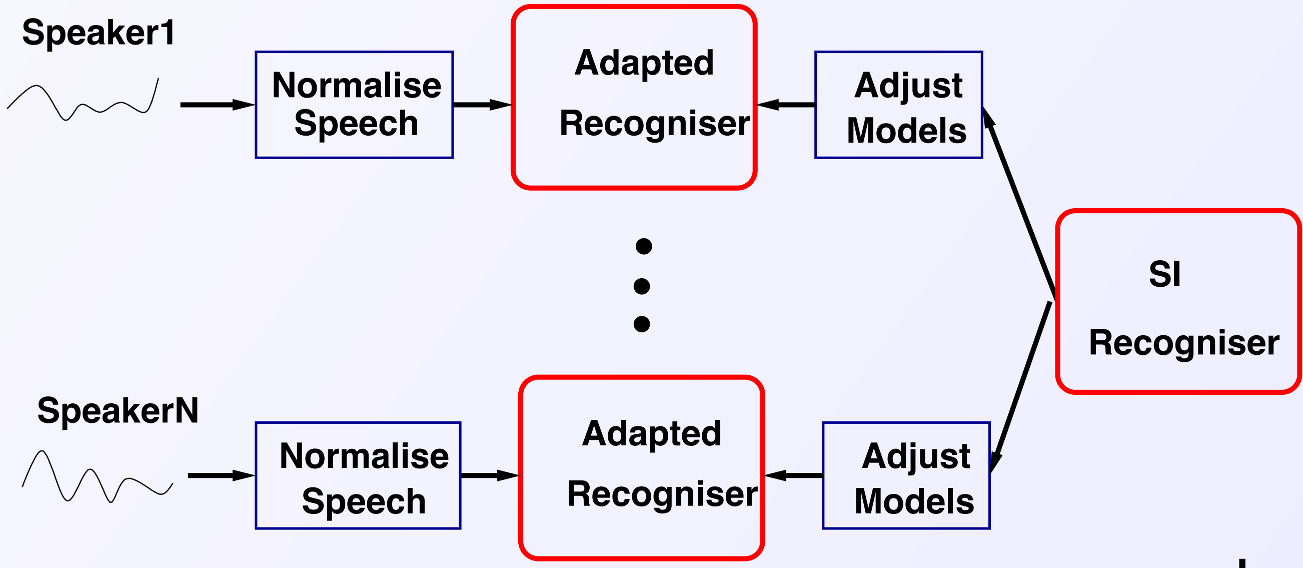
COM4511/COM6511 - Speech Technology

Handout 9
Adaptation

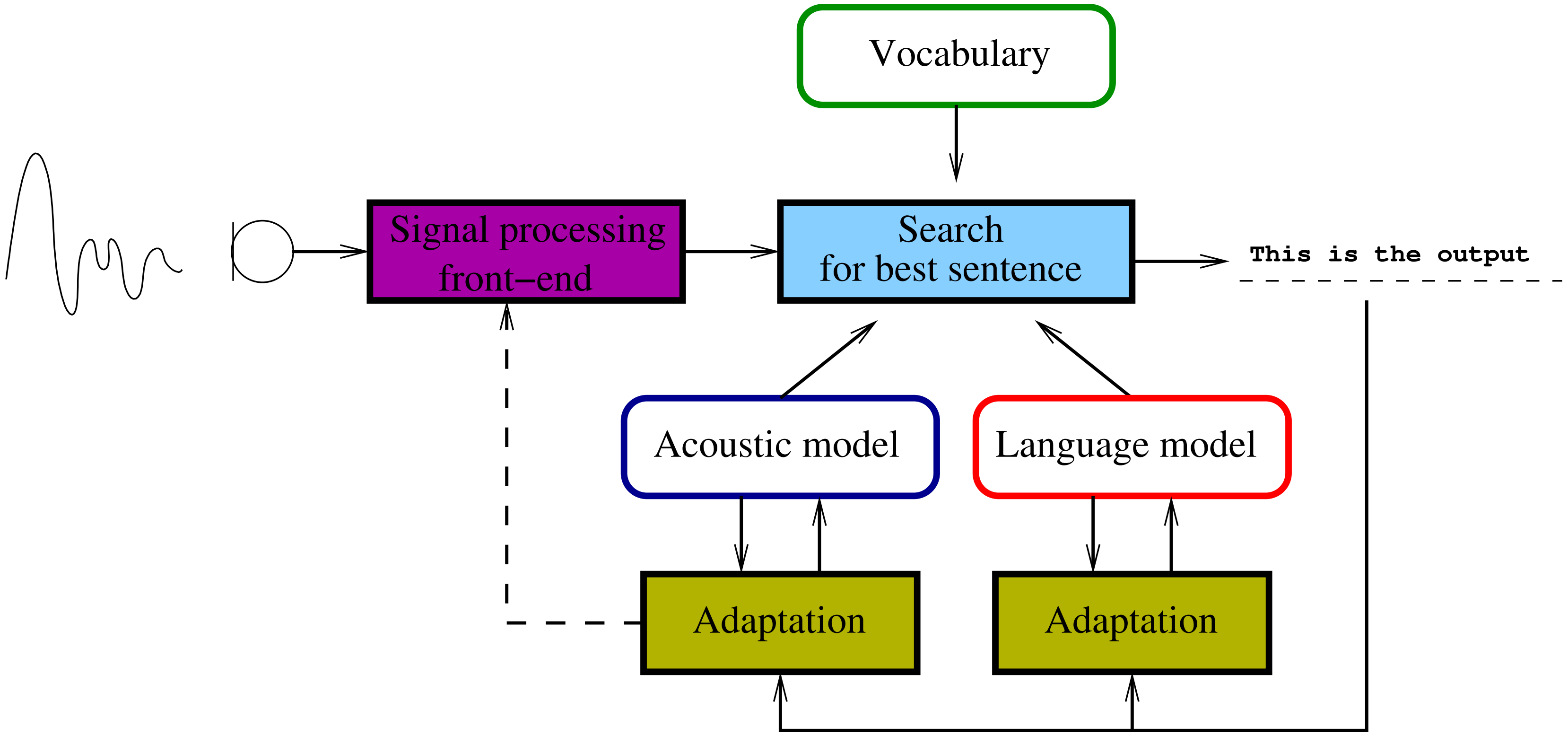


Thomas Hain
t.hain@sheffield.ac.uk
Spring Semester

SPandH
The University Of Sheffield.

Adaptation

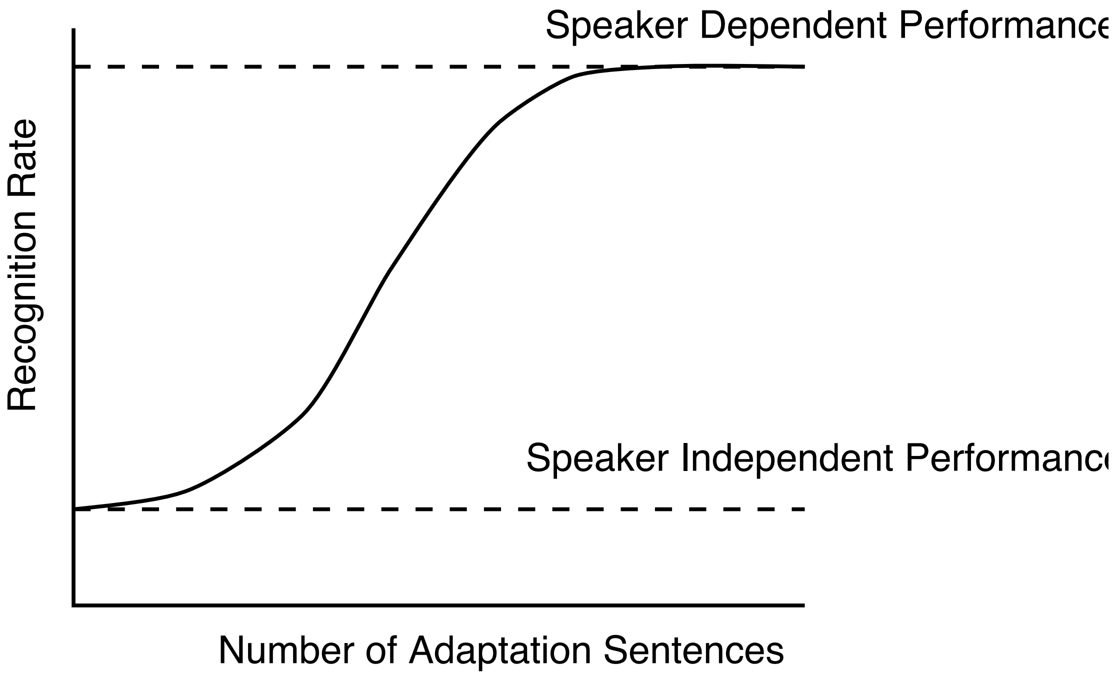
Recall: The output of a speech recognition system can be used to improve the performance in the current acoustic and linguistic condition.



Theoretically all knowledge source can be adapted, the vocabulary/dictionary, the acoustic and the language model. However, the most important techniques all target the variability of speech signals for different speakers. This class of adaptation techniques are commonly referred to as **speaker adaptation**. Other schemes are **environment adaptation**, **accent adaptation**, **topic adaptation**...

Speaker Adaptation

The aim of speaker adaptation is to improve the quality of the **speaker independent** (**SI**) speech recognition output by use of speaker specific information in the recognition process. With speaker specific **adaptation data** the components of the speech recogniser are changed to yield lower error rates. The system is called **speaker adaptive** (**SA**).



An adaptation scheme is good if

1. near speaker dependent (**SD**) performance can be achieved with sufficient data.
2. it is effective with small amounts of data.

Sometimes a distinction is made between adjusting the model parameters and modifying the features. For the latter we speak of **speaker normalisation**.

Speaker variations

It is common to distinguish between two types of speaker variations in speech signals:

- **Linguistic Differences**

  Accents account for large variations between speakers. For example the pronunciation of the word *tomato* in Received Pronunciation and general American. These effects may be handled in the lexicon as long as the change is consistent, often rules can be devised to predict the changes.

  In addition, there are idiosyncrasies associated with particular speakers. For example the pronunciation of the word *either*. Also, speaking rate is sometimes dependent on accent (can't be handled by lexicon).

