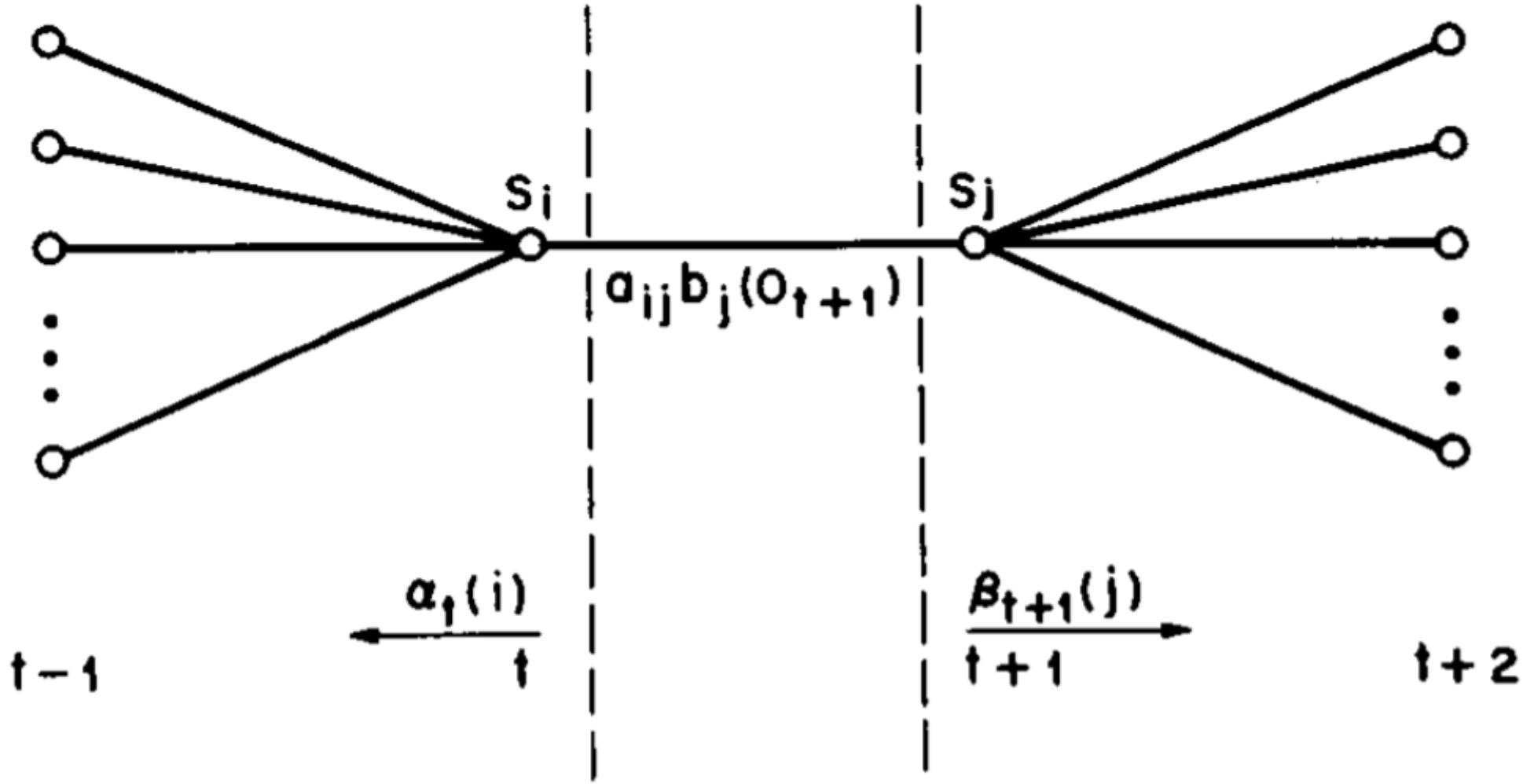

S_i
S_j
$a_{ij}b_j(O_{t+1})$
$\alpha_t(i)$
$\beta_{t+1}(j)$
$t-1$
t
$t+1$
$t+2$

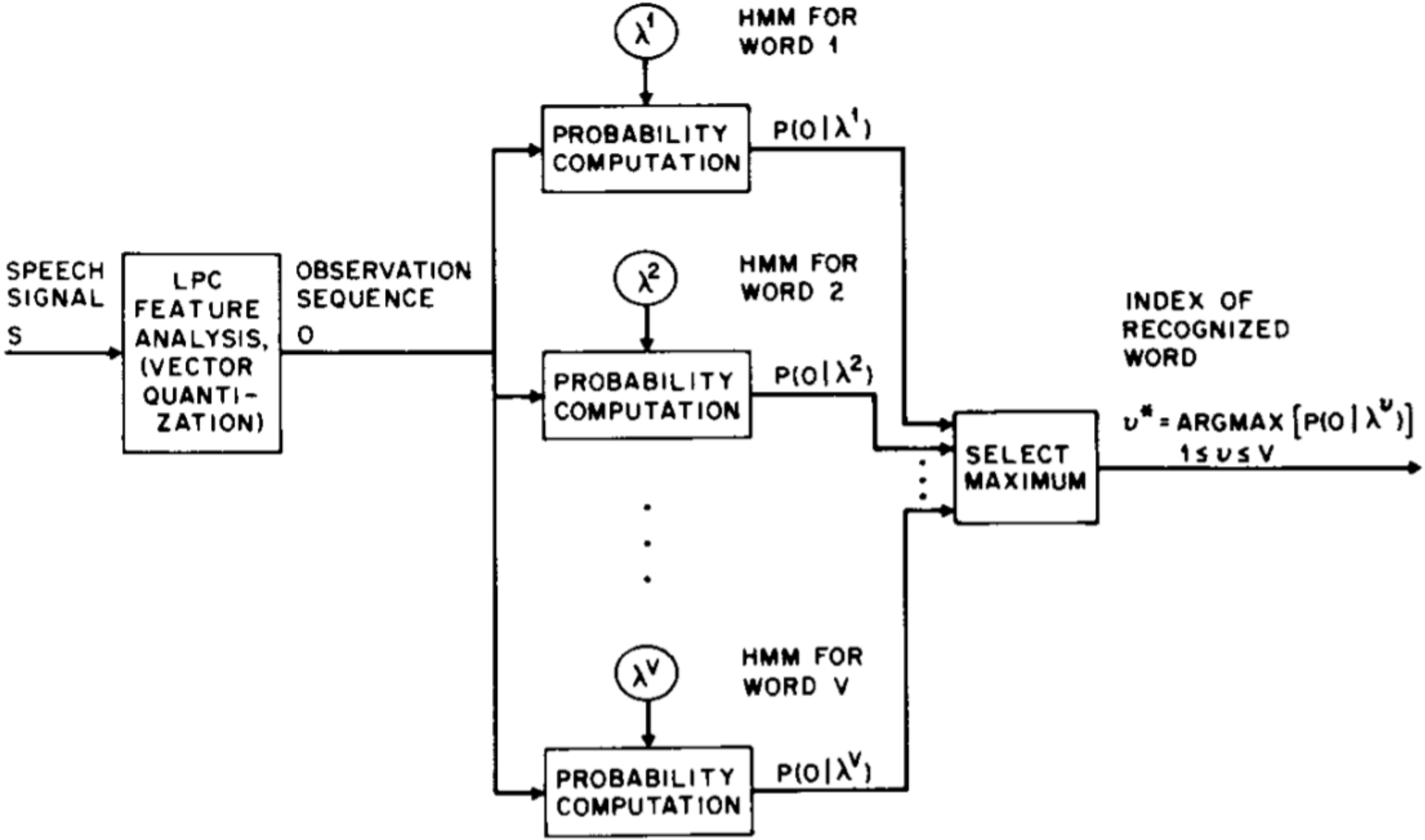
$$\bar{c}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j, k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j, k)}$$

$$\bar{\boldsymbol{U}}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j, k) \cdot (\boldsymbol{O}_t - \boldsymbol{\mu}_{jk})(\boldsymbol{O}_t - \boldsymbol{\mu}_{jk})'}{\sum_{t=1}^{T} \gamma_t(j, k)}$$

$$\bar{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j, k) \cdot \boldsymbol{O}_t}{\sum_{t=1}^{T} \gamma_t(j, k)}$$

$$\gamma_t(j, k) = \left[\frac{\alpha_t(j)\, \beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\, \beta_t(j)}\right]\left[\frac{c_{jk} \mathfrak{N}(\boldsymbol{O}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{U}_{jk})}{\sum_{m=1}^{M} c_{jm} \mathfrak{N}(\boldsymbol{O}_t, \boldsymbol{\mu}_{jm}, \boldsymbol{U}_{jm})}\right].$$

(The term $\gamma_t(j, k)$ generalizes to $\gamma_t(j)$ in the case of a simple mixture, or a discrete density.)


SPEECH SIGNAL
S
LPC FEATURE ANALYSIS, (VECTOR QUANTI-ZATION)
OBSERVATION SEQUENCE
O
λ^1
HMM FOR WORD 1
PROBABILITY COMPUTATION
$P(O|\lambda^1)$
λ^2
HMM FOR WORD 2
PROBABILITY COMPUTATION
$P(O|\lambda^2)$
λ^V
HMM FOR WORD V
PROBABILITY COMPUTATION
$P(O|\lambda^V)$
SELECT MAXIMUM
INDEX OF RECOGNIZED WORD
$v^* = \text{ARGMAX}_{1 \le v \le V} [P(O|\lambda^v)]$

