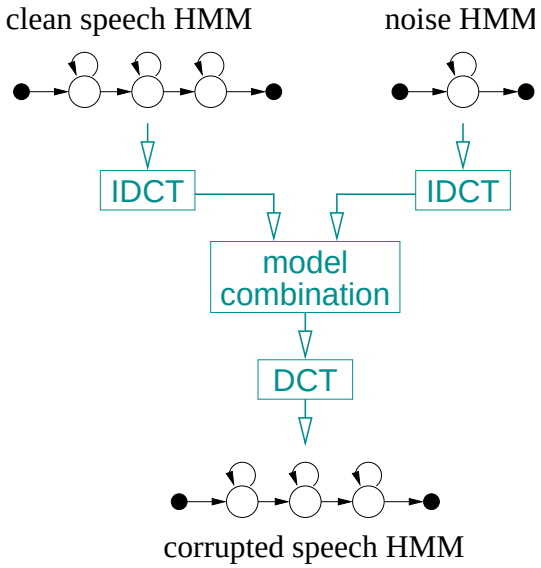
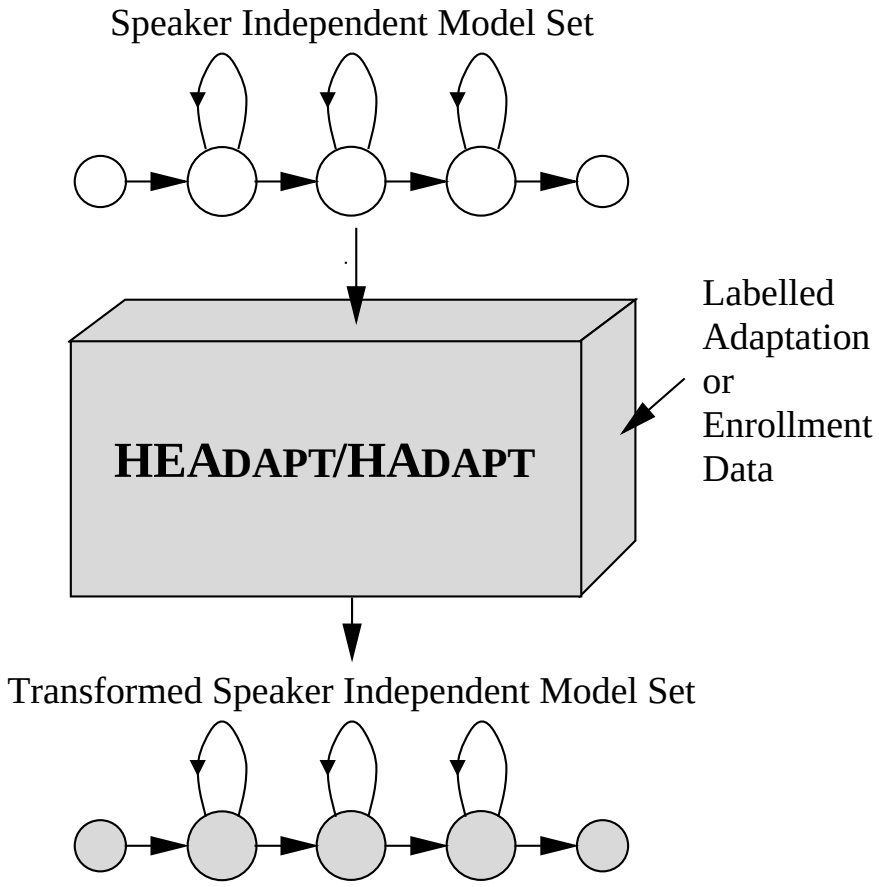
Speaker adaptation & noise robustness

Dr Philip Jackson

- Speaker adaptation
  - vocal-tract normalisation
  - ML regression & MAP
- Noise robustness
  - spectral subtraction
  - channel normalisation
  - noise masking
  - model combination



Speaker adaptation



Model adaptation, from Young et al. (2002).

Overview of adapting to a speaker

- *speaker independent* $\rightarrow$ *speaker dependent*
  - can halve recognition errors
  - but need data for training...
- **Normalisation & adaptation**
- **Availability of transcriptions**
- **Enrollment session**
- **Approaches**

Overview of adapting to a speaker

- *speaker independent* $\rightarrow$ *speaker dependent*
- **Normalisation & adaptation**
- **Availability of transcriptions**:
  - supervised
  - unsupervised
- **Enrollment session**:
  - static (off-line)
  - incremental (on-line)
- **Approaches**:
  - MLLR (transformation matrix)
  - MAP (new parameter estimates)

