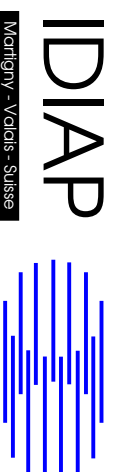


# Hard Problems in ASR

Hervé Bourlard

Swiss Federal Institute of Technology at Lausanne (EPFL)

IDIAP, Martigny, Switzerland



IMA Wokshop, September 18-22, 20000

## What I will not talk about

- Suitable levels of representation
- Mechanisms for learning
- Human-machine communication
- Speech disfluencies
- Language modeling

## What I will talk about

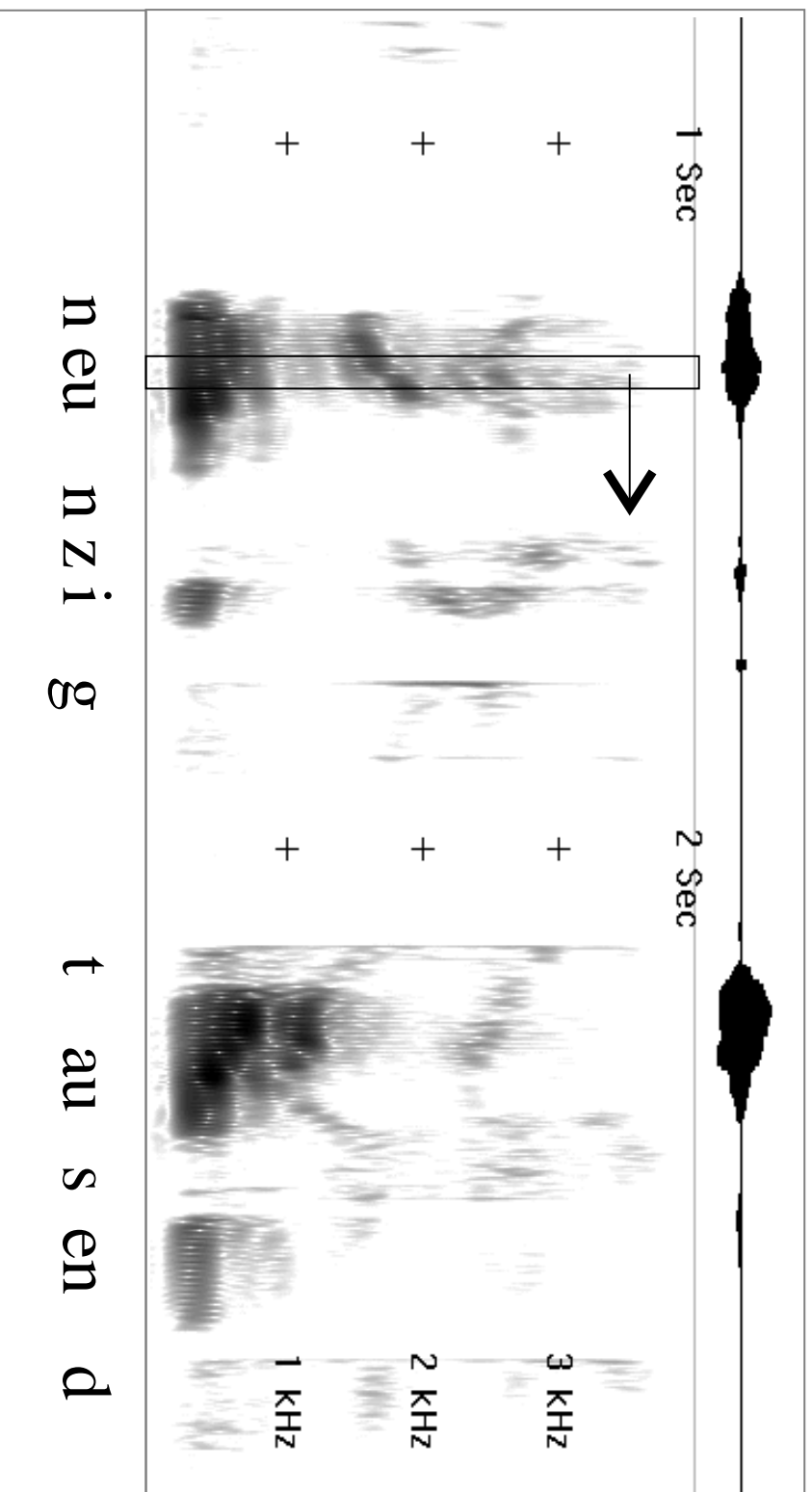
- Some of the simplest of the hard problems...
- “We should not look at ASR as a simple pattern recognition problem”

