A}\mathbf{x}_n + \mathbf{b}; \mu, \mathbf{\Sigma}) + \log(|\mathbf{A}|) \qquad \mathbf{A}' = \mathbf{A}^{-1}; \mathbf{b}' = \mathbf{A}\mathbf{b}$$

  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
  - Eigenvoices

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 λ_c mixing parameters need 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
- See Gales (2000) for more

Eigenvoices

- Basic idea Construct a speaker space from a set of SD HMMs
- Could regard each canonical model as forming a dimension of speaker space
- Generalize by computing PCA of sets of "supervectors" (concatenated mean vectors), to form speaker space: each dimension is an "eigenvoice"
- Represent a new speaker as a combination of eigenvoices
- Close relation to CAT
- Computationally intensive, does not scale well to large vocabulary systems
- See Kuhn et al (2000) for more

Feature normalization

- Basic idea: Transform the features to reduce mismatch between training and test
- *Cepstral Mean Normalization* (CMN): subtract the avergae feature value from each feature, so each feature has a mean value of 0. makes features robust to some linear filtering of the signal (channel variation)
- *Cepstral Variance Normalization* (CVN): Divide feature vector by standard deviation of feature vectors, so each feature vector element has a variance of 1
- Cepstral mean and variance normalisation, CMN/CVN:

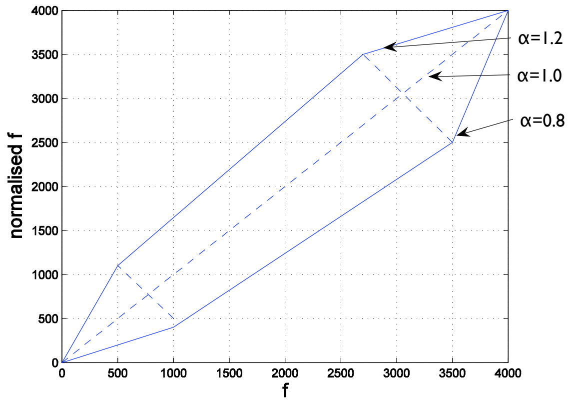
$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_i - \boldsymbol{\mu}(\mathbf{x})}{\boldsymbol{\sigma}(\mathbf{x})}$$

- Compute mean and variance statistics over longest available segments with the same speaker/channel
- Real time normalisation: compute a moving average

Vocal Tract Length Normalization (VTLN)

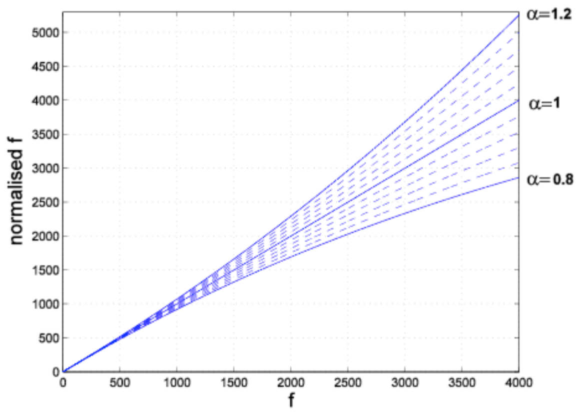
- Basic idea Normalize the acoustic data to take account of changes in vocal tract length
- Vocal tract length (VTL):
  - First larynx descent in first 2-3 years of life
  - VTL grows according to body size, and is sex-dependent
  - Puberty: second larynx descent for males
- VTL has large effect on the 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

Warping functions: Piecewise linear



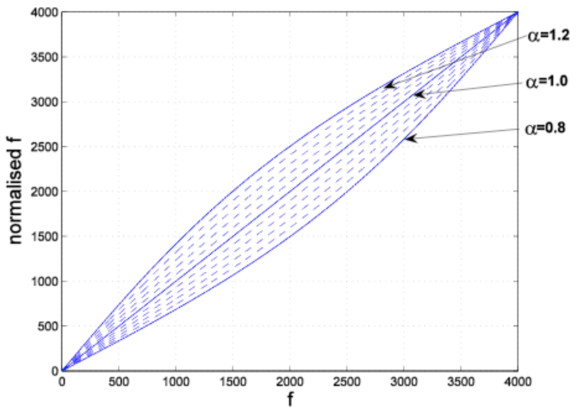
$$\hat{f} = \alpha f$$

Warping functions: Power function



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

Warping functions: Power function



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

Approaches to VTLN

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

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

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 ($\bar{F}_3$):

