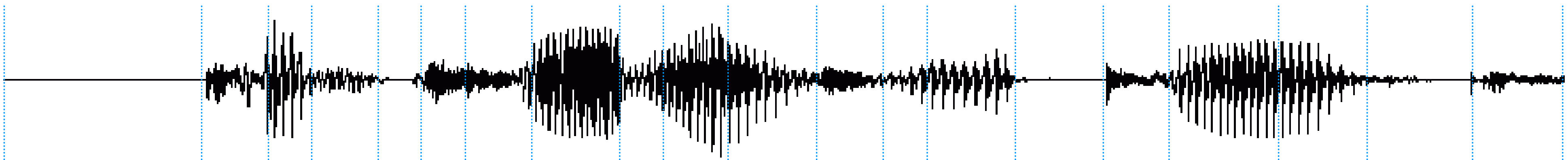
Lecture # 16
Session 2003

Segment-Based Speech Recognition

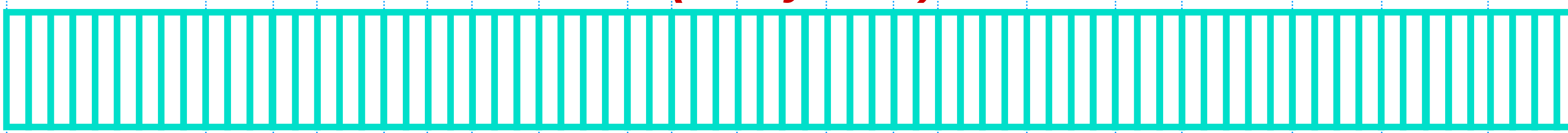
- **Introduction**
- **Searching graph-based observation spaces**
  - Anti-phone modelling
  - Near-miss modelling
- **Modelling landmarks**
- **Phonological modelling**

Segment-Based Speech Recognition

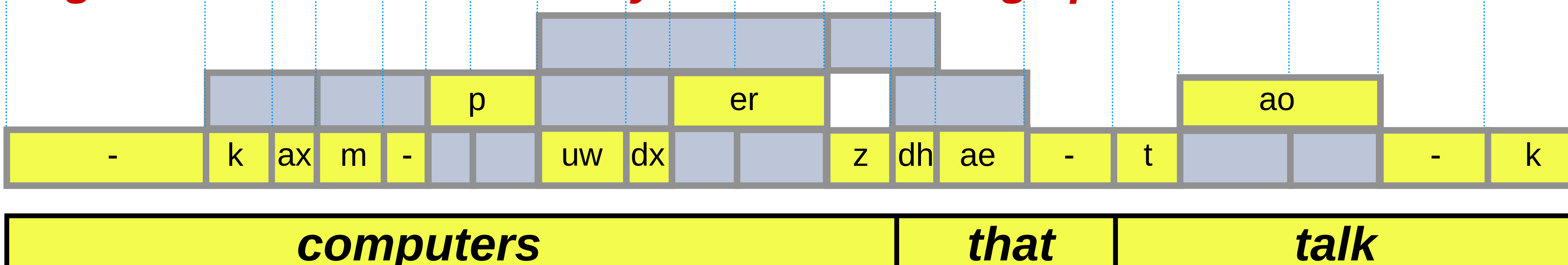
Waveform



Frame-based measurements (every 5ms)