## What we have got so far

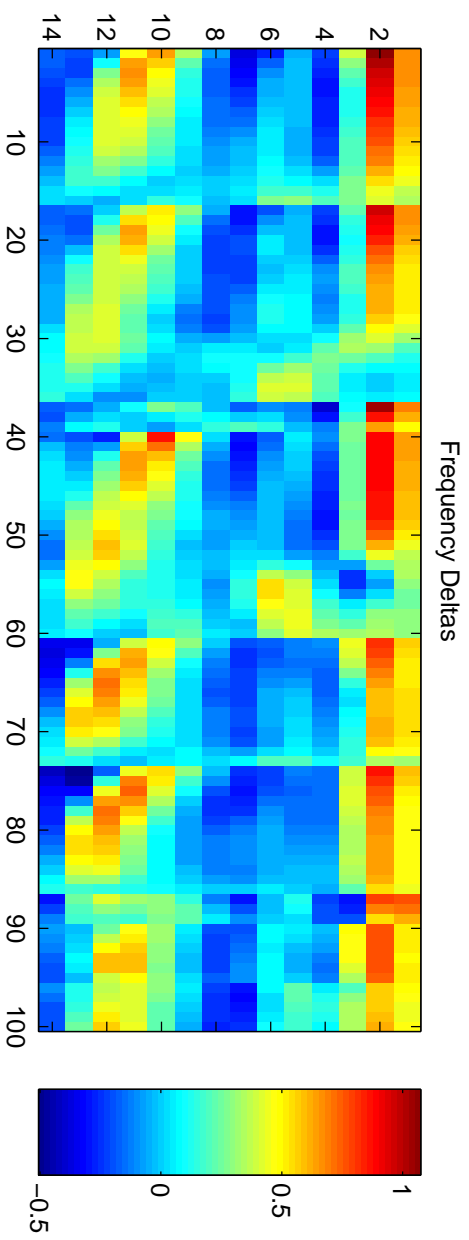
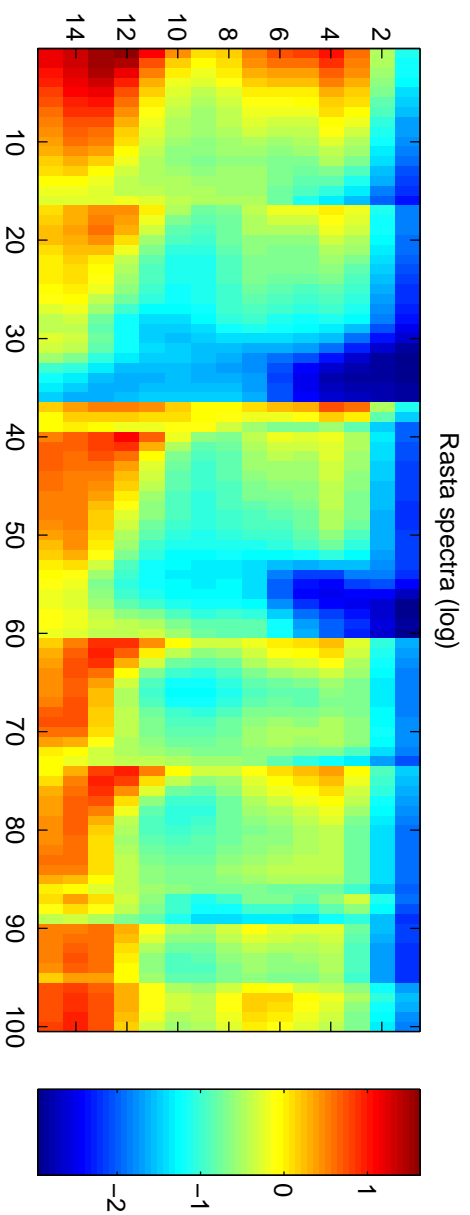
- A good mathematical formalism (HMM) and a powerful optimization algorithm (EM):
  - Integrates acoustic and linguistic *priors*: BUT those priors are not well known and, theoretically as well as empirically, are strongly conditioned on the acoustic.
  - Ignorance based modeling; performance depends on stationary properties and generalization to unseen data.
  - Units/levels *implicit* in model structure: BUT we do not know the right model structure (units/levels).