Probabilistic approach – statistical sequence recognition

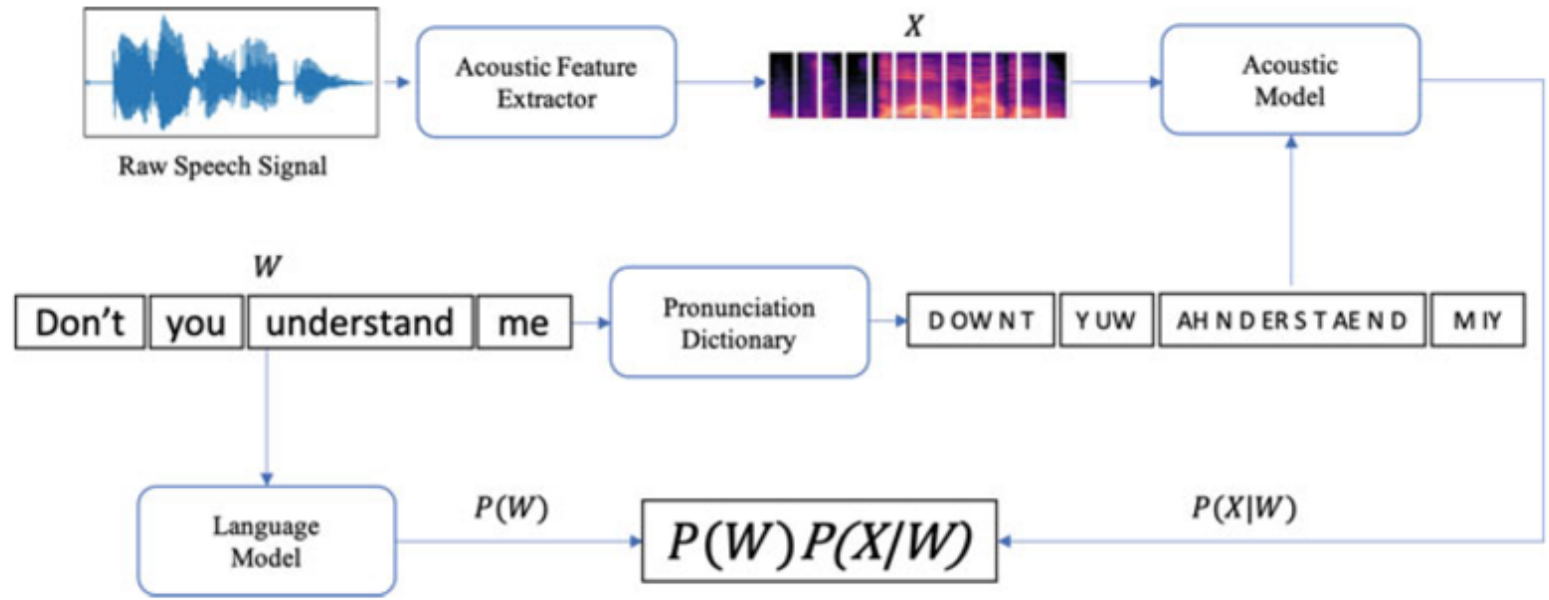
Using Bayes' Theorem

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)}$$

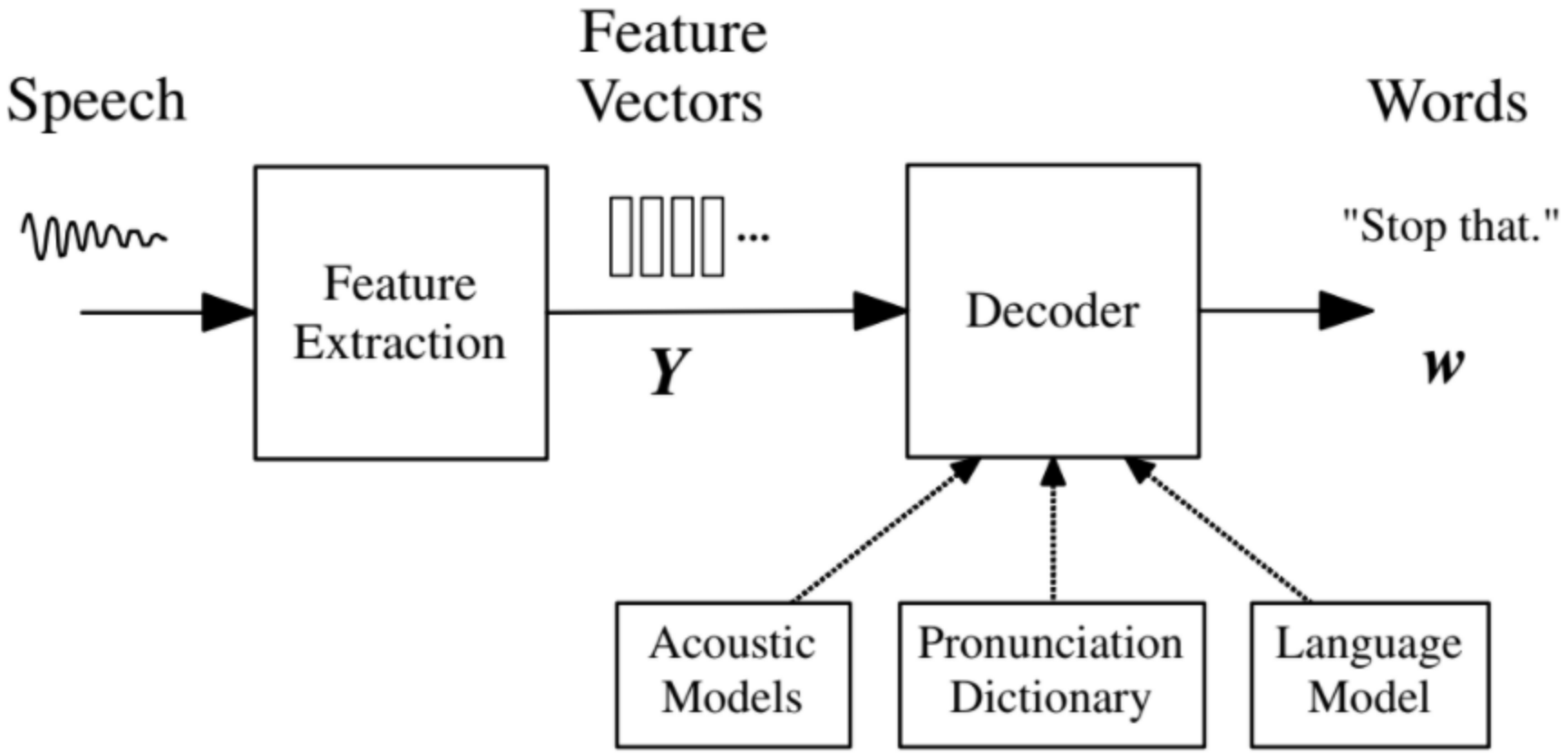
$$W^* = \underset{W \in V^*}{\operatorname{argmax}} P(X|W)P(W)$$