Segment network created by interconnecting spectral landmarks



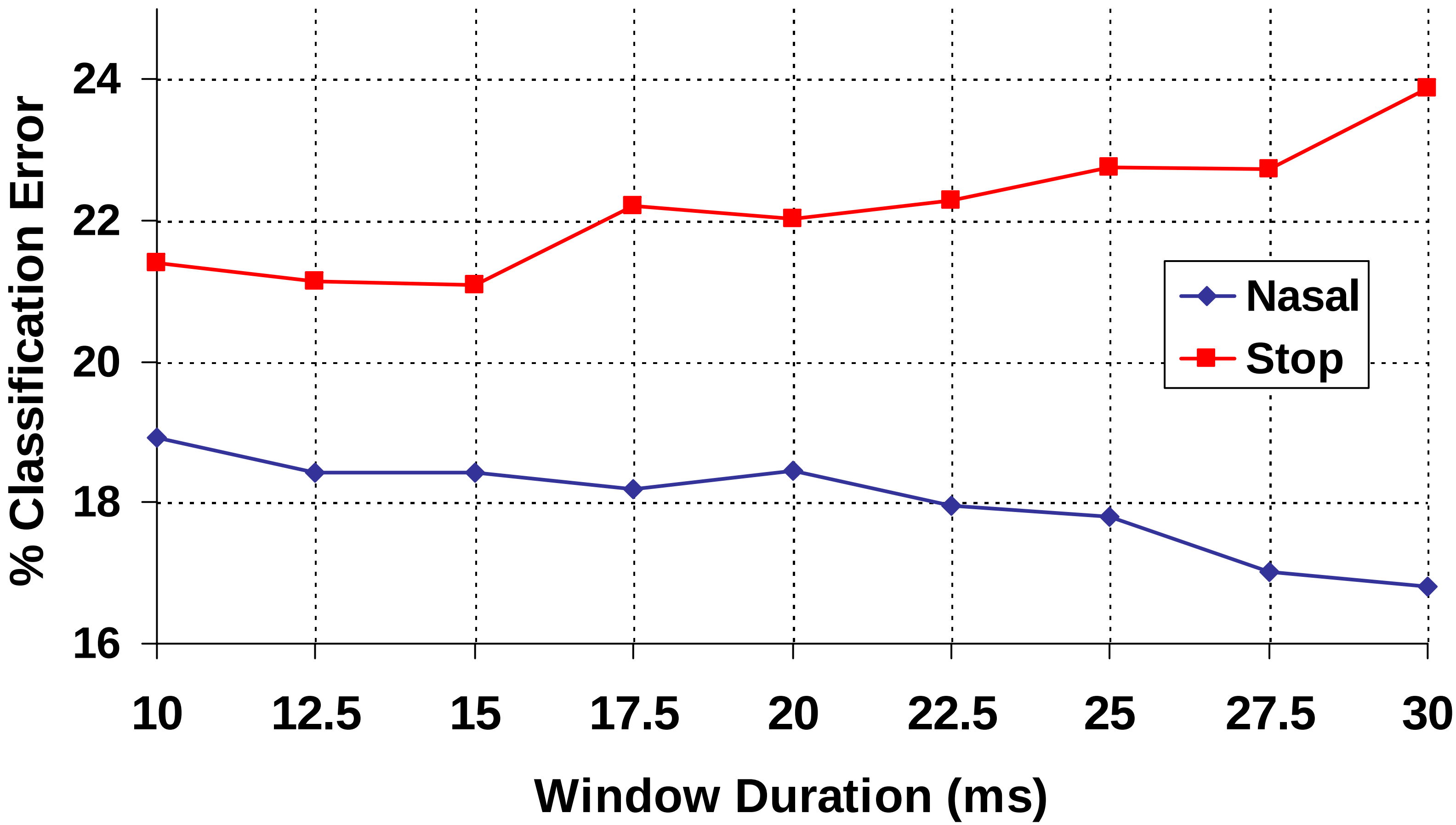
Probabilistic search finds most likely phone & word strings

Segment-based Speech Recognition

- **Acoustic modelling is performed over an entire segment**
- **Segments typically correspond to phonetic-like units**
- **Potential advantages:**
  - Improved joint modelling of time/spectral structure
  - Segment- or landmark-based acoustic measurements
- **Potential disadvantages:**
  - Significant increase in model and search computation
  - Difficulty in robustly training model parameters