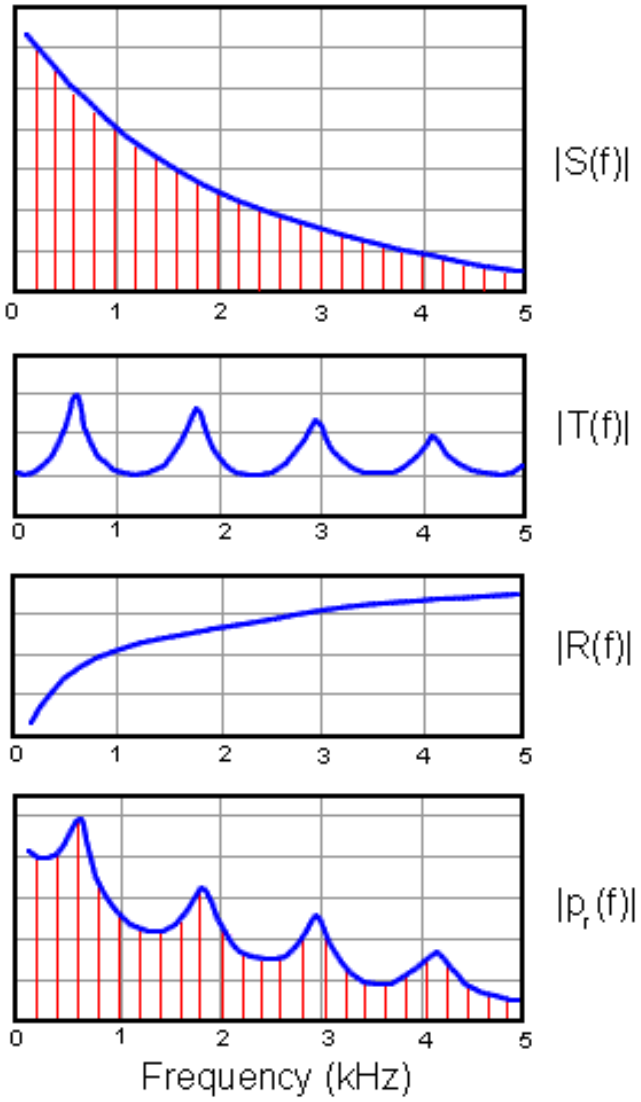
Creating speaker-dependent models

- **Speaker normalisation**
  - vocal tract length normalisation (VTLN)
- **Speaker adaptation**
  - Maximum likelihood linear regression (MLLR)
  - Maximum a posteriori (MAP)

	static (supervised)	incremental (unsupervised)
MLLR	$\surd$	$\surd$
MAP	$\surd$	

Methods for performing speaker adaptation.

Formant frequencies and vocal-tract length

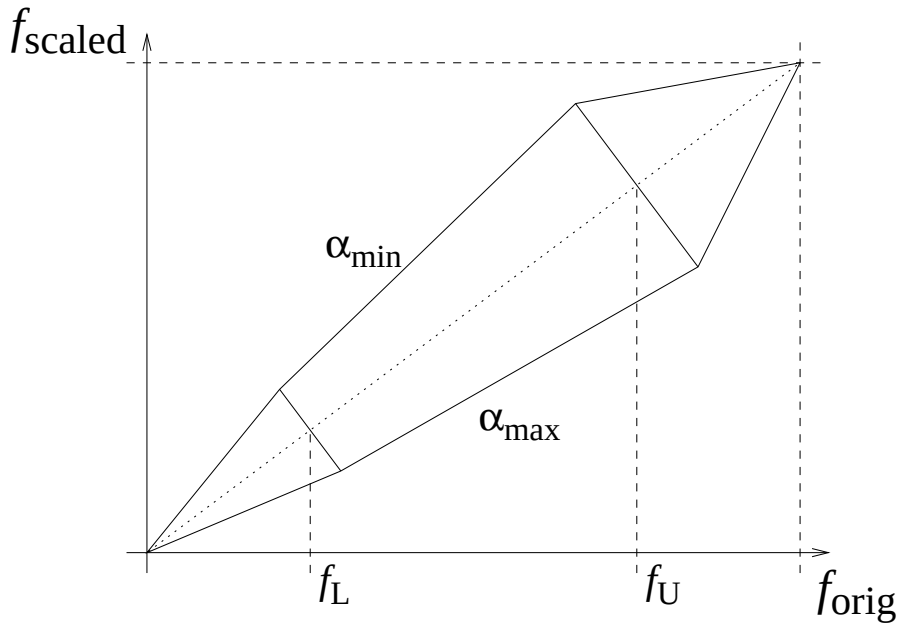


A speaker's natural formant frequencies are inversely proportional to the length of their vocal tract:

$$F_n = \frac{c}{4l}(2n-1)$$

Vocal-tract length normalisation

- Compensates for vocal-tract length's effect on formants
- Frequency warping (piecewise linear):



VTLN frequency-warping functions, from Young et al. (2002).

Speaker adaptation by MLLR

Maximum-likelihood linear regression (MLLR) uses matrices $\mathbf{W}$ and $\mathbf{H}$ to transform the means and variances into new speaker-dependent models:

$$\hat{\boldsymbol{\mu}}_j = \mathbf{W}\bar{\boldsymbol{\mu}}_j \tag{1}$$

The Gaussian covariance $\boldsymbol{\Sigma}_j$ is adapted using

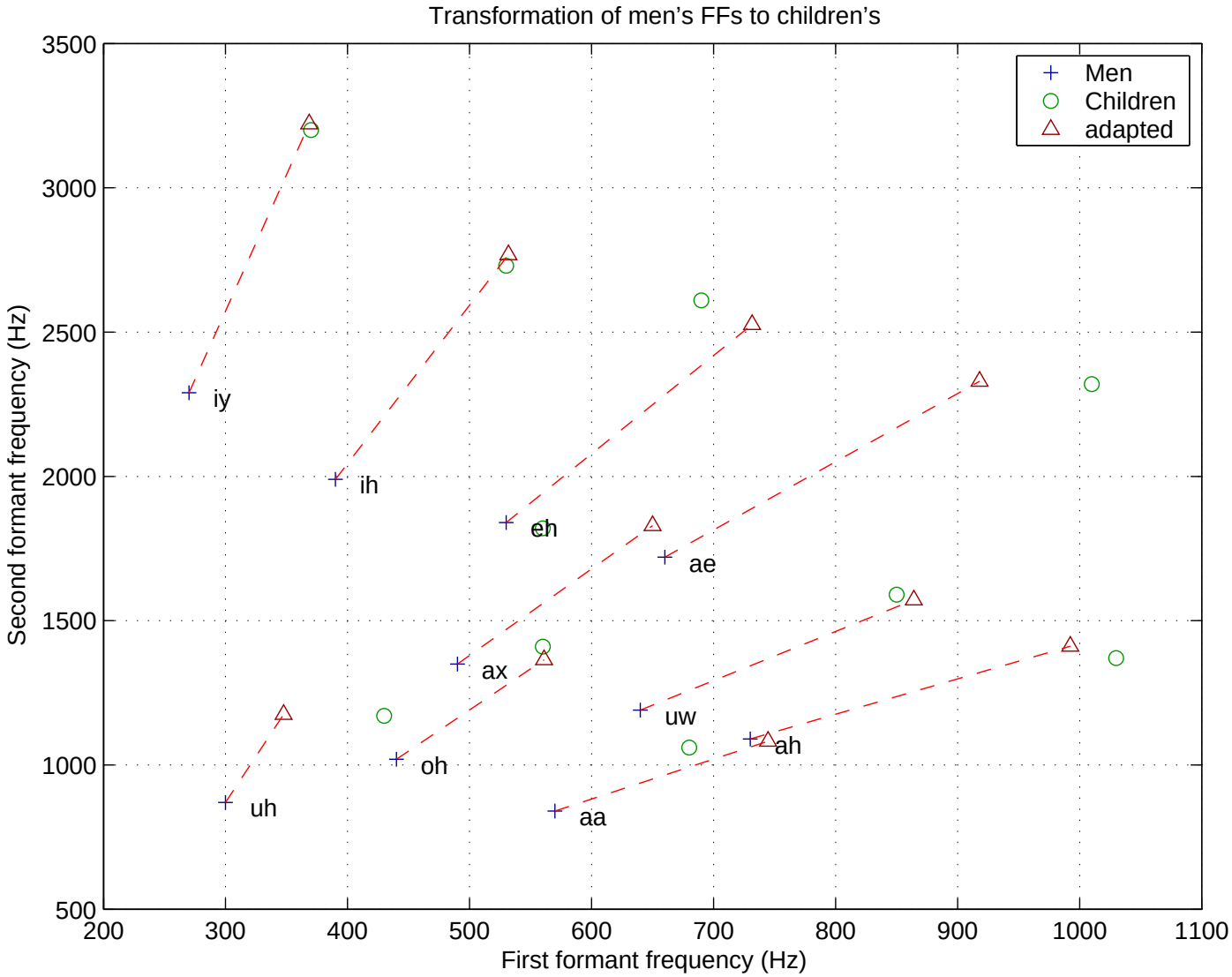
$$\hat{\boldsymbol{\Sigma}}_j = \mathbf{B}_j^{\mathsf{T}}\,\mathbf{H}\,\mathbf{B}_j \tag{2}$$

where $\mathbf{H}$ is the linear transformation to be estimated and $\mathbf{B}_j$ is the inverse of the Cholesky factor* of $\boldsymbol{\Sigma}_j^{-1}$, so

$$\boldsymbol{\Sigma}_j^{-1} = \mathbf{C}_j\mathbf{C}_j^{\mathsf{T}} \qquad \text{and} \qquad \mathbf{B}_j = \mathbf{C}_j^{-1}$$

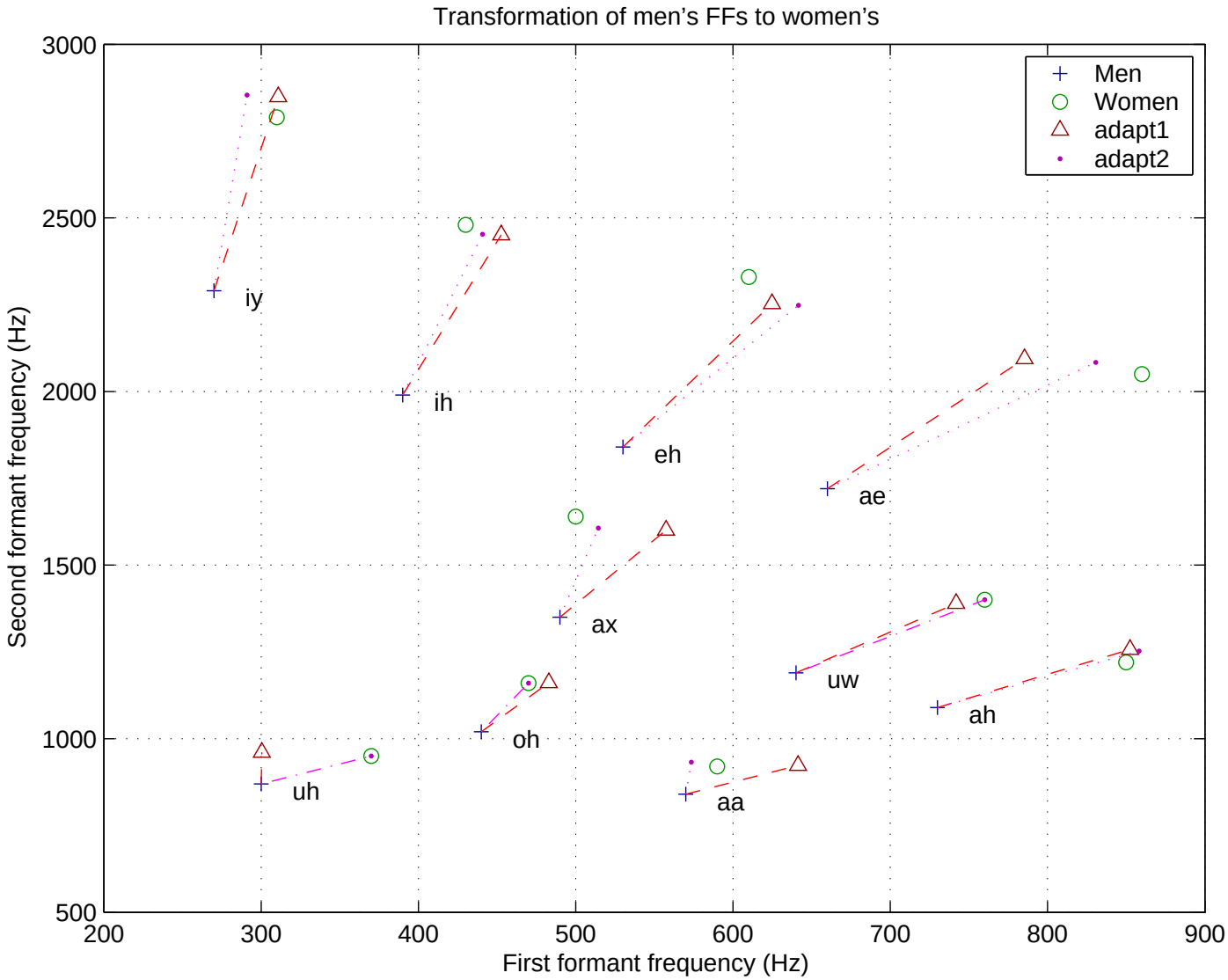
* The Cholesky factor can be thought of as the "square-root" of the inverse covariance matrix (Deller et al. 1993, p. 313).