- Mathematical definition of recognition:
  - Most likely interpretation; efficient decoding (search) algorithms
  - BUT: best fit / most likely explanation of *any* input; will hallucinate
  - If the right acoustic features are used (additivity), signal decomposition/combination possible, but not widely used.

# Just a look at the signal...



# Another look at the signal...



# SIGNAL PROCESSING

Ideal speech features: reversibility (preserving useful information), additivity (in time domain), smoothness (especially in the case of AR modeling).

- AR coefficients do not preserve time-domain additivity (idem MFCC and LSP coeff)
- Reflection coefficients  $k$ 's are smooth and reversible but they are not additive (nonlinear functions involved in their calculation)
- DFT is reversible and additive, but not really smooth... and doesn't yield optimal performance.

## Signal processing

- Bad news for ML based ASR: orthogonalized data likelihood increases with noise level!
- ASR should be able to dynamically extract and deal with multiple (possibly correlated) data streams exhibiting different properties

# ACOUSTIC MODELING

- Speech signals are complex dynamical systems, which cannot be considered as piecewise stationary...
- Standard HMMs model dynamics and formant trajectories through:
  - Context dependencies
  - Increasing the numbers of Gaussians PDFs
  - Increasing the number of HMM statesresulting in a constantly increasing need for more training data.

- Noise robustness can only be achieved by increasing the training material or estimating/adapting the PDFs to the characteristics of the current noisy data.
- What we need:
  - Principled approach to model *conditional* densities
  - “Optimal” decomposition and combination into (piecewise stationary) data streams, still able to deal with their possible (temporal and accross stream) correlation, and performing ASR based on the most reliable features.