- **Physiological Differences**

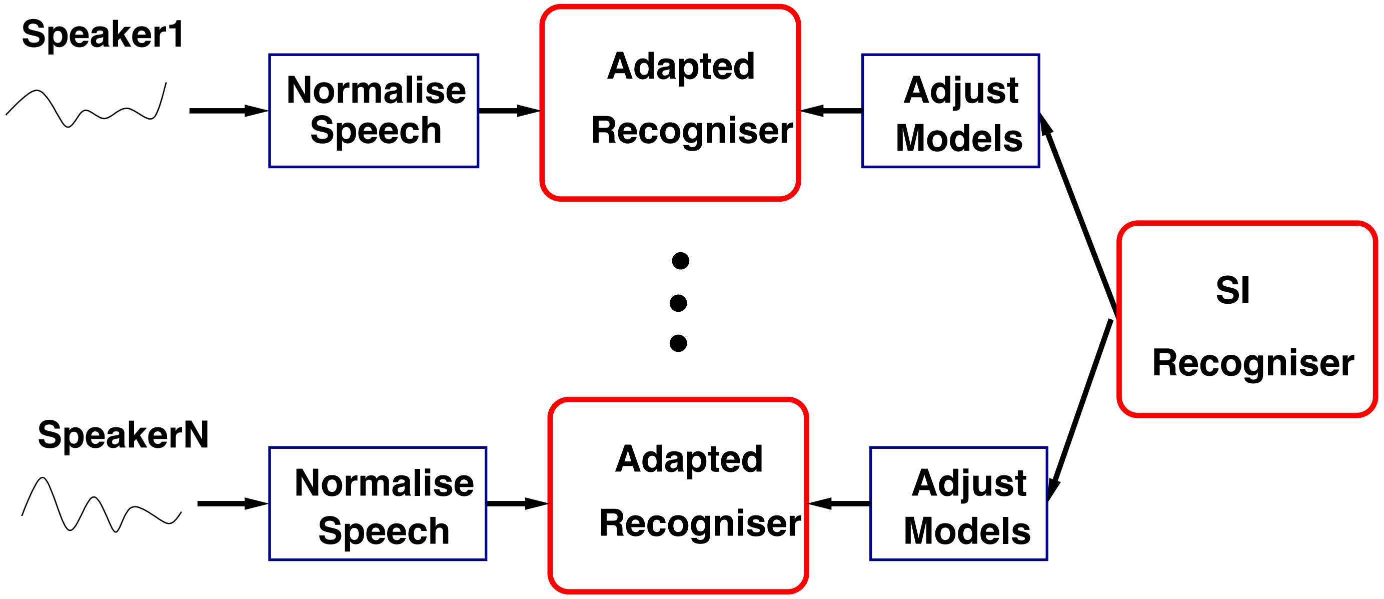
  The most obvious physiological difference is gender which has a profound impact on the whole structure of the speech signal (fundamental frequency, formants). Here we talk about **inter-speaker** variability. Various transitory effects will also alter the speech, for example having a cold. These effects increase the within a speaker, **intra-speaker**, variability.

Adaptation to the Speaker Acoustics

The aim of speaker adaptation is to rapidly adapt a recognition system to a particular speaker. A sample of speech from the new speaker is used to generate an adaptation mapping and all further processing of the speaker uses this mapping.

Types of acoustic adaptation:

1. **Speaker Classification**: An appropriate model set for a particular speaker is selected.
2. **Feature Normalisation**: Uses mappings of the feature vector space to make all speakers appear similar.
3. **Model Adaptation**: Attempts to re-estimate the model parameters for a particular speaker.



Modes

Speaker adaptation may be performed in a variety of different modes.

- **Supervised**: The transcription of the adaptation data is known
- **Unsupervised**: The transcription is unknown. If required it must therefore be estimated.
- **Static** (or Block): All of the adaptation data is presented before the final adapted system is produced.
- **Dynamic** (or Incremental): An adapted system is produced after only part of the adaptation data has been presented. The adapted system may be further refined as more adaptation data is presented.

The particular mode of adaptation has consequences for the adaptation scheme in terms of

- **Computational load**
  Is it necessary to update the models more than once?
- **Accuracy of adaptation**
  Do I need to estimate the transcription?

Speaker Clustering

A large number of model sets are generated, each one corresponding to a different speaker. The adaptation process is then to decide which model set the current speaker belongs to, a speaker identification problem.

In practice it is not possible to generate a model set for an individual speaker (training data limitations). Speakers are therefore clustered into similar **speaker groups** and model sets are trained for these speaker groups. During adaptation the most suitable speaker group is selected. Hence the speaker adaptation problem is separated into clustering task and a classification problem.

In practice there are problems with this approach.

1. The spectral variation of speakers, even within a cluster, may be large.
2. The new speaker may be not well represented by the training set of speakers.
3. It is assumed that the differences between speakers are uniform across a model set.