Acoustic Model

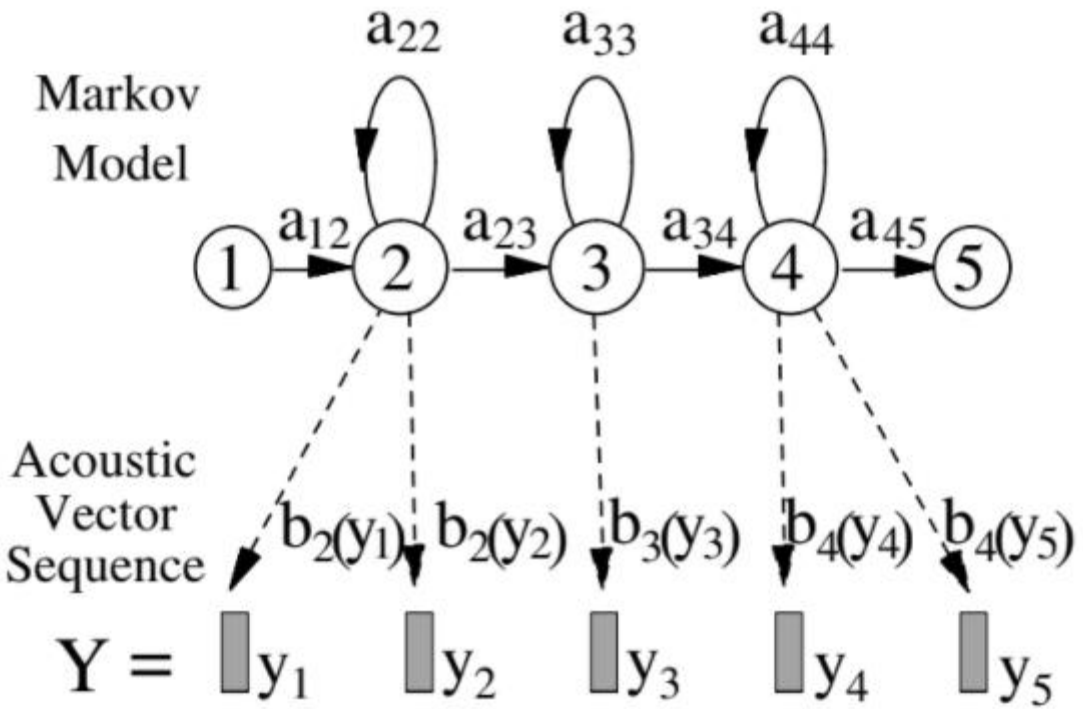
Language Model



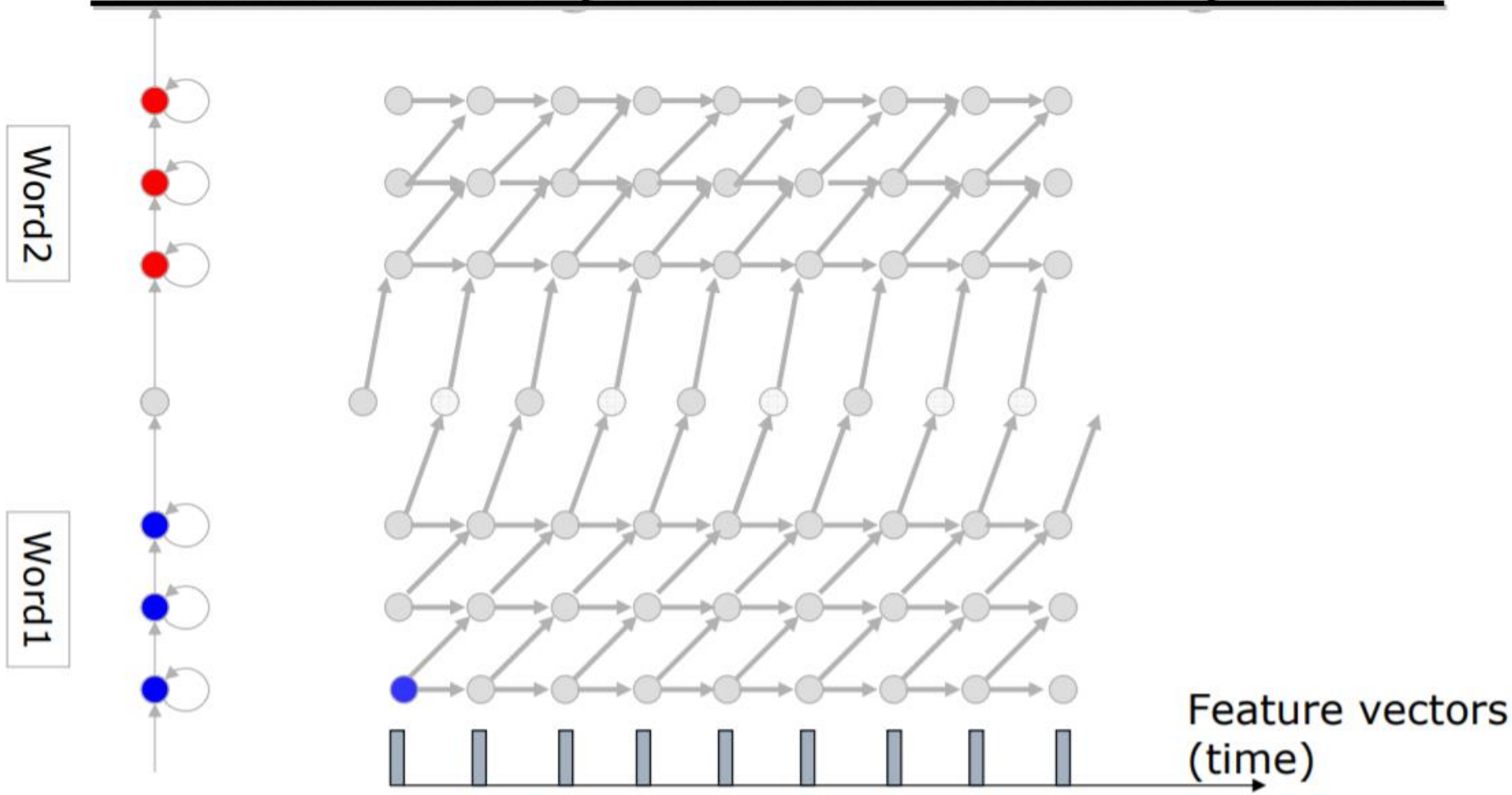
Architecture of HMM-based recognizer



HMM-based phone model

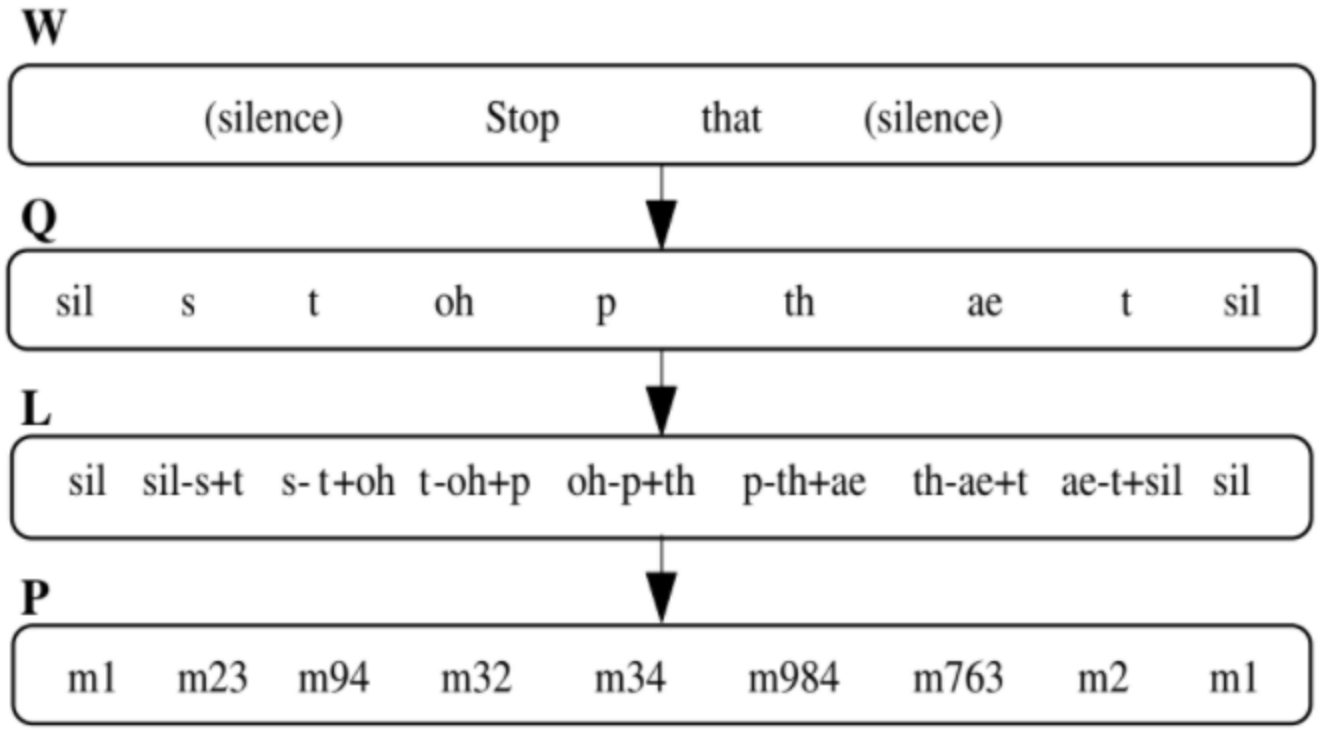