Hierarchical Acoustic-Phonetic Modelling

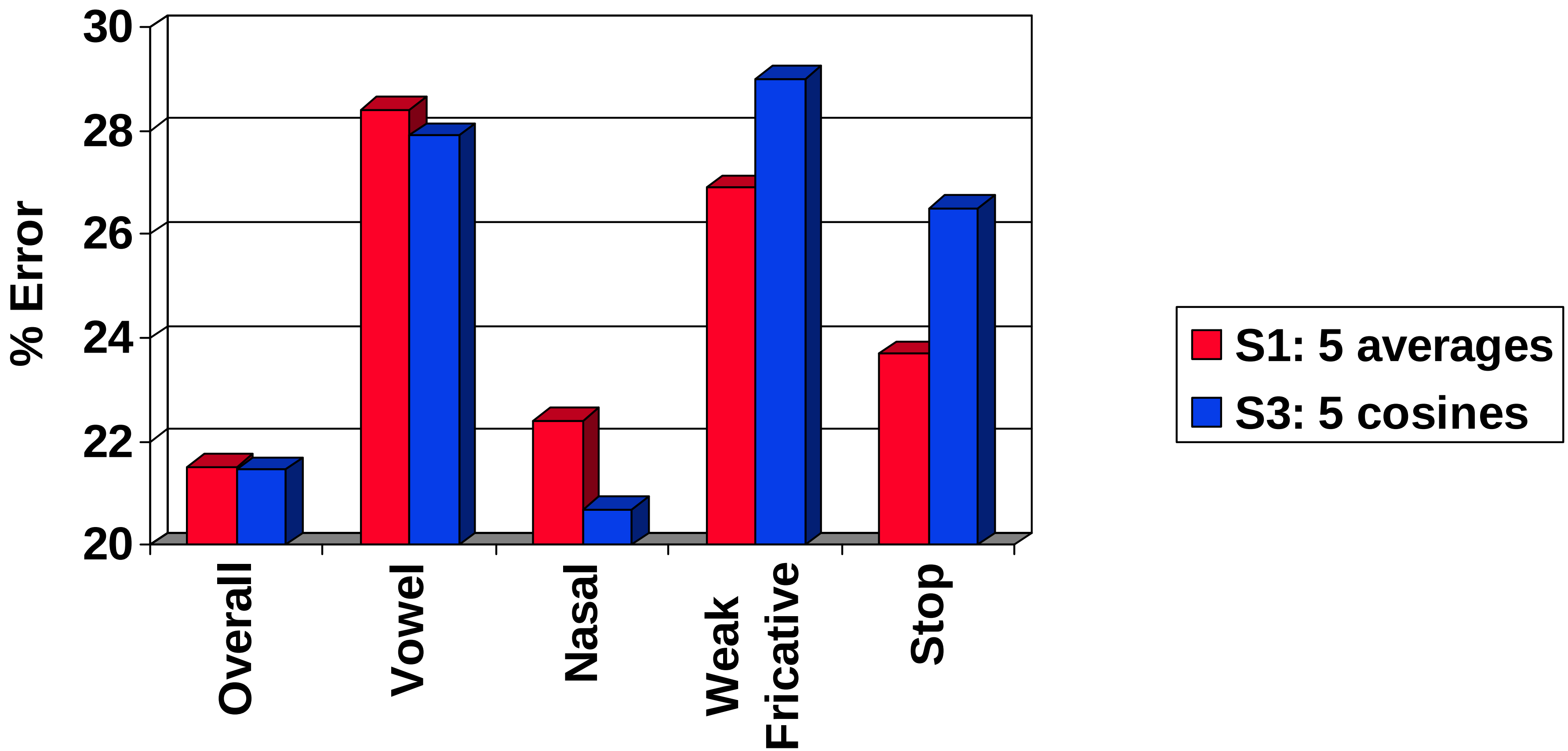
- **Homogeneous measurements can compromise performance**
  - Nasal consonants are classified better with a longer analysis window
  - Stop consonants are classified better with a shorter analysis window



- **Class-specific information extraction can reduce error**

Committee-based Phonetic Classification

- **Change of temporal basis affects within-class error**
  - Smoothly varying cosine basis better for vowels and nasals
  - Piecewise-constant basis better for fricatives and stops



- **Combining information sources can reduce error**

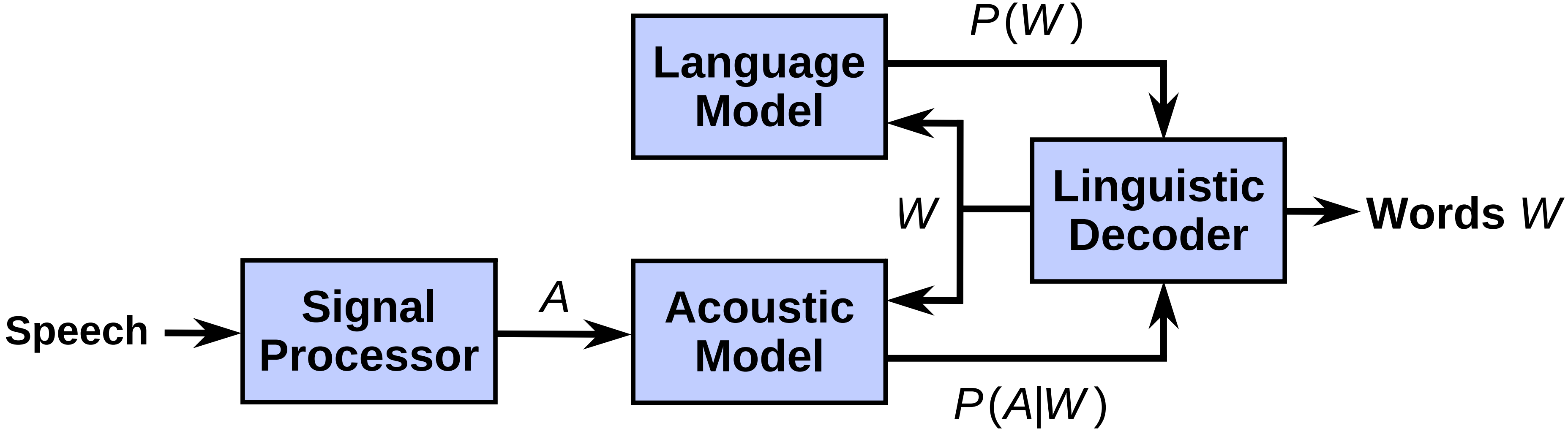
Phonetic Classification Experiments
(A. Halberstadt, 1998)