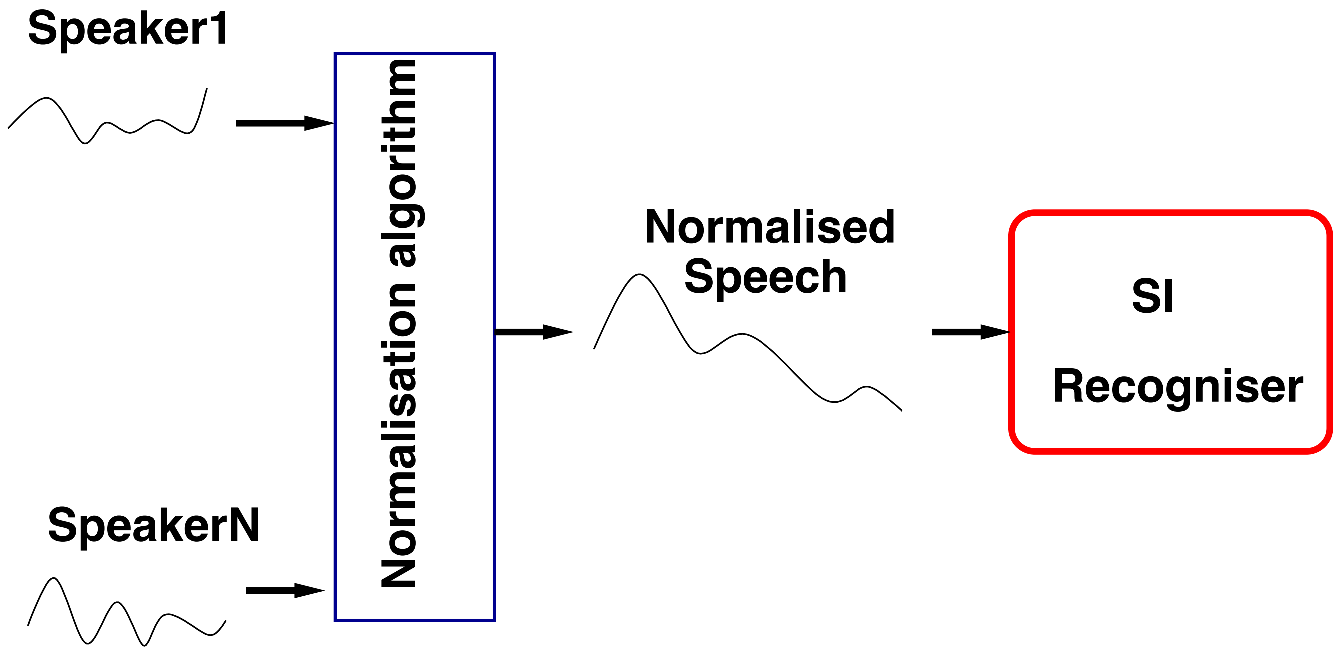
Improved techniques such as eigenvoices or cluster adaptive models exist ....

Feature normalisation

In feature normalisation parameters in the feature extraction are adjusted to optimal values for a certain speaker. The aim is to lower or even remove the inter-speaker variability of the speech signals. Note that with this approach normally high efficiency is reached, but the due the small number of parameters results are rarely close to SD performance.

Several techniques have been used. Among the most well known are

1. **Vocal Tract Length Normalisation**: The "length" of the vocal tract is estimated and spectral shifts calculated accordingly.
2. **Linear Cepstral Transformations**: A general linear transformation of the features is used.



Algorithmically both schemes are related (i.e. VTLN can be approximated by linear transforms).

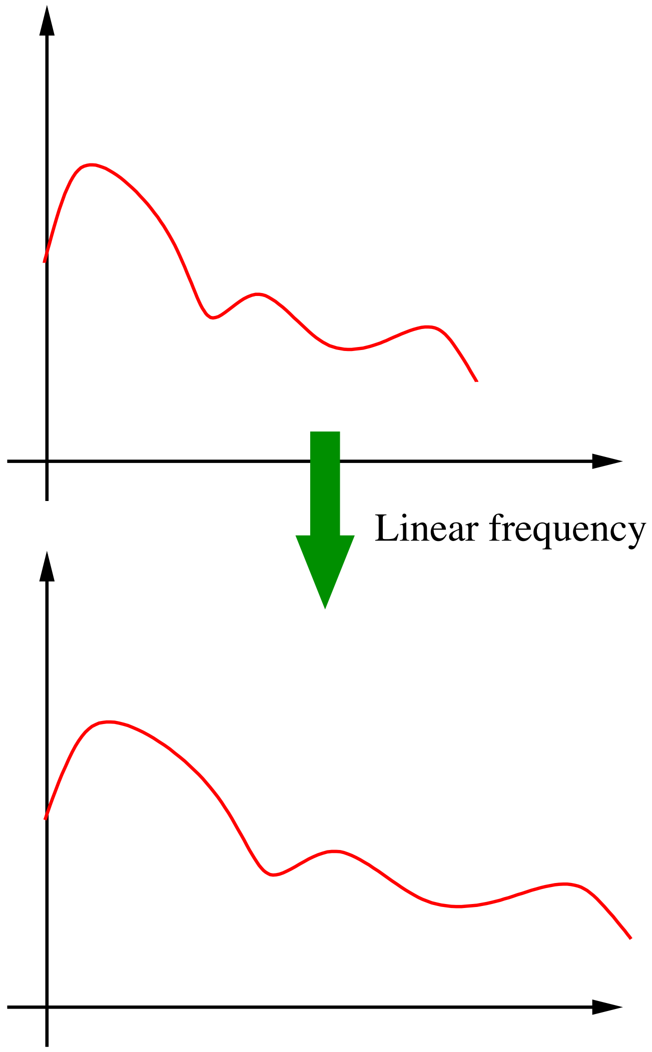
Vocal-Tract Length Normalisation

VTLN is a simple normalisation scheme that acts on the frequency axis. Schemes for estimating the warping factor include:

- **Discrete search**: A set of discrete warping factors are examined, typically in the range 0.88 to 1.12. The one yielding the maximum likelihood is selected.
- **Direct measurement** Found by (trying to) measure, for example, formant frequencies.