Example of linear-regression transform (MLLR-1)



Example transformation of adult male to child vowel formant frequencies by linear regression.

Example of linear-regression transforms (MLLR-2)



Linear transforms of adult male to female vowel formants:
one class (△), and two classes based on lip rounding (·). Q.10

Speaker adaptation by MAP of means

The maximum *a posteriori* (MAP) update formula is:

$$\hat{\boldsymbol{\mu}}_j = \frac{\tau}{\tau + \Gamma_j}\bar{\boldsymbol{\mu}}_j + \frac{\Gamma_j}{\tau + \Gamma_j}\dot{\boldsymbol{\mu}}_j \tag{3}$$

where τ is a weighting of prior knowledge (from training) to adaptation data, and Γ_j is the accumulated occupation likelihood of adaptation data for state j:

$$\Gamma_j = \sum_{t=1}^{T} \gamma_j^{(t)}$$

$\bar{\boldsymbol{\mu}}_j$ is the speaker-independent mean, and $\dot{\boldsymbol{\mu}}_j$ is mean of the adaptation data $\mathcal{O}$:

$$\dot{\boldsymbol{\mu}}_j = \frac{\sum_{t=1}^{T} \gamma_j^{(t)} \boldsymbol{o}_t}{\sum_{t=1}^{T} \gamma_j^{(t)}}$$

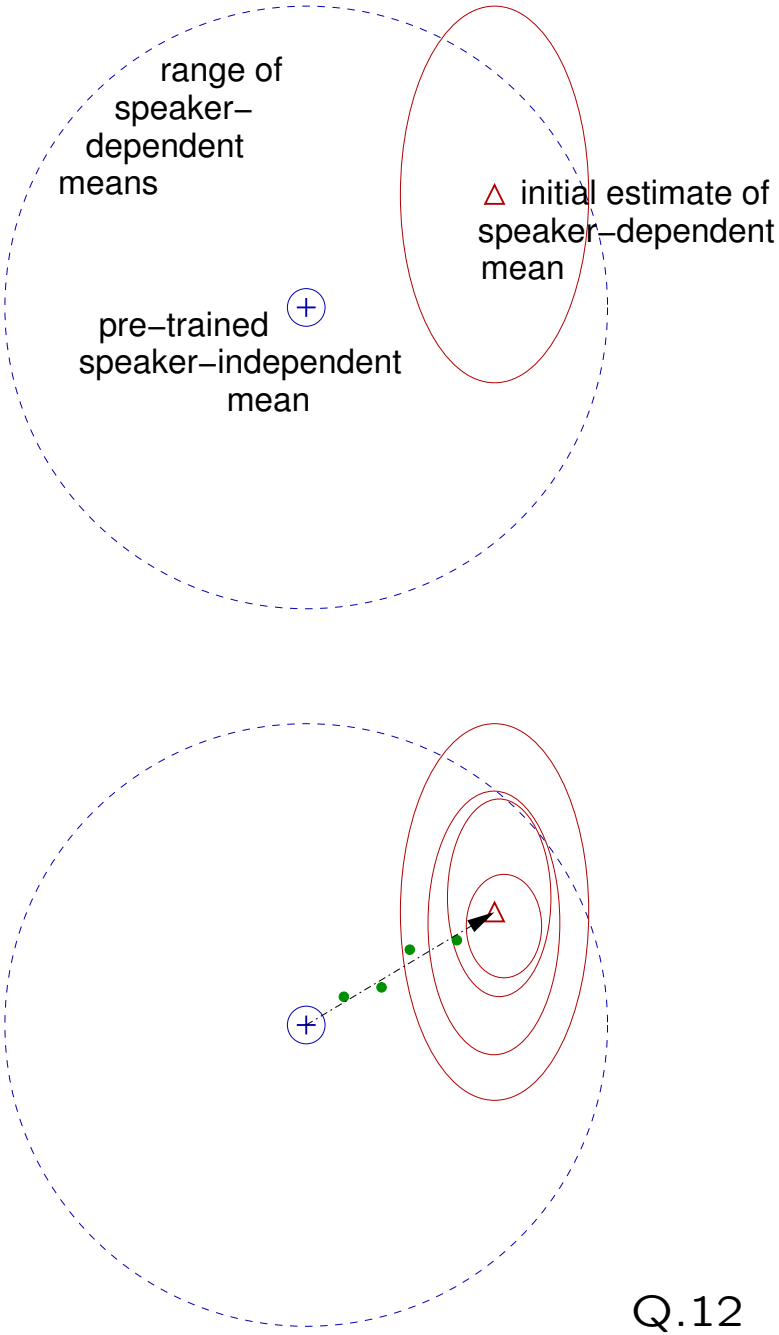
Thus, if Γ_j is small, the MAP-estimated model stays near the speaker-independent model; it moves away after more observations, as Γ_j increases. Q.11