- **TIMIT acoustic-phonetic corpus**
  - Context-independent classification only
  - 462 speaker training corpus, 24 speaker core test set
  - Standard evaluation methodology, 39 common phonetic classes
- **Several different acoustic representations incorporated**
  - Various time-frequency resolutions (Hamming window 10-30 ms)
  - Different spectral representations (MFCCs, PLPCCs, etc)
  - Cosine transform vs. piecewise constant basis functions
- **Evaluated MAP hierarchy and committee-based methods**

Method	% Error
Baseline	21.6
MAP Hierarchy	21.0
Committee of 8 Classifiers	18.5*
Committee with Hierarchy	18.3

* Development set performance

Statistical Approach to ASR



- **Given acoustic observations, *A*, choose word sequence, *W**, which maximizes *a posteriori* probability, *P(W |A)***

$$W^* = \underset{W}{\operatorname{argmax}} P(W \mid A)$$

- **Bayes rule is typically used to decompose *P(W |A)* into acoustic and linguistic terms**

$$P(W \mid A) = \frac{P(A \mid W) P(W)}{P(A)}$$

ASR Search Considerations

- **A full search considers all possible segmentations, *S*, and units, *U*, for each hypothesized word sequence, *W***

$$W^* = \arg\max_W P(W \mid A) = \arg\max_W \sum_S \sum_U P(WUS \mid A)$$

- **Can seek best path to simplify search using dynamic programming (e.g., Viterbi) or graph-searches (e.g., *A**)**

$$W^*, U^*, S^* \approx \arg\max_{W,U,S} P(WUS \mid A)$$

- **The modified Bayes decomposition has four terms:**

$$P(WUS \mid A) = \frac{P(A \mid SUW)\, P(S \mid UW)\, P(U \mid W)\, P(W)}{P(A)}$$

In HMM's these correspond to acoustic, state, and language model probabilities or likelihoods

Examples of Segment-based Approaches

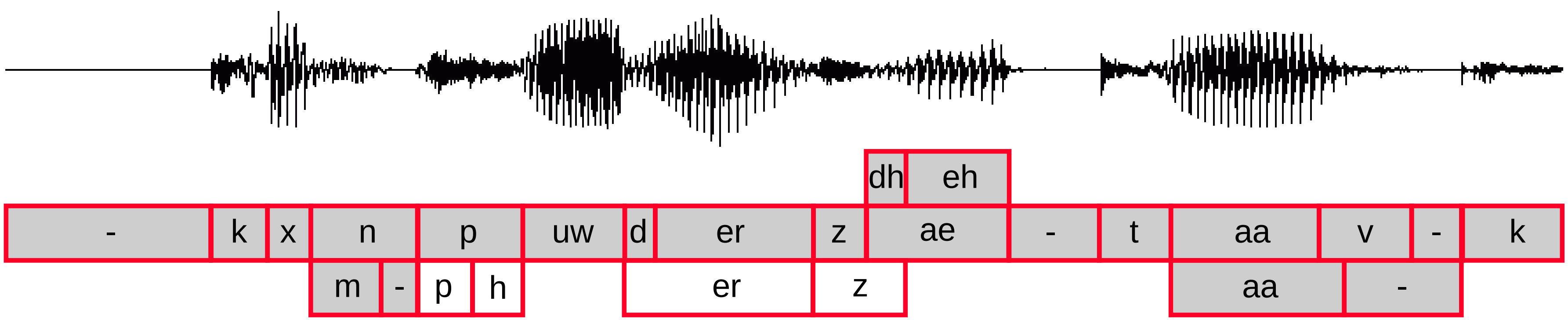
- **HMMs**
  - Variable frame-rate (Ponting et al., 1991, Alwan et al., 2000)
  - Segment-based HMM (Marcus, 1993)
  - Segmental HMM (Russell et al., 1993)
- **Trajectory Modelling**
  - Stochastic segment models (Ostendorf et al., 1989)
  - Parametric trajectory models (Ng, 1993)
  - Statistical trajectory models (Goldenthal, 1994)
- **Feature-based**
  - FEATURE (Cole et al., 1983)
  - SUMMIT (Zue et al., 1989)
  - LAFF (Stevens et al., 1992)