Mapping can be applied to data by

- **Time-domain resampling**: Resample the time-domain waveform using the new warping factor.
- **Filter-bank shifting**: Alter the canter frequency of the filterbanks according to the warping factor.
- **Cepstral transform**: similar effects can be achieved by directly manipulating the cepstra



Model-Based Adaptation

Model-based adaptation approaches have become increasingly popular. The aim is now to alter the parameters of, for example, a speaker independent model set to be more representative of a particular speaker.
Two approaches are commonly used

1. **MAP**
   Maximum A-Posteriori (MAP) approaches have been extensively used for speaker adaptation (also used for smoothing).
2. **Linear Regression**
   A linear transformation is applied to the Gaussian mean (and possibly variance) parameters.

One problem that occurs with these model adaptation approaches is that of **unobserved Gaussians** i.e. there are no frames of adaptation data that correspond to a particular part of the model. This is a general problem in speaker adaptation. When dealing with small amounts of adaptation data, few (sometimes very few) Gaussians will be observed in the adaptation data.

MAP Estimation

In Maximum Likelihood training the set of model parameters $\mathcal{M}$ are found by maximising the likelihood $p(\mathbf{O}|\mathcal{M})$ of the training data $\mathbf{O}$. MAP Estimation attempts to find the model parameters by optimising

$$p(\mathcal{M}|\mathbf{O}) = \frac{p(\mathbf{O}|\mathcal{M})p_0(\mathcal{M})}{p(\mathbf{O})}$$

The influence of the **prior**, $p_0(\mathcal{M})$, may be explicitly seen. When a non-informative prior is used the MAP estimate becomes the ML estimate. MAP estimation techniques have been used to adapt the means, variances and mixture weights for CDHMM systems. Here, only **mean adaptation** will be described.

MAP estimation will, given sufficient adaptation data, yield the same performance as a speaker-dependent system! The MAP estimate of the mean of state j is given by

$$\mu_{MAP} = \frac{\sigma_j^2 \mu_{pj} + \sigma_{pj}^2 \sum_{t=1}^{T} L_j(t) o(t)}{\sigma_j^2 + \sigma_{pj}^2 \sum_{t=1}^{T} L_j(t)}$$

where σ_j^2 is the assumed known variance, μ_{pj} and σ_{pj}^2 are the prior mean and variance of the mean, and $L_i(t) = P(q(t) = i|\mathbf{O})$ is the state level posterior probability.

MAP for smoothing and adaptation

One needs to determine the prior mean and variance, μ_{pj} and σ^2_{pj}. When used for smoothing these were simply, for example, the context independent model parameters. In speaker adaptation they may be estimated from S speaker dependent (or multi-speaker) models.

$$\mu_{pj} \quad = \quad \sum_{s=1}^{S} c_s \mu_j^{(s)} \qquad\qquad \sigma^2_{pj} = \sum_{s=1}^{S} c_s (\mu_j^{(s)} - \mu_{pj})^2$$

where c_k is the "weight" assigned to the k^{th} cluster state.

Alternatively the ratio $\frac{\sigma^2_j}{\sigma^2_{pj}} = \tau$ is set. Values of τ in the range 1-100 are normally used.

This **only** allows **observed Gaussians** to be updated. However, it is known that high correlations exist between various states of the model sets and various modifications to MAP have been suggested to handle this.

Least Squares Linear Regression

A second class of model adaptation is based on a linear transform of the model parameters. Thus

$$\hat{\boldsymbol{\mu}}_j^{(s)} = \boldsymbol{A}^{(s)}\boldsymbol{\mu}_j + \boldsymbol{b}^{(s)}$$

where $\boldsymbol{A}^{(s)}$ is an $n \times n$ matrix and $\boldsymbol{b}^{(s)}$ is an $n \times 1$ vector. These describe the transform from the SI model to the particular speaker. This is sometimes written as

$$\hat{\boldsymbol{\mu}}_j^{(s)} = \boldsymbol{W}^{(s)}\boldsymbol{\xi}_j \qquad\qquad \boldsymbol{\xi} = \left[\begin{array}{cccc} 1 & \mu_{j1} & \dots & \mu_{jn} \end{array}\right]^T$$

where $\boldsymbol{W}^{(s)}$ is an $n \times (n+1)$ matrix and $\boldsymbol{\xi}_j$ is the extended mean vector.

Again there is the problem of unobserved Gaussians and also the **problem of limited training data** for each transformation. It is therefore necessary to tie the transformations over a set of Gaussians. Initially consider a single transform for all Gaussians.