MAP adaptation

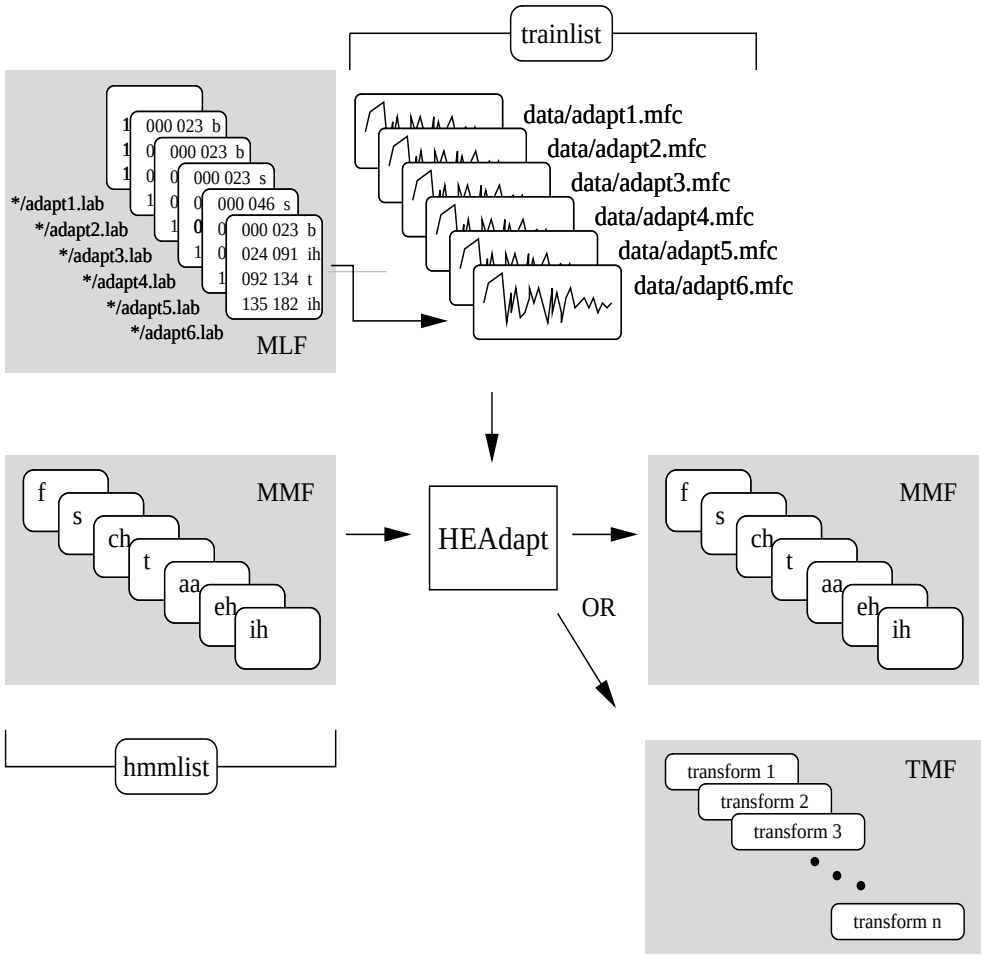
Adaptation starts from the pre-trained speaker-independent mean, $\bar{\mu}$.

As adaptation data arrive, the speaker-dependent estimate, $\dot{\mu}$, is initially uncertain.

As estimates become more reliable, the MAP-adapted mean, $\hat{\mu}$, approaches the speaker-dependent mean.



Practical speaker adaptation of models



HMM adaptation from enrollment data, from Young et al. (2002).