Viterbi through a Non-Emitting State

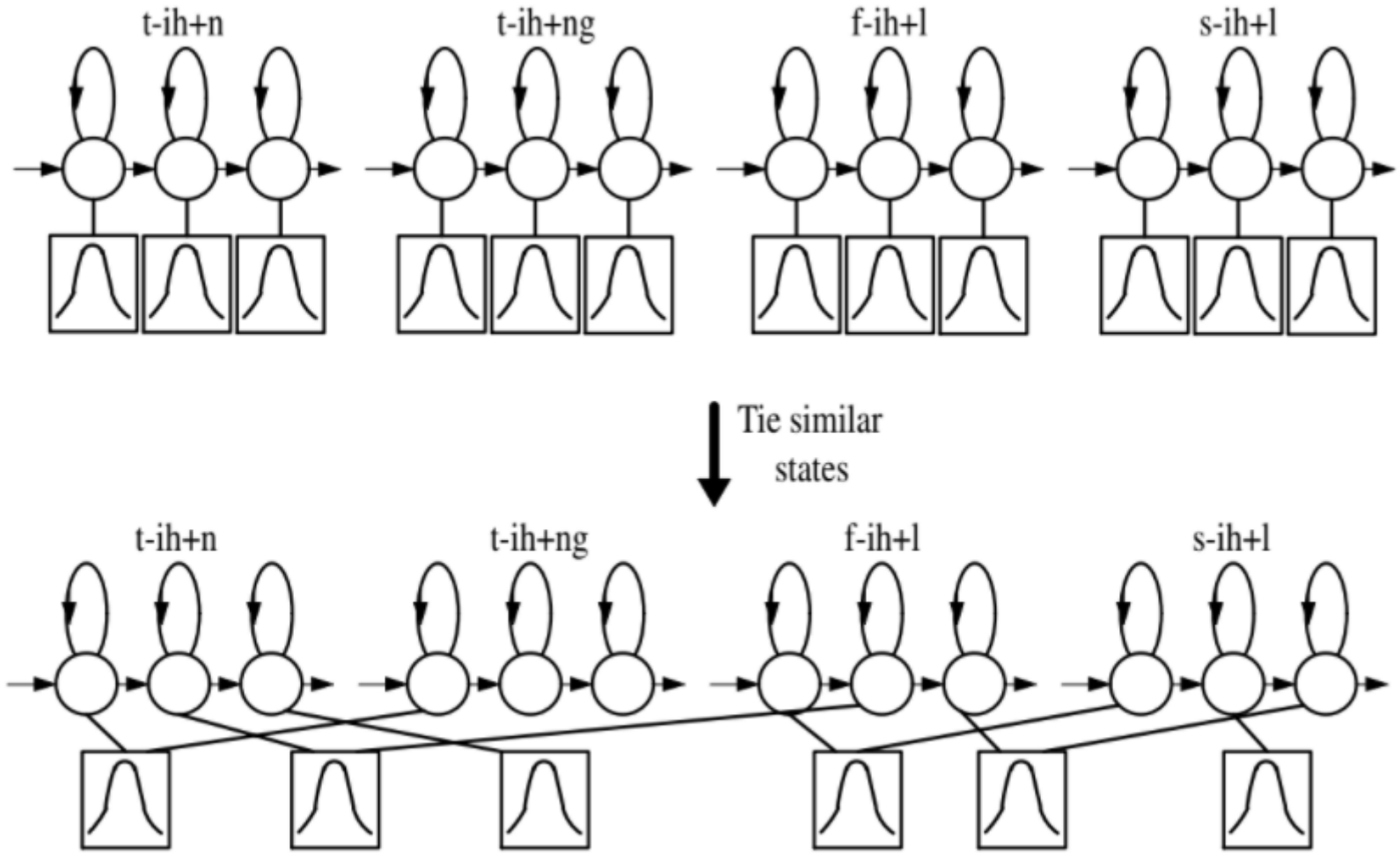


http://www.ee.iisc.ac.in/new/people/faculty/prasantg/downloads/Bhiksha_Tutorial_ConnectedWordModel.pdf

Context dependent phone modeling