One simple estimation scheme for $\boldsymbol{W}$ is *least squares*. This has a simple solution that

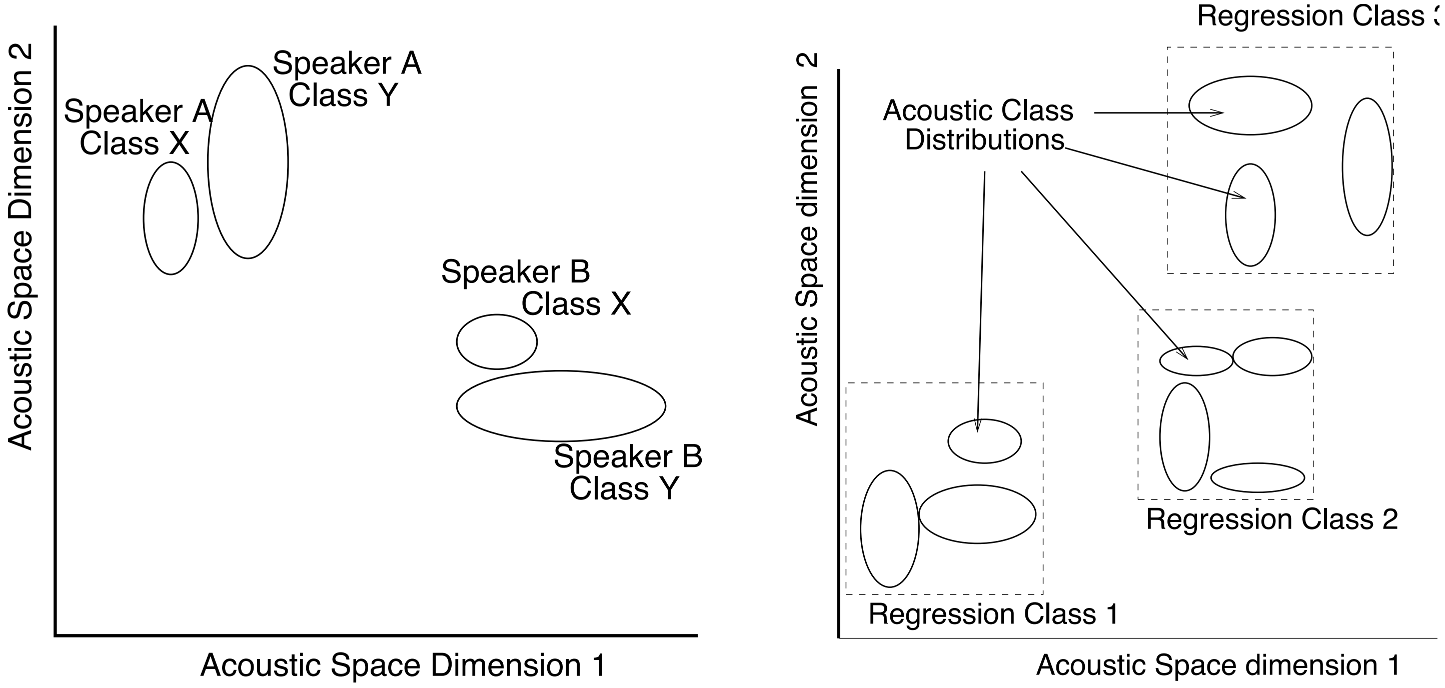
$$\boldsymbol{W}^{(s)} = \left(\sum_{j=1}^{M}\sum_{t=1}^{T} L_j(t)\boldsymbol{o}(t)\boldsymbol{\xi}_j^T\right) \times \left(\sum_{j=1}^{M}\sum_{t=1}^{T} L_j(t)\boldsymbol{\xi}_j\boldsymbol{\xi}_j^T\right)^{-1}$$

Maximum Likelihood Linear Regression

The transformation matrix $\boldsymbol{W}^{(s)}$ may also be estimated such that the likelihood of the adaptation data is maximised (hence Maximum Likelihood Linear Regression). The Expectation-Maximisation Algorithm can be used for this purpose. (Note: LSLR is identical to MLLR when $\boldsymbol{\Sigma}_j = \boldsymbol{I}$ for all states).

MLLR has been found to outperform LSLR (and naturally fits into iterative EM schemes).

Rather than using a single transform multiple transforms may be used (using so-called regression class trees).



A matrix multiplication is essentially a rotation and scaling operation. Hence, for all classes associated with single transform (for example all triphones with a certain centre phone), we assume that if the two classes are “close” in acoustic space for speaker A, then they will also be “close” in acoustic space for speaker B.

Transformation Matrices

The threshold used in the regression class tree will be a function of the number of parameters in the linear transformation.

- ➜ **Simple Offset**: The transformation is

$$\hat{\boldsymbol{\mu}}_j^{(s)} = \boldsymbol{\mu}_j + \boldsymbol{b}^{(s)}$$

  Number of parameters is n.

- ➜ **Diagonal**: Only the leading diagonal of the matrix and the bias are non-zero. Here all elements are independent of one another.

$$\hat{\mu}_{ji}^{(s)} = a_{ii}^{(s)} \mu_{ji} + b_i^{(s)}$$

Number of parameters is $2n$.

Transformation Matrices (2)

- **Full**: All elements of the matrix may be non-zero. Number of parameters is $n(n+1)$.
- **Block Diagonal**: The matrix has a form like

$$\mathbf{A}^{(s)} = \begin{pmatrix} \mathbf{A}_s^{(s)} & \mathbf{0}^n & \mathbf{0}^n \\ \mathbf{0}^n & \mathbf{A}_\Delta^{(s)} & \mathbf{0}^n \\ \mathbf{0}^n & \mathbf{0}^n & \mathbf{A}_{\Delta^2}^{(s)} \end{pmatrix}$$

  The number of parameters, assuming a split into 3 equal blocks and including the bias is $\left(\frac{n}{3}\right)^2 + n$

Block diagonal and full transformation matrices have been found to consistently outperform the diagonal case.