Speaker adaptation summary

- Aim to improve recognition preformance by tailoring the system to an individual's voice
- Speaker normalisation: reduces speaker variation
  - vocal-tract length normalisation (VTLN)
- Speaker adaptation: customises the models
  - maximum-likelihood linear regression (MLLR)
  - maximum a posteriori (MAP) re-estimation

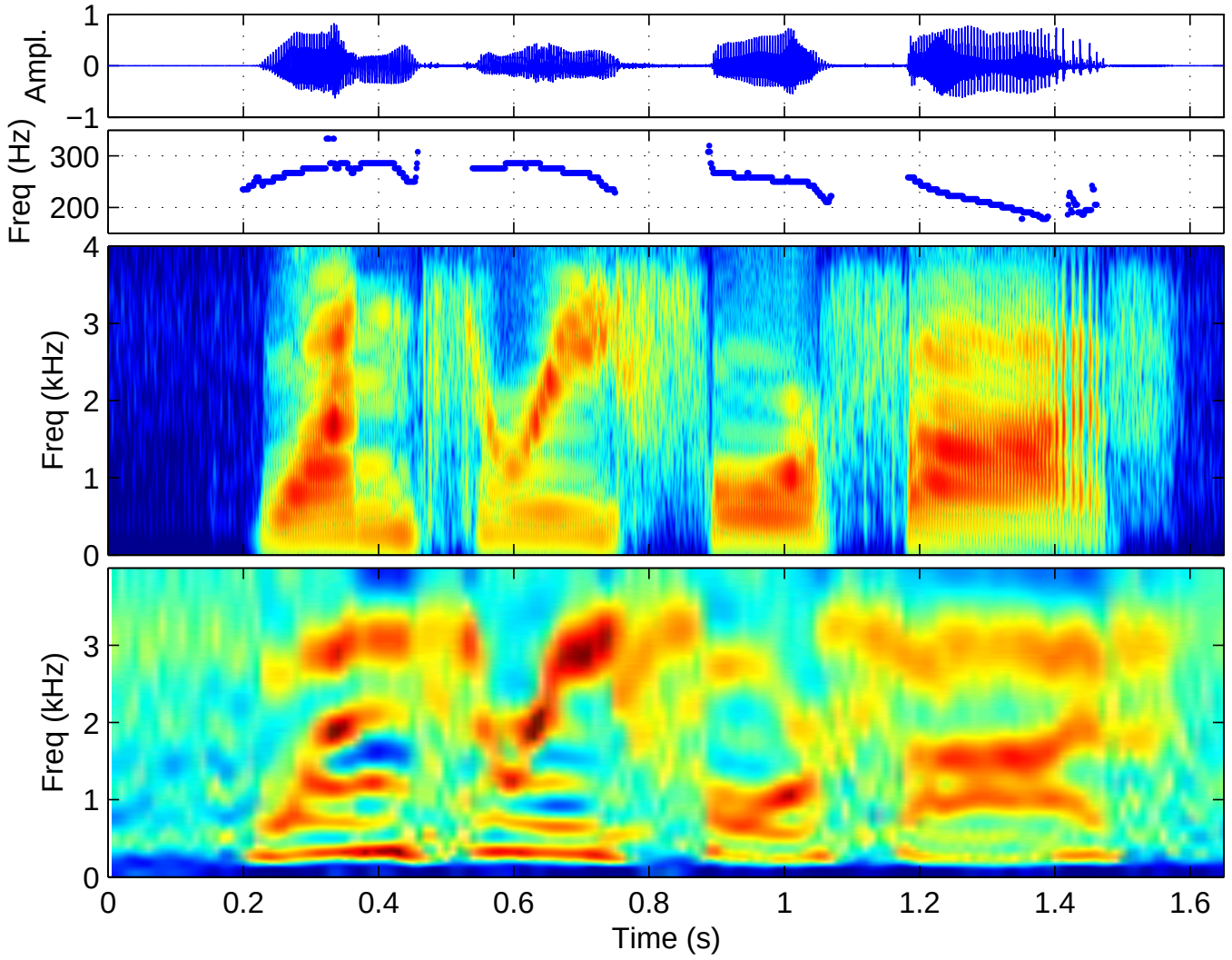
Robust ASR

- **Real systems** differ from those in the lab:
  - microphones
  - noisy transmission channel
  - background noise
  - reverberation
- **Speaker's response to noise**:
  - Lombard effect
- **Recognition performance**:
  - better in 'clean' conditions than in noisy ones
  - worst with mismatched training/test conditions