## Product of Errors Rule

Let  $X = (X^1, X^2)$  be two data streams

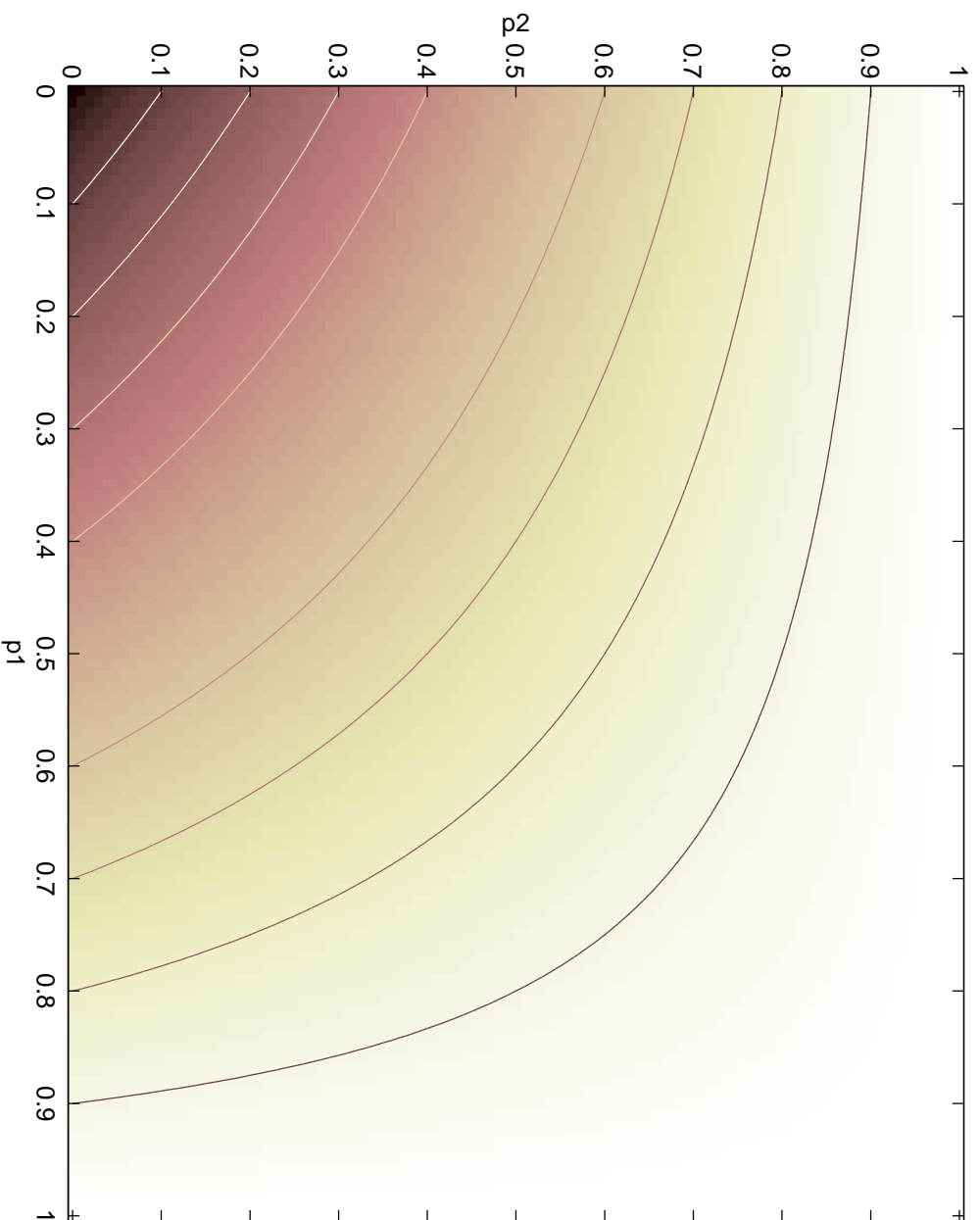
$$\begin{aligned} e_T(q_j | X^1, X^2) &= e_1(q_j | X^1) e_2(q_j | X^2) \\ &= (1 - P_1(q_j | X^1))(1 - P_2(q_j | X^2)) \end{aligned}$$

Consequently:

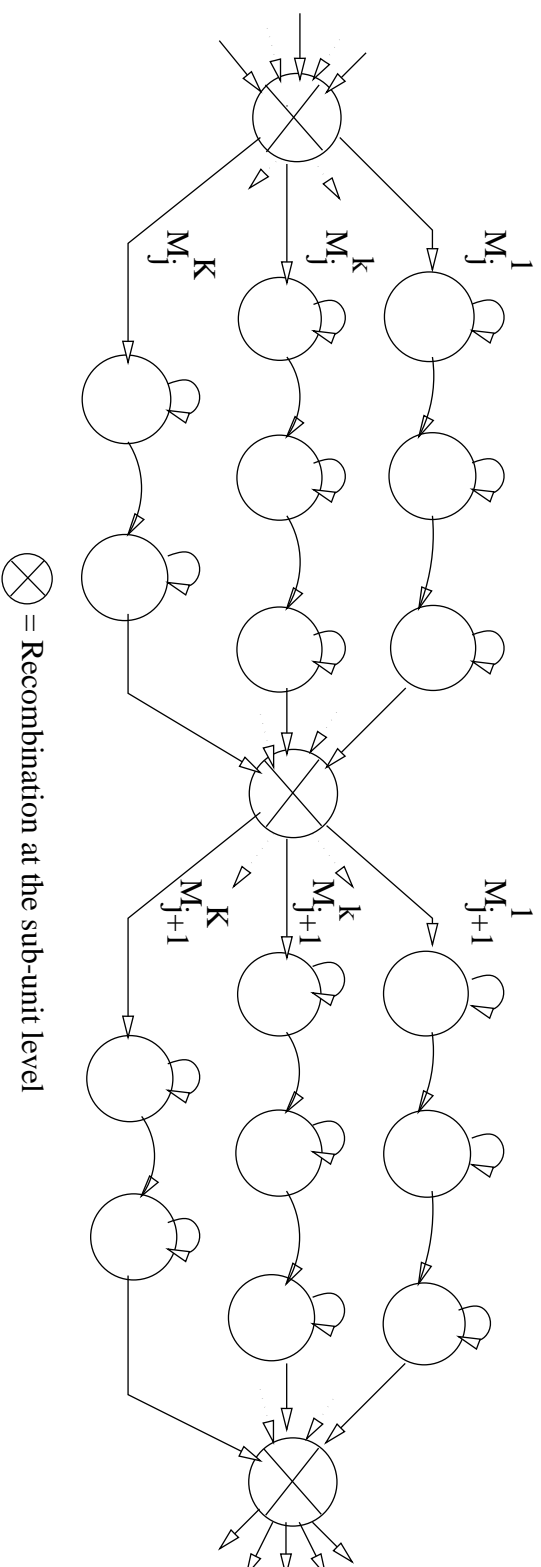
$$\begin{aligned} P_T(q_j | X^1, X^2) &= 1 - e_T(q_j | X^1, X^2) \\ &= 1 - \sum_{k=1}^2 P_k(q_j | X^k) + \prod_{k=1}^2 P_k(q_j | X^k) \end{aligned}$$