Formation of the tied state phone model



Tree-Based State Tying for High Accuracy Acoustic Modelling

S.J. Young, J.J. Odell, P.C. Woodland

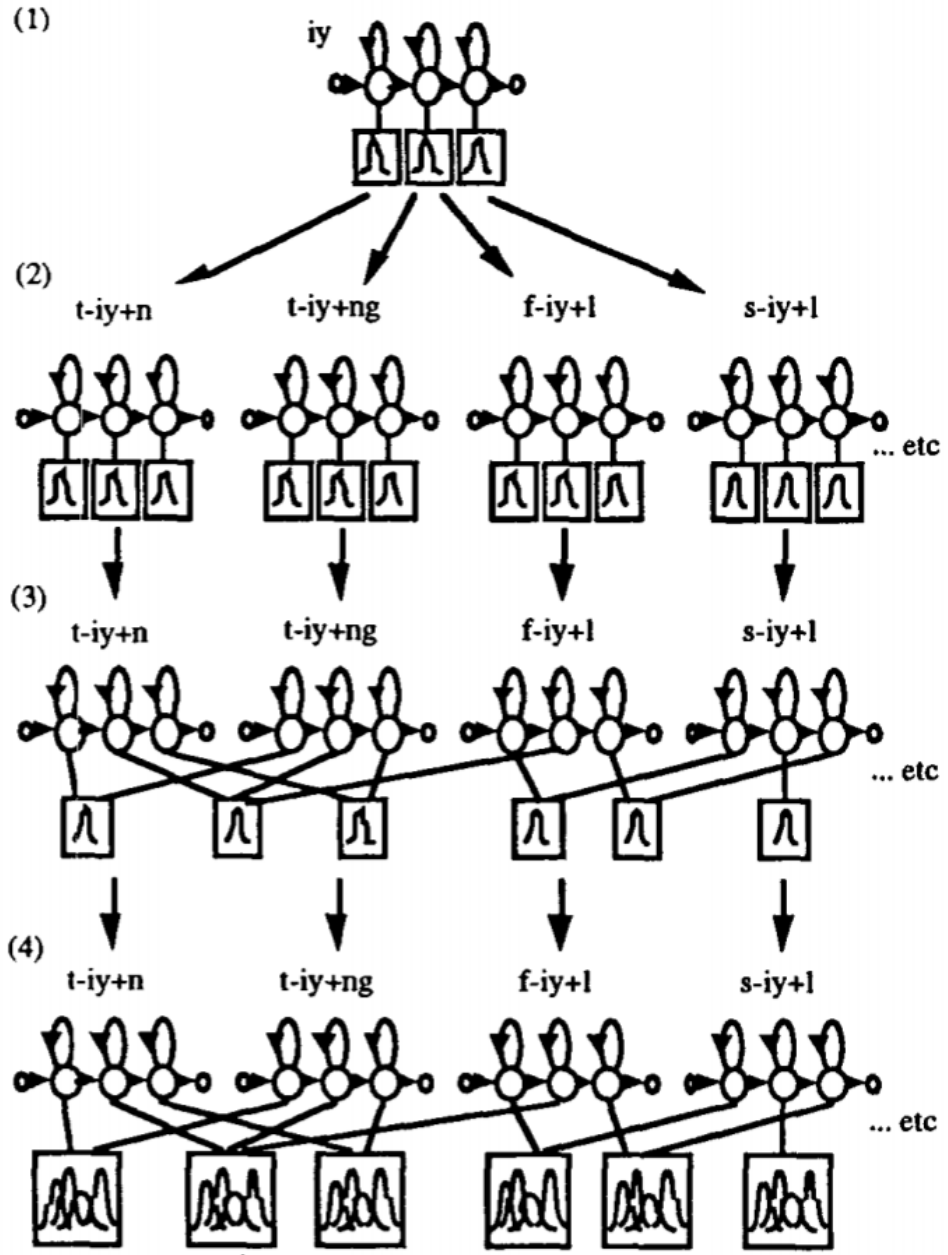
Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, England