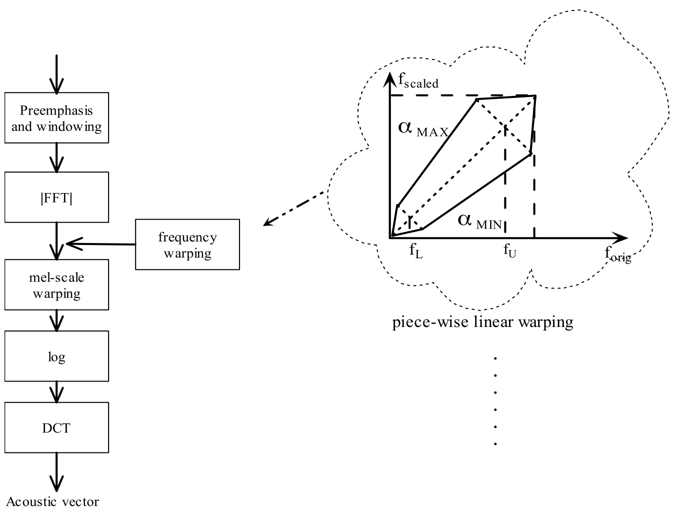
$$\alpha_s = \frac{\bar{F}_{3,s}}{\bar{F}_3}$$

- Wegmann et al (1996) used a generic voiced speech model, estimated using maximum likelihood. During training, estimation of warping factors was alternated with estimating the phone models using the warped data
- These approaches require an accurate estimation of voiced parts of the speech signal

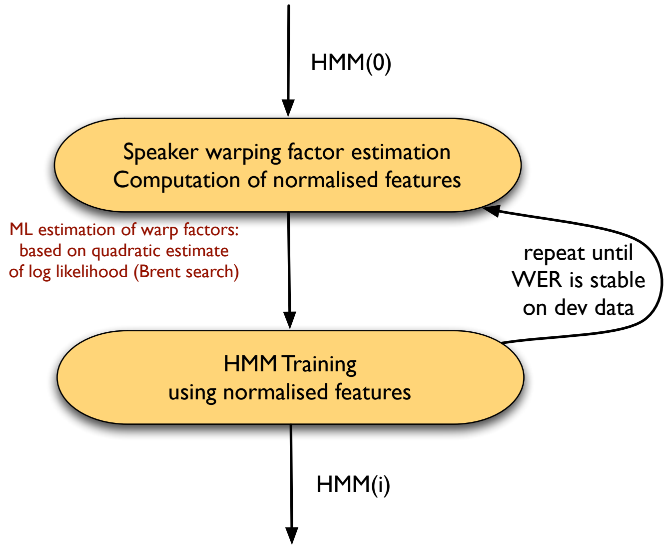
Model-based VTLN

- 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
- The process may be iterated
- Model-based VTLN does not directly estimate vocal tract size, rather it estimates an optimal frequency warping, which may be affected by other factors (eg $F0$)
- Exhaustive search for the optimal warping factor would be expensive
  - Approximate the log likelihood wrt α as a quadratic, and find the maximum using a line search (Brent's method)