Performance

1. **Test Set**: Unlimited vocabulary, two sets of 20 unknown speakers uttering about 15 sentences each (Dev and Eval), “clean” environment.
2. **Adaptation**: Incremental MLLR adaptation using full transforms and a regression class tree to determine number of transforms.

System	Adaptation	WER (%)	
	MLLR	Dev Data	Eval Data
GI	-	9.5	9.2
GD	-	9.2	8.6
GI	μ	8.0	8.3
GD	μ, Σ	7.9	8.1

Source CU-HTK WSJ system (Phil Woodland/Mark Gales)

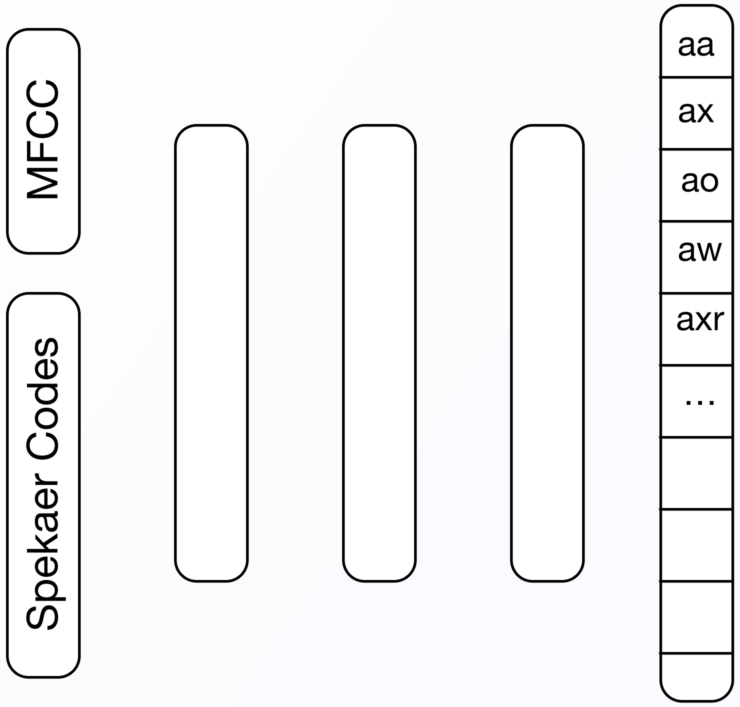
Adaptation for Neural Networks

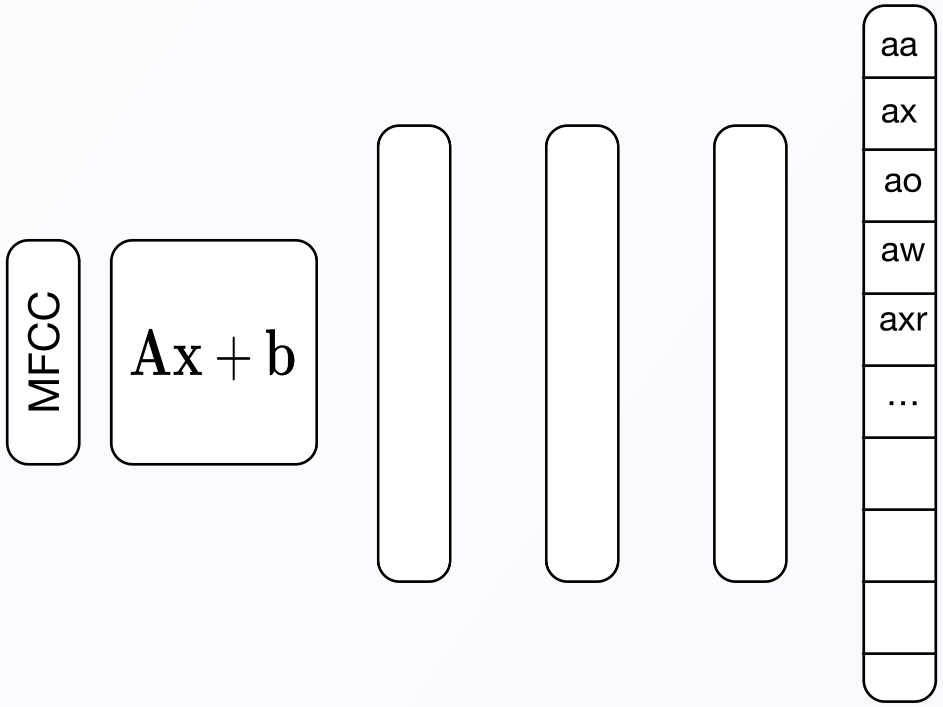
- Similar principles hold:
  - Normalisation:
    Input normalisation using the same techniques as for HMM / GMMs, i.e. Matrices
  - Model adaptation:
    Providing input to the model can use for adaptation.
  - Adaptive training:
    Change parts of the model depending on speakers.