This rule can easily be generalized to  $K$  streams, yielding  $2^K - 1$  terms.

# Product of Errors Rule



# Multiband/multistream processing and recognition



Combination rule:

If:

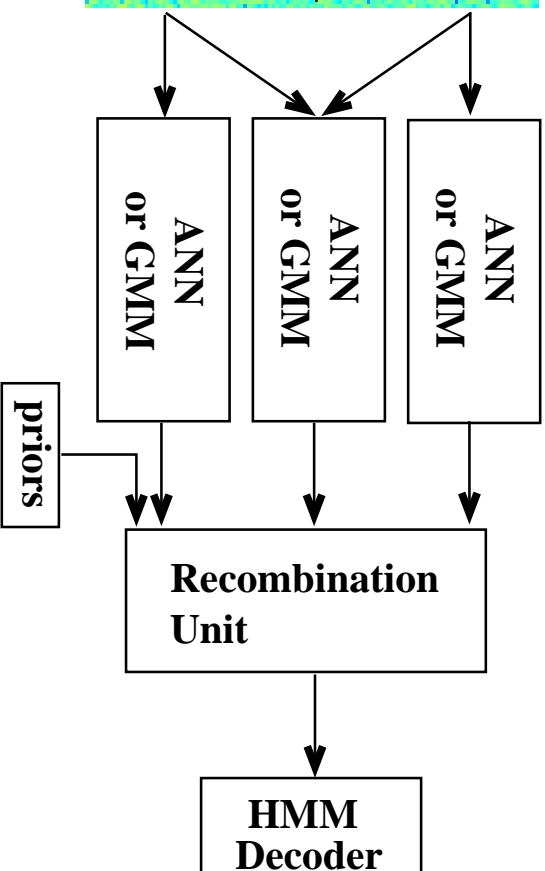
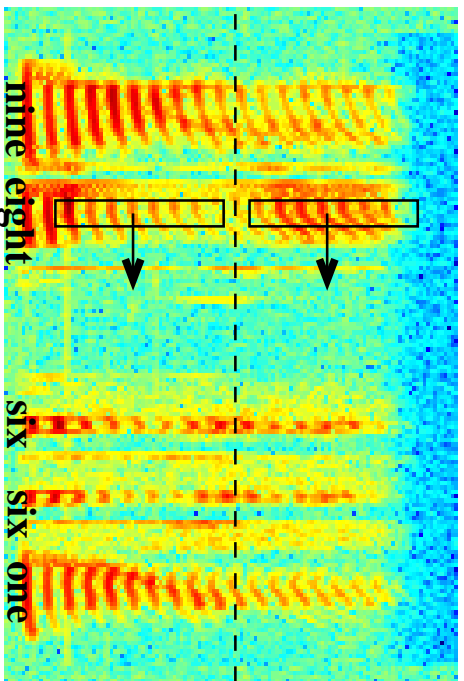
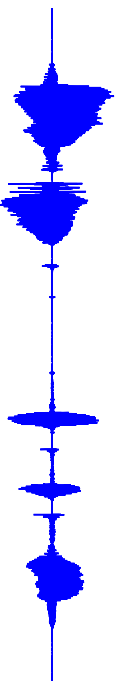
- $K$  = number of streams
- $\mathcal{S} = \{S^\ell\}$  is the set of all possible ( $L = 2^K$ ) subband combinations, including the empty set