Segment-based Modelling at MIT

- **Baseline segment-based modelling incorporates:**
  - Averages and derivatives of spectral coefficients (e.g., MFCCs)
  - Dimensionality normalization via principal component analysis
  - PDF estimation via Gaussian mixtures
- **Example acoustic-phonetic modelling investigations, e.g.,**
  - Alternative probabilistic classifiers (e.g., Leung, Meng)
  - Automatically learned feature measurements (e.g., Phillips, Muzumdar)
  - Statistical trajectory models (Goldenthal)
  - Hierarchical probabilistic features (e.g., Chun, Halberstadt)
  - Near-miss modelling (Chang)
  - Probabilistic segmentation (Chang, Lee)
  - Committee-based classifiers (Halberstadt)

SUMMIT Segment-Based ASR

- **SUMMIT speech recognition is based on phonetic segments**
  - Explicit phone start and end times are hypothesized during search
  - Differs from conventional frame-based methods (e.g., HMMs)
  - Enables segment-based acoustic-phonetic modelling
  - Measurements can be extracted over landmarks and segments

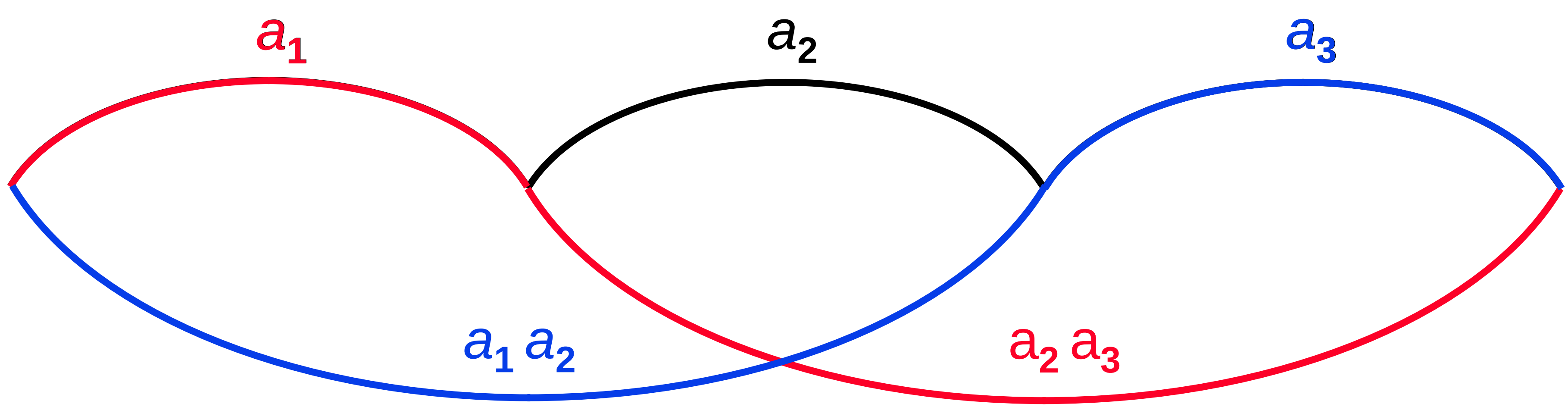


- **Recognition is achieved by searching a phonetic graph**
  - Graph can be computed via acoustic criterion or probabilistic models
  - Competing segmentations make use of different observation spaces
  - Probabilistic decoding must account for graph-based observation space