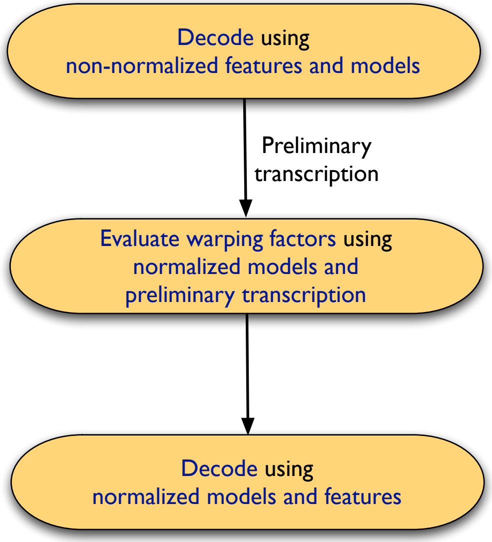
Model-based VTLN



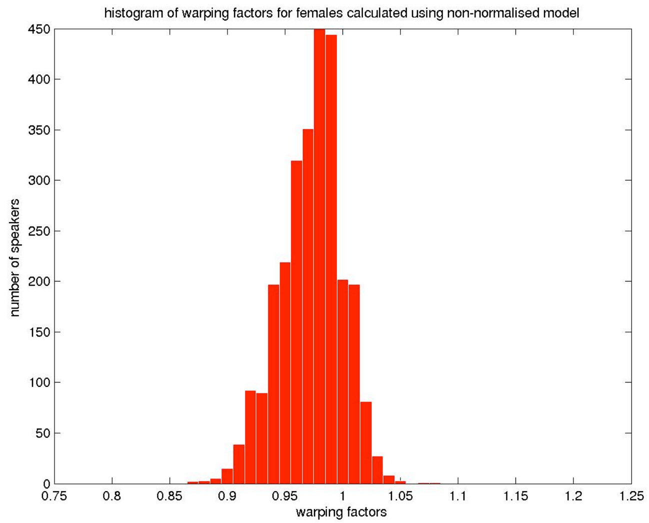
VTLN: Training



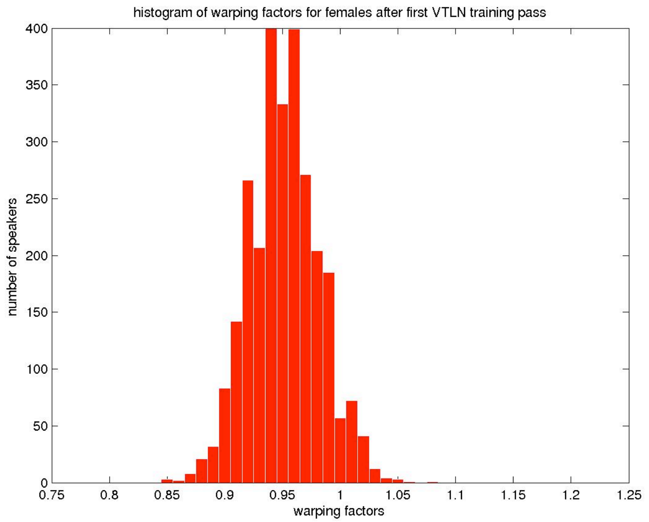
VTLN: Recognition



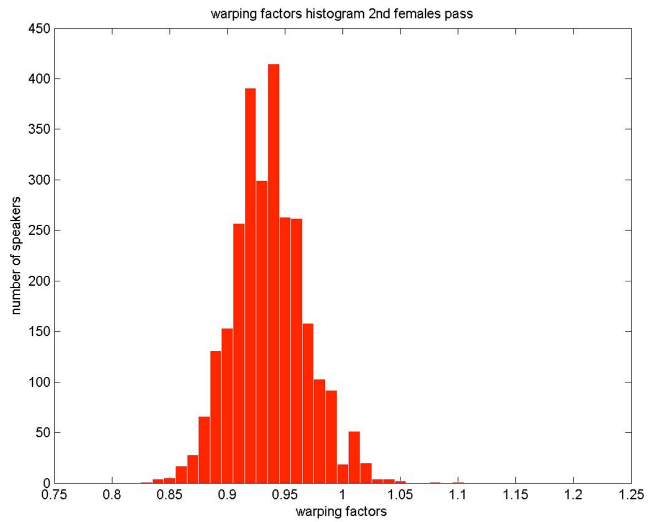
VTLN: Warp factor estimation, females, non-normalized



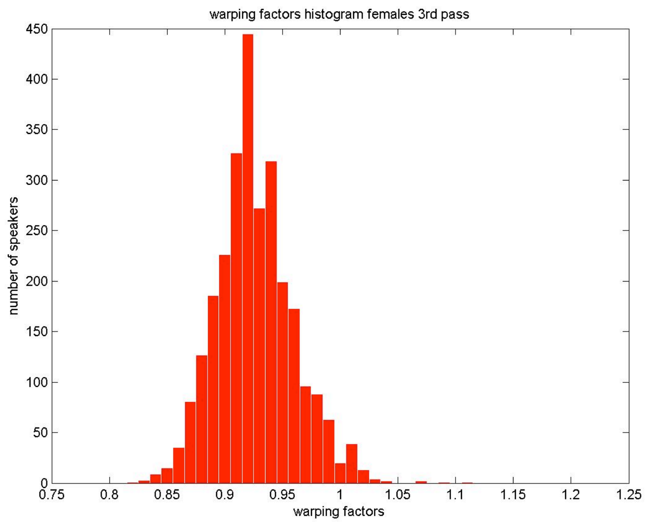
VTLN: Warp factor estimation, females, pass 1



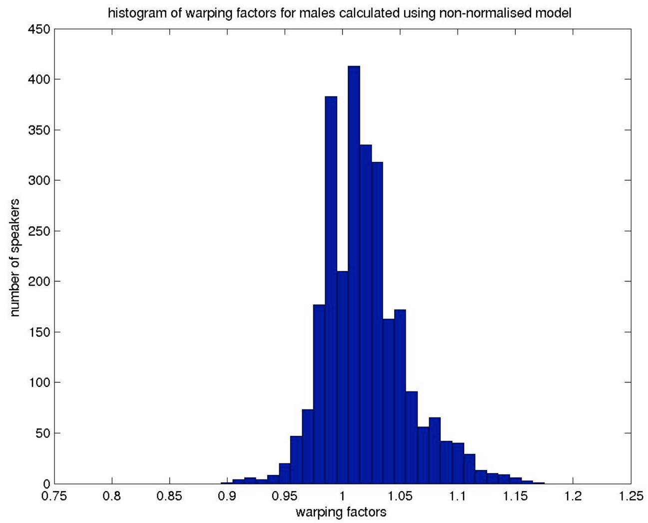
VTLN: Warp factor estimation, females, pass 2