ABSTRACT

The key problem to be faced when building a HMM-based continuous speech recogniser is maintaining the balance between model complexity and available training data. For large vocabulary systems requiring cross-word context dependent modelling, this is particularly acute since many such contexts will never occur in the training data. This paper describes a method of creating a tied-state continuous speech recognition system using a phonetic decision tree. This tree-

Management task when using the standard word pair grammar and 20,000 when no grammar is used. For the 20k Wall Street Journal task, around 55,000 triphones are needed. However, only 6600 triphones occur in the Resource Management training data and only 18,500 in the SI84 section of the Wall Street Journal training data.

Traditional methods of dealing with these problems involve sharing models across differing contexts to form



1. An initial set of a 3 state left-right monophone models with single Gaussian output probability density functions is created and trained.

2. The state output distributions of these monophones are then cloned to initialise a set of untied context dependent triphone models which are then trained using Baum-Welch re-estimation. The transition matrix is not cloned but remains tied across all the triphones of each phone.

3. For each set of triphones derived from the same monophone, corresponding states are clustered. In each resulting cluster, a typical state is chosen as exemplar and all cluster members are tied to this state.

4. The number of mixture components in each state is incremented and the models re-estimated until performance on a development test set peaks or the desired number of mixture components is reached.

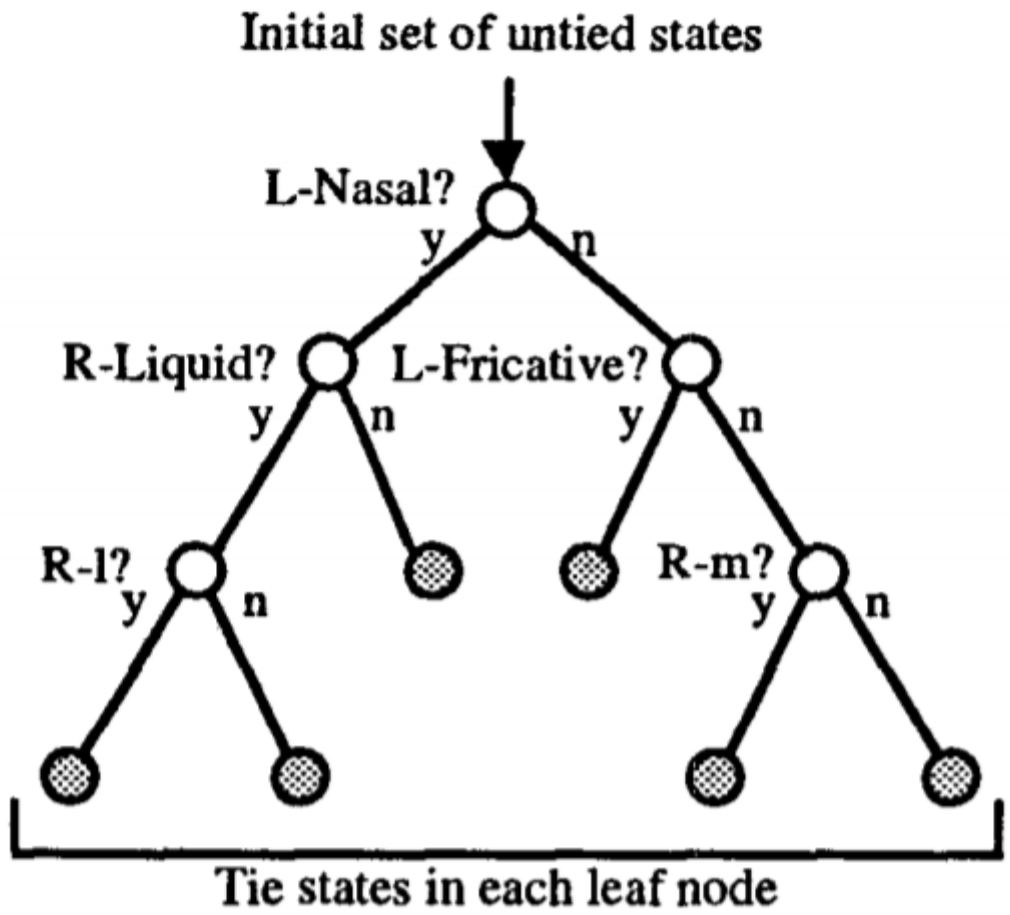


Figure 2: Example of a phonetic decision tree

TREE-BASED CLUSTERING

A phonetic decision tree is a binary tree in which a question is attached to each node. In the system described here, each of these questions relates to the phonetic context to the immediate left or right. For example, in Fig. 2, the question "Is the phone on the left of the current phone a nasal?" is associated with the root node of the tree. One tree is constructed for each state of each phone to cluster all of the corresponding states of all of the associated triphones. For example, the tree shown in Fig. 2 will partition its states into six subsets corresponding to the six terminal nodes. The states in each subset are tied to form a single state and the questions and the tree topology are chosen to maximise the likelihood of the training data given these tied states whilst ensuring that there is sufficient data associated with each tied state to estimate the parameters of a mixture Gaussian PDF.