Robust ASR

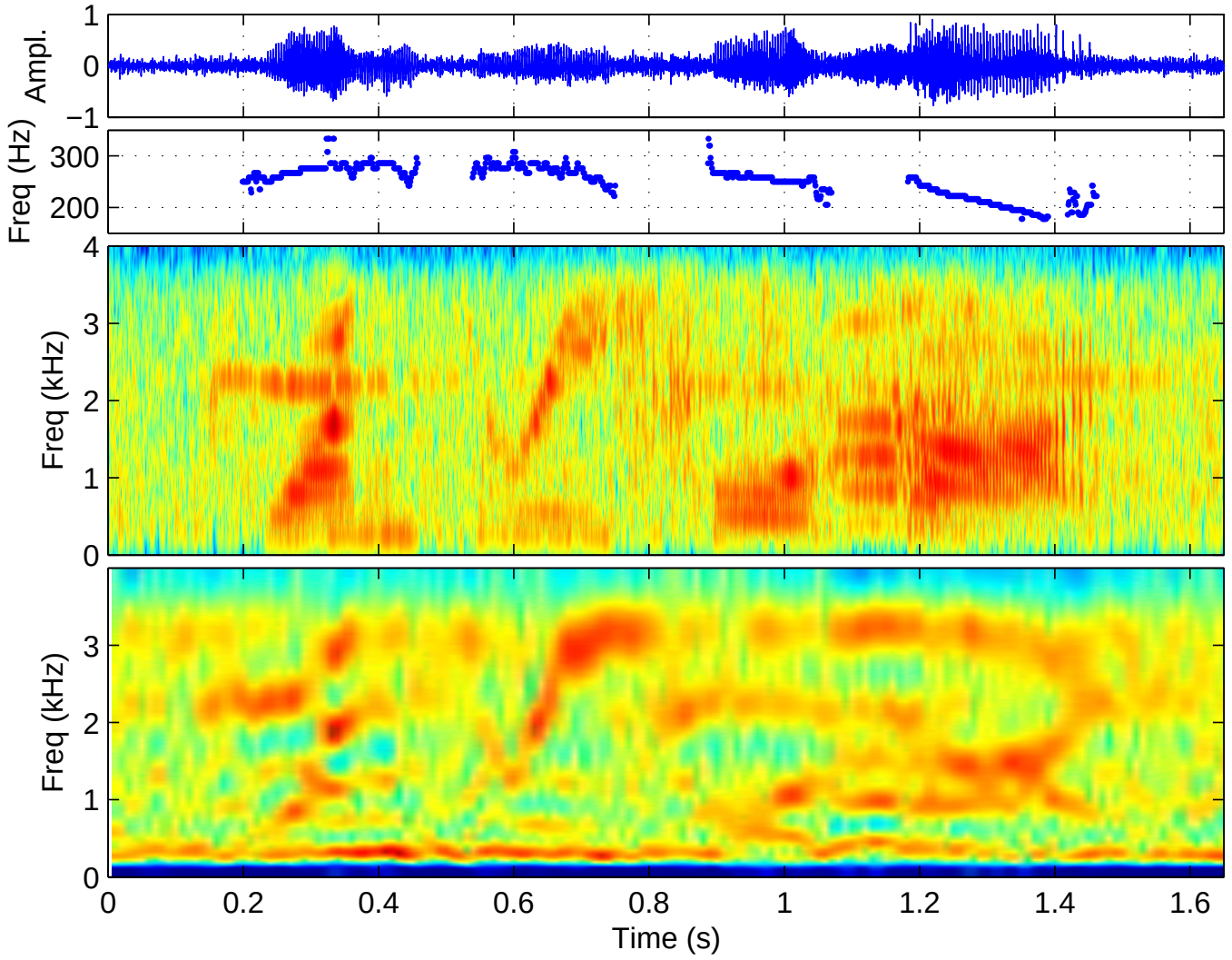
- **Real systems** differ from those in the lab:
  - microphones
  - noisy transmission channel
  - background noise
  - reverberation
- **Speaker's response to noise**:
  - Lombard effect
- **Recognition performance**:
  - better in 'clean' conditions than in noisy ones
  - worst with mismatched training/test conditions
- **Solutions**:
  - attempt to train system with *matched* data
  - remove unreliable information from acoustics

Clean speech spectrograms



Clean recording of female utterance "1-3-4-5" (from top): acoustic waveform, fundamental frequency f_0, wide-band spectrogram, and spectrogram derived from the MFCCs. Q.17

Noisy speech spectrograms



Noise-corrupted female utterance "1-3-4-5" (from top): acoustic waveform, fundamental frequency f_0, wide-band spectrogram, and spectrogram derived from the MFCCs. Q.18

Approaches to noise robustness

- **Subtraction & normalisation**
  - additive noise
  - convolutional noise
- **Noise modelling**
  - integrating a model of noise into recognizer

Feature-based approaches

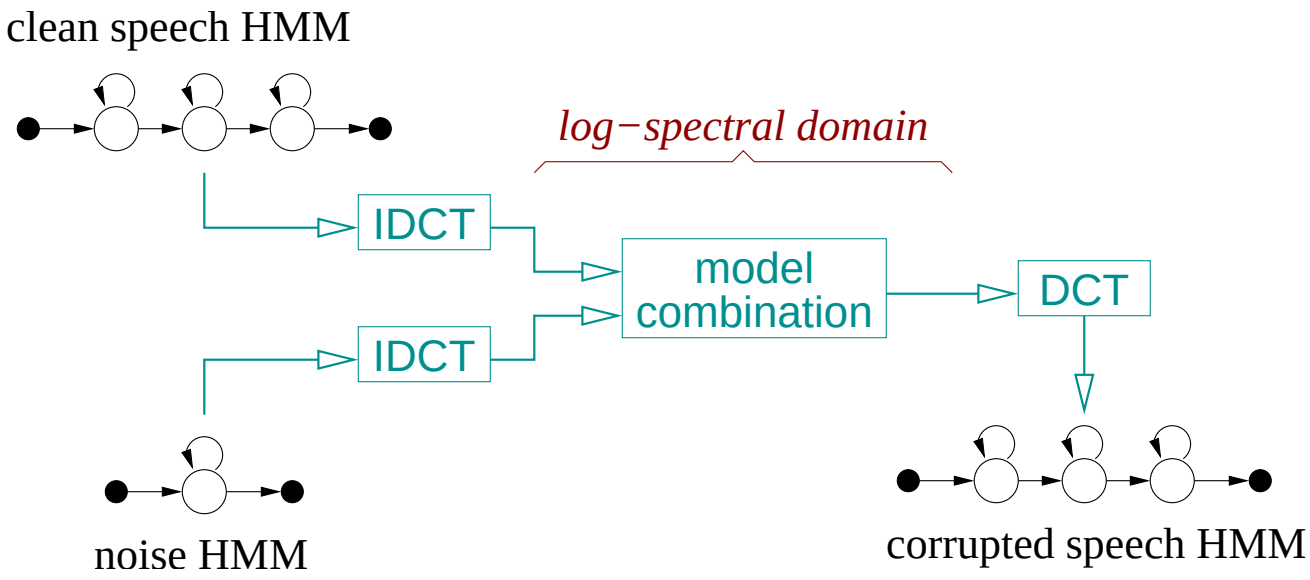
- **Spectral subtraction** (additive)
  - use silence, or a second microphone
  - negative values set to zero
    Done *before* binning step in feature extraction
- **Channel normalisation** (convolutional)
  - use speech, has changing spectral characteristics
  - compute mean cepstrum over longer time scale
    (e.g., a few seconds)
  - cepstral mean subtraction (CMS)
    a.k.a. cepstral mean normalisation (CMN)
- Band-pass filtering of spectral characterstics
  - relative spectral processing (**RASTA** features)
- Difference features are inherently robust