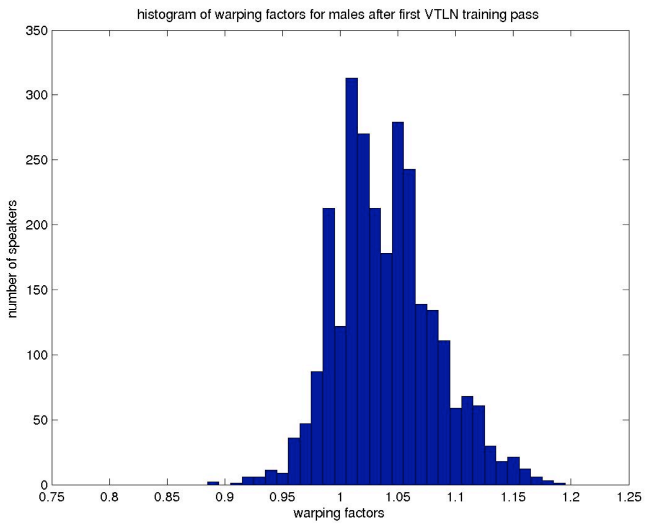
VTLN: Warp factor estimation, females, pass 3



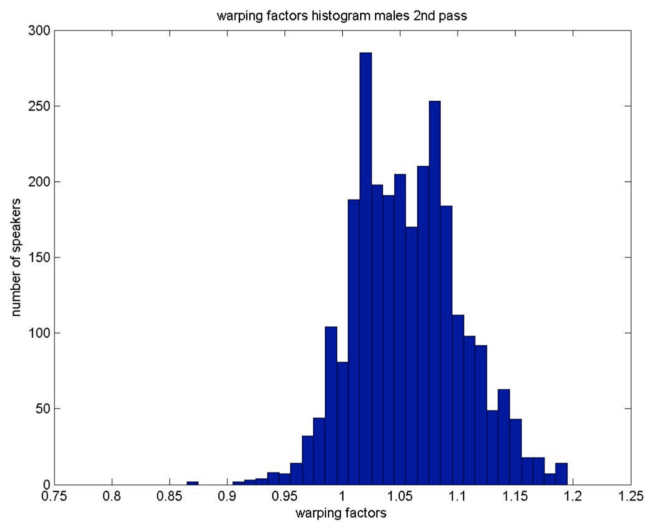
VTLN: Warp factor estimation, males, non-normalized



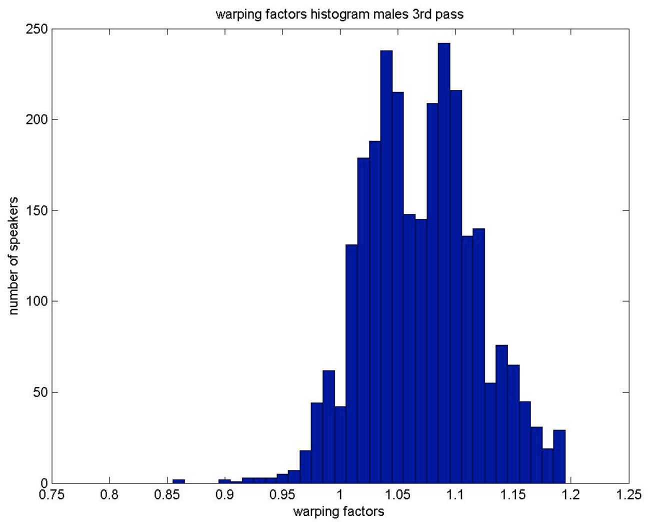
VTLN: Warp factor estimation, males, pass 1



VTLN: Warp factor estimation, males, pass 2



VTLN: Warp factor estimation, males, pass 3



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

Summary

Speaker Adaptation

- 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

Reading

- Gales and Young (2007), sec. 5. Good overview.
- Woodland (2001). Review paper.
- Gales (1998). Best overview of the MLLR family.
- Garau et al (2005). VTLN (for meetings speech).
- Kuhn et al (2000). Eigenvoices.