Steps for tied-state context-dependent acoustic model

(1) A flat-start monophone set is created in which each base phone is a monophone single-Gaussian HMM with means and covariances equal to the mean and covariance of the training data.

(2) The parameters of the Gaussian monophones are re-estimated using 3 or 4 iterations of EM.

(3) Each single Gaussian monophone q is cloned once for each distinct triphone $x - q + y$ that appears in the training data.

(4) The resulting set of *training-data* triphones is again re-estimated using EM and the state occupation counts of the last iteration are saved.

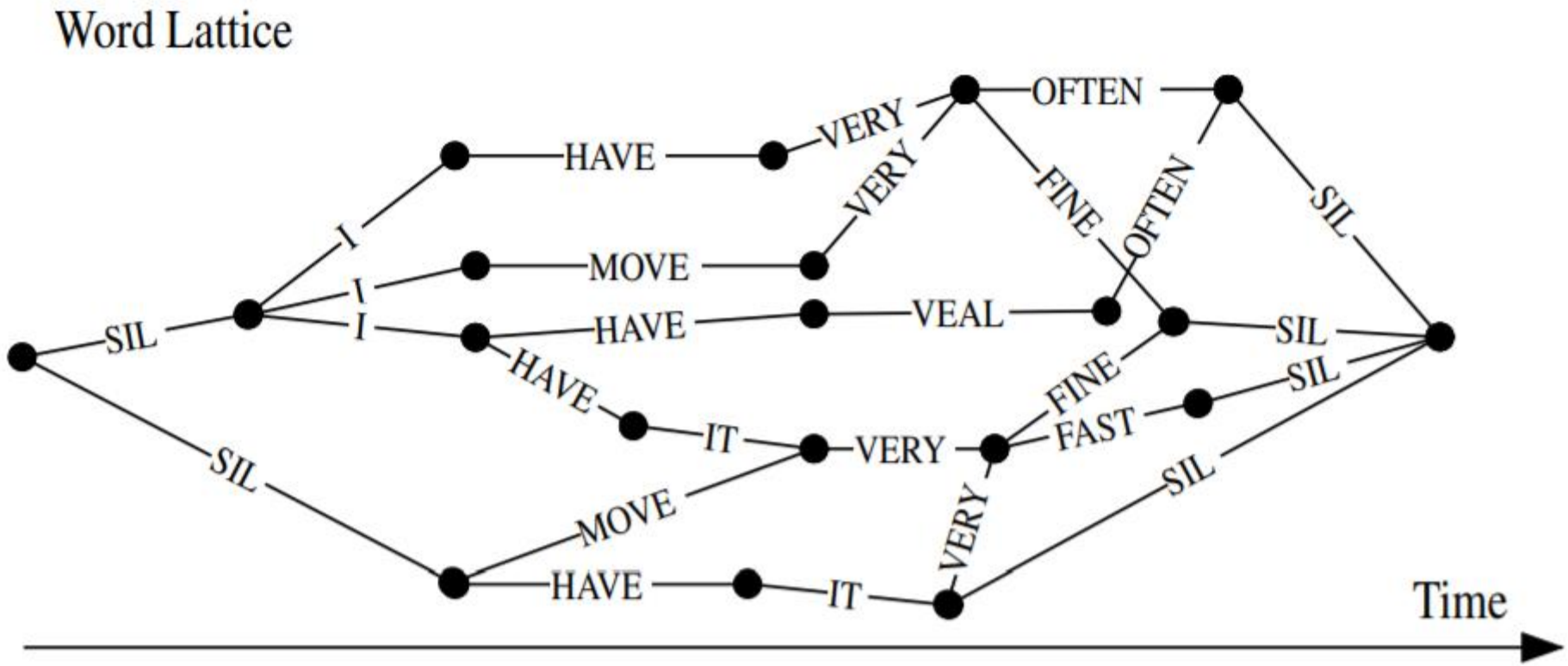
(5) A decision tree is created for each state in each base phone, the training-data triphones are mapped into a smaller set of tied-state triphones and iteratively re-estimated using EM.

Language Model

$$P(\boldsymbol{w}) = \prod_{k=1}^{K} P(w_k | w_{k-1}, \ldots, w_1).$$

$$P(\boldsymbol{w}) = \prod_{k=1}^{K} P(w_k | w_{k-1}, w_{k-2}, \ldots, w_{k-N+1}),$$

$$P(w_k | w_{k-1}, w_{k-2}) \approx \frac{C(w_{k-2} w_{k-1} w_k)}{C(w_{k-2} w_{k-1})}.$$


Word Lattice
SIL
I
I
I
HAVE
MOVE
HAVE
HAVE
IT
SIL
MOVE
HAVE
IT
VERY
VERY
VERY
OFTEN
FINE
OFTEN
SIL
VEAL
SIL
FINE
FAST
SIL
VERY
SIL
Time