Then:

$$P(q_j|x, \Theta) = \sum_{\ell=1}^L P(q_j, E_\ell|x, \Theta) = \sum_{\ell=1}^L P(q_j|S^\ell, \Theta_\ell)P(E_\ell|x)$$

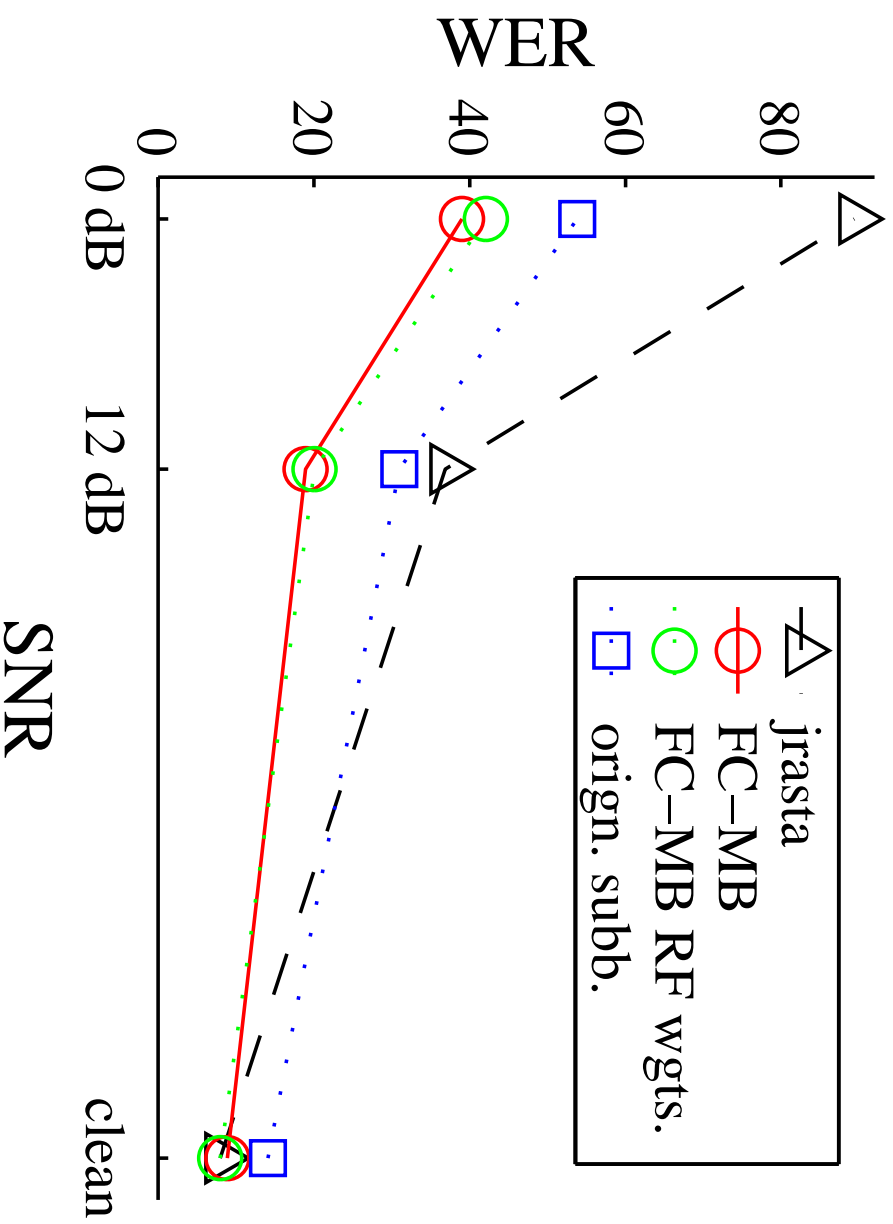
... similar to product of errors rule.