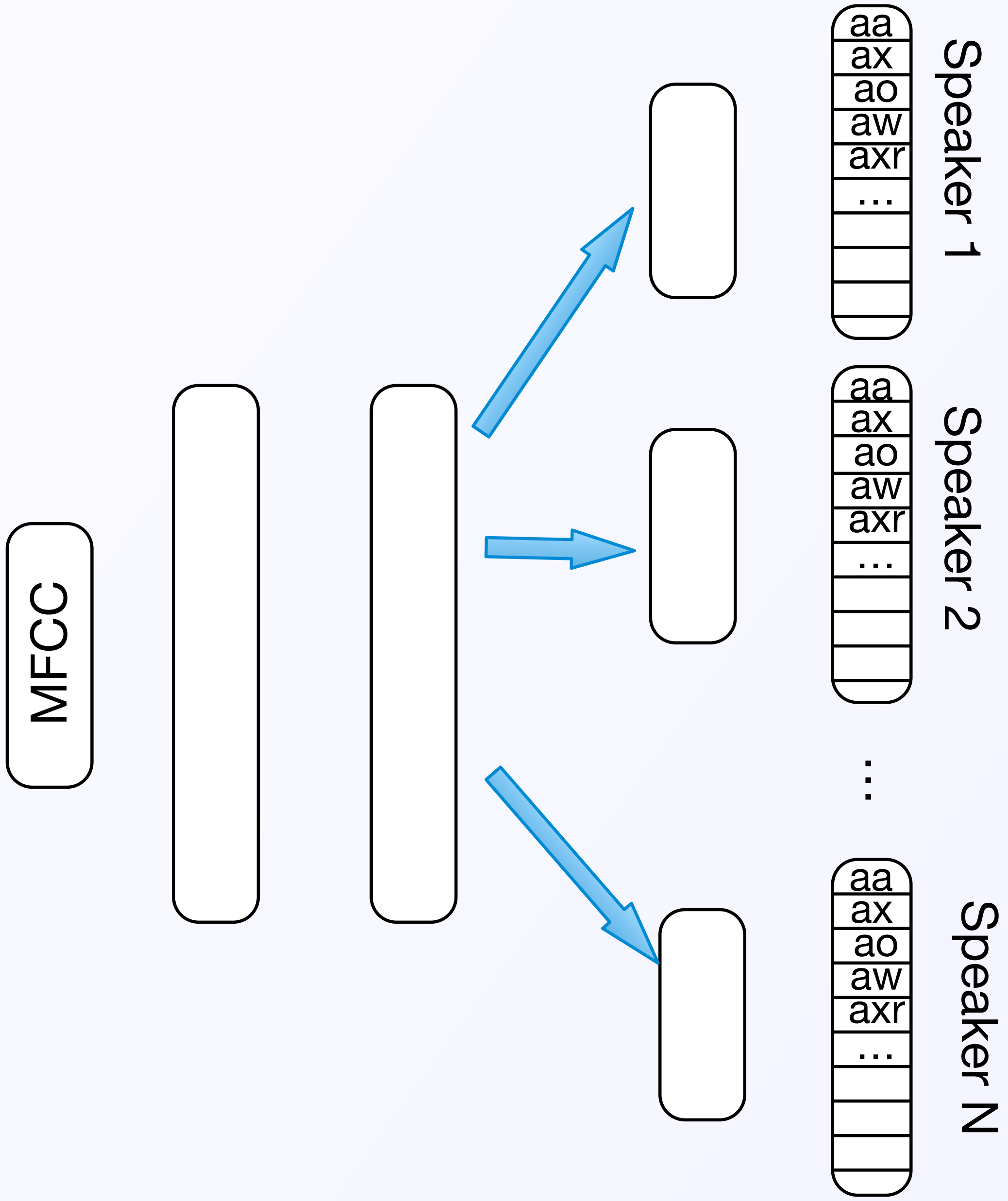
Speaker representations

- Additional input into neural networks
- If the speakers are known
  - 1 out of N encoding of speaker identity (1 hot)
- If the speakers are unknown
  - Vector representation of speaker idenity
  - Speakers with close acoustic properties should be close in vector space.
    - iVectors
      N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. Audio, Speechand Language Processing, vol. 19, no. 4, May 2011.
    - xVectors
      D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," ICASSP 2018, Calgary, AB, 2018, pp. 5329-5333.
    - *HVectors*
      Yanpei Shi, Qiang Huang, Thomas Hain (2018). H-VECTORS: Utterance-level Speaker Embedding Using A Hierarchical Attention Model. In Proc. ICASSP 2020.

Adaptation of (Deep) Neural Networks







Other adaptation schemes

- ➜ **Language model adaptation**

  The adaptation of language models is more difficult due to considerable data sparsity effects. For example only a few trigrams would be observed in the adaptation data. Hence the effect on a model with many millions of parameters may be minor. One particular simple adaptation scheme is the use of **unigram caches**. Here a fixed size window over recently recognised words is used to obtain a word unigram distribution. This is then combined with the standard (possibly trigram) language model. Despite the fact that considerable reductions in perplexity can be observed (10-20%) the impact on WER usually remains low.

- ➜ **Lexicon adaptation**

  Schemes for lexicon adaptation are rarely used due to data sparsity effects. In theory the dictionary can be modified to allow for speaker specific accents. As in practice the acoustic models and the dictionary both encode this type of information it is often more effective to use standard acoustic model adaptation such as MLLR.