Model-based approaches

- **Shared concepts**
  - prob. calculation weights most reliable features
  - compare observation and noise-estimate levels
- **Noise masking**:
  - apply same spectral subtraction to model as to the observations before probability calculation
- **Missing data**:
  - treat masked features as 'missing'
  - eliminate these from probability calculation
- **Explicit noise model**:
  - stationary (one-state HMM)
  - non-stationary (multiple-state HMM), which required all pairings to be considered in decoding.

Parallel model combination (PMC)

- 'clean'-speech and noise models combined to produce 'corrupted'-speech model
- if noise model has more than one state, then corrupted model has more states than clean model
- training procedure must be modified to estimate all parameters in clean-speech and noise models



Representation of parallel model combination, adapted from Gales & Young (1996).

Unknown noise corruption

Ways of modeling unknown corrupting noise:

1. **Prediction**
   - PMC is predictive (of corrupted speech model)
   - assumes independence of speech and noise
2. **Adaptation**
   - adaptation techniques (e.g., MLLR & MAP) can adjust model parameters to deal with changes in speaking style and model variance

Noise robustness summary

- Feature-based approaches
  - spectral subtraction
  - channel normalisation
  - noise-robust features (RASTA)
- Model-based approaches
  - noise masking & missing data
  - parallel model combination (PMC)
  - adaptation

Appendix

Speaker adaptation by MLLR

Maximum-likelihood linear regression (MLLR) uses matrices $\mathbf{W}$ and $\mathbf{H}$ to transform the means and variances into new speaker-dependent models:

$$\hat{\boldsymbol{\mu}}_j = \mathbf{W}\bar{\boldsymbol{\mu}}_j$$

ML solution for extended mean vector, $\zeta_j = \begin{bmatrix} 1 \\ \bar{\boldsymbol{\mu}}_j \end{bmatrix}$, gives

$$\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_j^{(t)}\boldsymbol{\Sigma}_j^{-1}\boldsymbol{o}_t\zeta_j^{\mathsf{T}} = \sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_j^{(t)}\boldsymbol{\Sigma}_j^{-1}\mathbf{W}\zeta_j\zeta_j^{\mathsf{T}} \qquad (4)$$

where occupation likelihood $\gamma_j^{(t)}$ at time t is computed for adaptation data $\mathcal{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$.

To train transformations flexibly and robustly, states are divided into groups in practice (e.g., by CART) and a transform matrix $\mathbf{W}_\rho$ is estimated for each group's set of Gaussians $\{m_1, m_2, \ldots, m_M\}$, where $\rho = 1..M$.

MLLR derivation for the means

To help solve eq. (4) for $\mathbf{W}$, two new terms are defined:

$$\mathbf{Z} = \sum_{t=1}^{T} \sum_{j=1}^{N} \gamma_j^{(t)} \mathbf{\Sigma}_j^{-1} \boldsymbol{o}_t \zeta_j^{\mathsf{T}}$$

and $\mathbf{G}^{(i)}$ with elements

$$g_{mn}^{(i)} = \sum_{j=1}^{N} v_{ii}^{(j)} d_{mn}^{(j)}$$

where

$$\mathbf{V}^{(j)} = \sum_{t=1}^{T} \gamma_j^{(t)} \mathbf{\Sigma}_j^{-1} \qquad \text{and} \qquad \mathbf{D}^{(j)} = \zeta_j \zeta_j^{\mathsf{T}}$$

Hence, the elements of $\mathbf{W}$ can be calculated from

$$\mathbf{w}_i^{\mathsf{T}} = \mathbf{G}_i^{-1} \mathbf{z}_i^{\mathsf{T}} \tag{5}$$

where $\mathbf{w}_i$ is the ith vector of $\mathbf{W}$, and $\mathbf{z}_i$ of $\mathbf{Z}$.

MLLR derivation for diagonal covariances

The Gaussian covariance $\boldsymbol{\Sigma}_j$ is adapted using

$$\hat{\boldsymbol{\Sigma}}_j = \mathbf{B}_j^{\mathsf{T}} \mathbf{H} \mathbf{B}_j$$

where $\mathbf{H}$ is the linear transformation to be estimated and $\mathbf{B}_j$ is the inverse of the Cholesky factor* of $\boldsymbol{\Sigma}_j^{-1}$, so

$$\boldsymbol{\Sigma}_j^{-1} = \mathbf{C}_j \mathbf{C}_j^{\mathsf{T}} \qquad \text{and} \qquad \mathbf{B}_j = \mathbf{C}_j^{-1}$$

After rewriting the auxiliary function, the transform matrix is estimated as

$$\mathbf{H} = \frac{\sum_{j=1}^{N} \mathbf{C}_j^{\mathsf{T}} \left[\gamma_j^{(t)} (\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_j)(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_j)^{\mathsf{T}} \right] \mathbf{C}_j}{\gamma_j^{(t)}} \tag{6}$$

* The Cholesky factor can be thought of as the "square-root" of the inverse covariance matrix (Deller et al. 1993, p. 313).