# Multiband processing and recognition

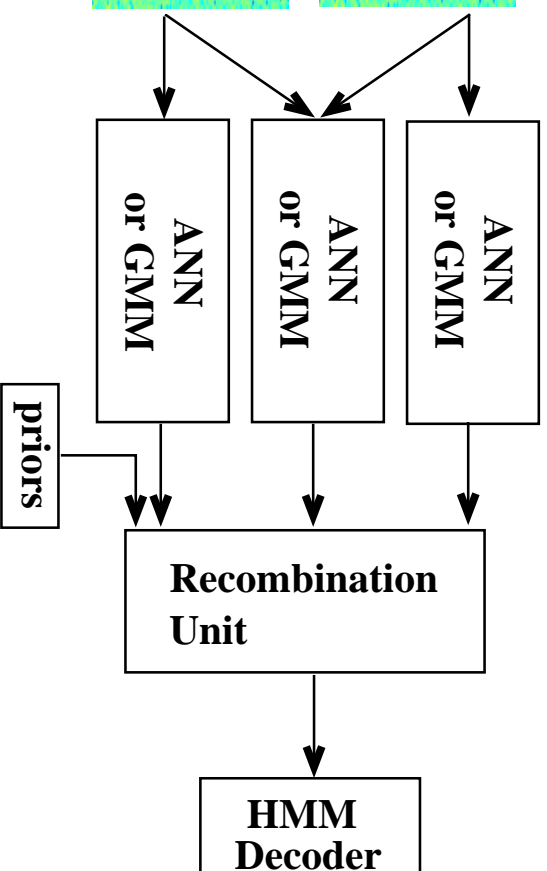
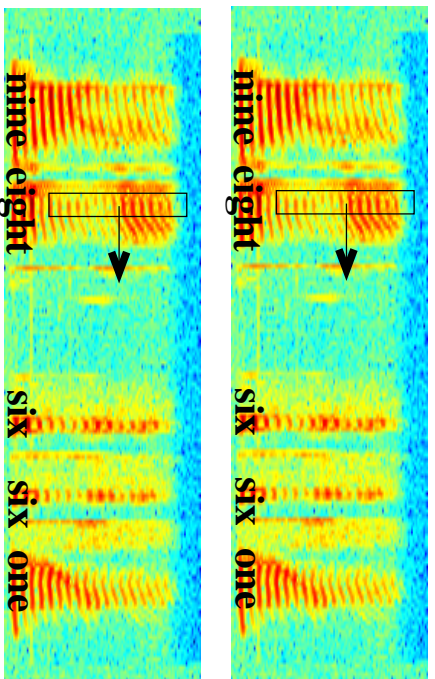
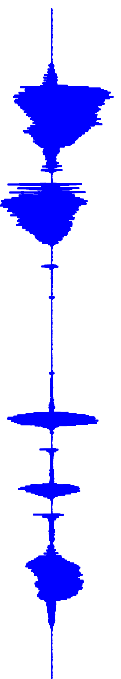


# Typical WER

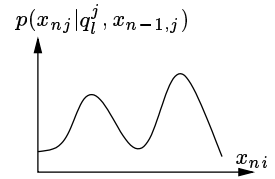
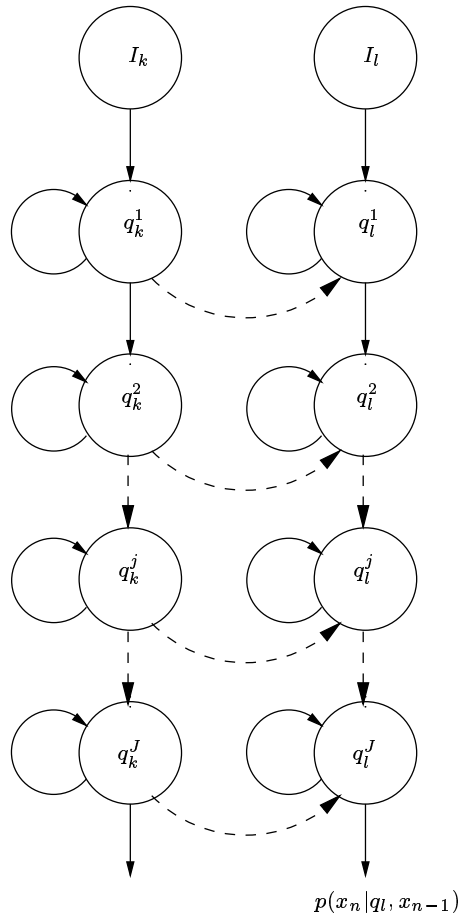
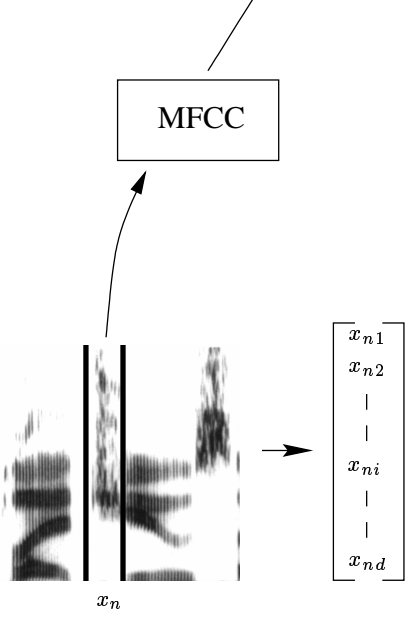
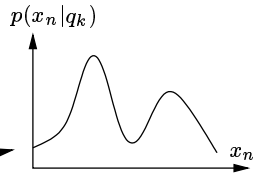
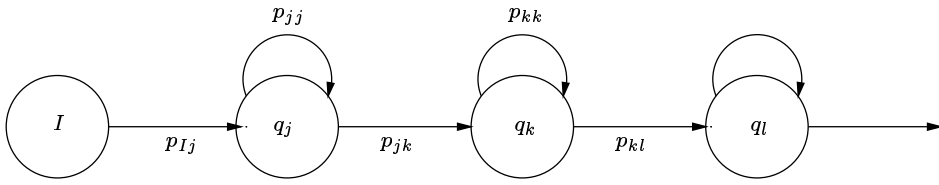
## Siren Noise



# Multistream processing and recognition



# **HMM12: Modeling the temporal and frequency structure**



## Advantages of HMMV2

- Automatic (nonlinear) spectral warping (vocal tract normalization)
- “Adaptive” mel scale
- Formant tracking
- Easier to include long term features
- Optimal “stream merging” during training and recognition to maximize likelihood
- In case of subband recognition: dynamic subband processing and combination
- Easy control of the number of parameters

# LEXICAL MODELLING

Current lexical models have deficiencies:

- Elementary sub-units (phones) are not appropriate...
- Modeling pronunciation variants based on ML criterion doesn't work well... pronunciation variants should be *conditioned on the acoustic parameters* (speaking rate, pitch, etc)... BUT HMMs are not really good at modeling conditional densities.
- For difficult cases, phonetic (or frame) recognition rates can decrease after having added lexical (and syntactical) constraints!!!
- How well could we do if we are provided ALL the information (acoustics, models, and *actual* phonetic transcription) but the word sequence? Cf. Fred Jelinek: ASR as a code breaking activity...

# LANGUAGE MODELLING

- This is a VERY HARD problem!!!
- *Language models should be conditioned on the acoustics*
- “Fudge factor”:
  - Should be different for each utterance (N best hypotheses or N recognizers?).
  - Should be a factor of (1) the reliability of the acoustic model and (2) the reliability of the LM, for each specific utterance.

- Language modeling can be considered as a mixture of experts problem:

$$\begin{aligned} P(M|X, \mathcal{L}) &= P(M, E_X|X, \mathcal{L}) + P(M, E_{\mathcal{L}}|X, \mathcal{L}) \\ &\approx P(E_X|X, \mathcal{L})P(M|E_X, X) + \\ &\quad P(E_{\mathcal{L}}|X, \mathcal{L})P(M|E_{\mathcal{L}}, \mathcal{L}) \end{aligned}$$