"Frame-based" Speech Recognition

- **Observation space, *A*, corresponds to a temporal sequence of acoustic frames (e.g., spectral slices)**

$$A = \{a_1 a_2 a_3\}$$



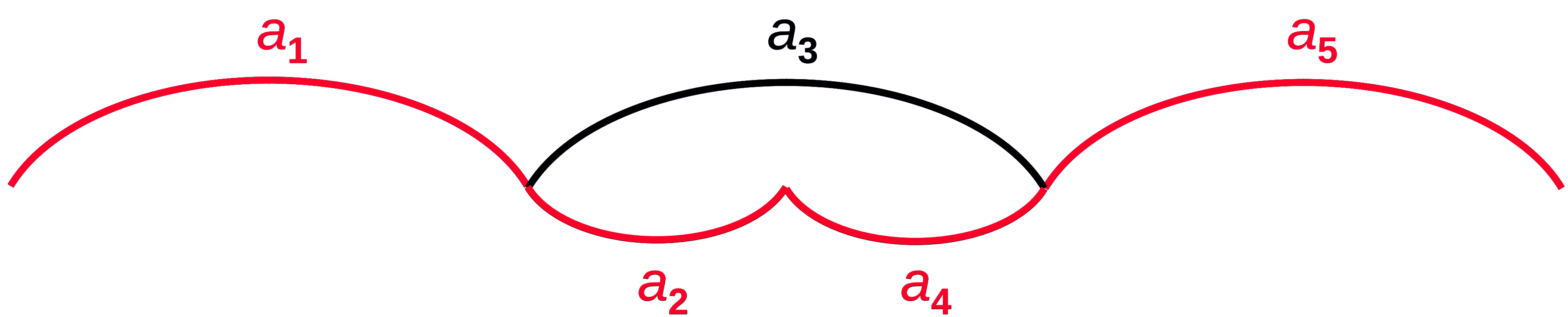
- **Each hypothesized segment, s_i, is represented by the series of frames computed between segment start and end times**
- **The acoustic likelihood, $P(A|SW)$, is derived from the same observation space for all word hypotheses**

$$P(a_1 a_2 a_3 |SW) \Leftrightarrow P(a_1 a_2 a_3 |SW) \Leftrightarrow P(a_1 a_2 a_3 |SW)$$

“Feature-based” Speech Recognition

- **Each segment, s_i, is represented by a single feature vector, a_i**

$$A = \{a_1 a_2 a_3 a_4 a_5\}$$



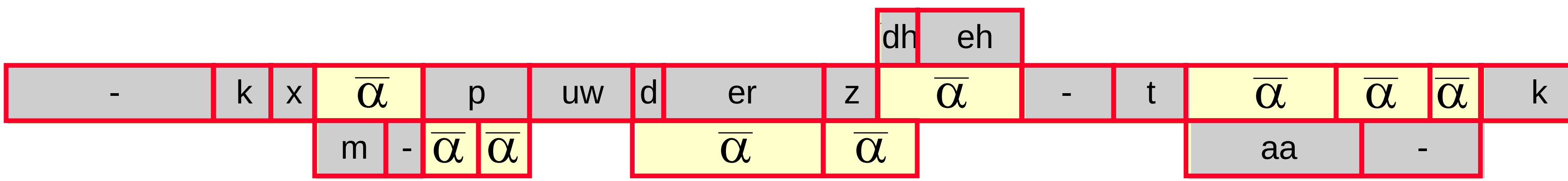
$X = \{a_1 a_3 a_5\}$ $Y = \{a_2 a_4\}$ $X = \{a_1 a_2 a_4 a_5\}$ $Y = \{a_3\}$

- **Given a particular segmentation, S, A consists of X, the feature vectors associated with S, as well as Y, the feature vectors associated with segments not in S:** $A = X \cup Y$
- **To compare different segmentations it is necessary to predict the likelihood of both X and Y:** $P(A|SW) = P(XY|SW)$

$$P(a_1 a_3 a_5\, a_2 a_4 \,|SW) \Leftrightarrow P(a_1 a_2 a_4 a_5\, a_3 \,|SW)$$

Searching Graph-Based Observation Spaces: The Anti-Phone Model

- **Create a unit, $\bar{\alpha}$, to model segments that are not phones**
- **For a segmentation, *S*, assign anti-phone to extra segments**
  - All segments are accounted for in the phonetic graph
  - Alternative paths through the graph can be legitimately compared



- **Path likelihoods can be decomposed into two terms:**
  1. The likelihood of all segments produced by the anti-phone (a constant)
  2. The ratio of phone to anti-phone likelihoods for all path segments
- **MAP formulation for most likely word sequence, *W*, given by:**

$$W^* = \operatorname*{argmax}_{W,S} \prod_{i}^{N_S} \frac{P(x_i \mid u_i)}{P(x_i \mid \bar{\alpha})} P(s_i \mid u_i) P(U \mid W) P(W)$$

Modelling Non-lexical Units: The Anti-phone

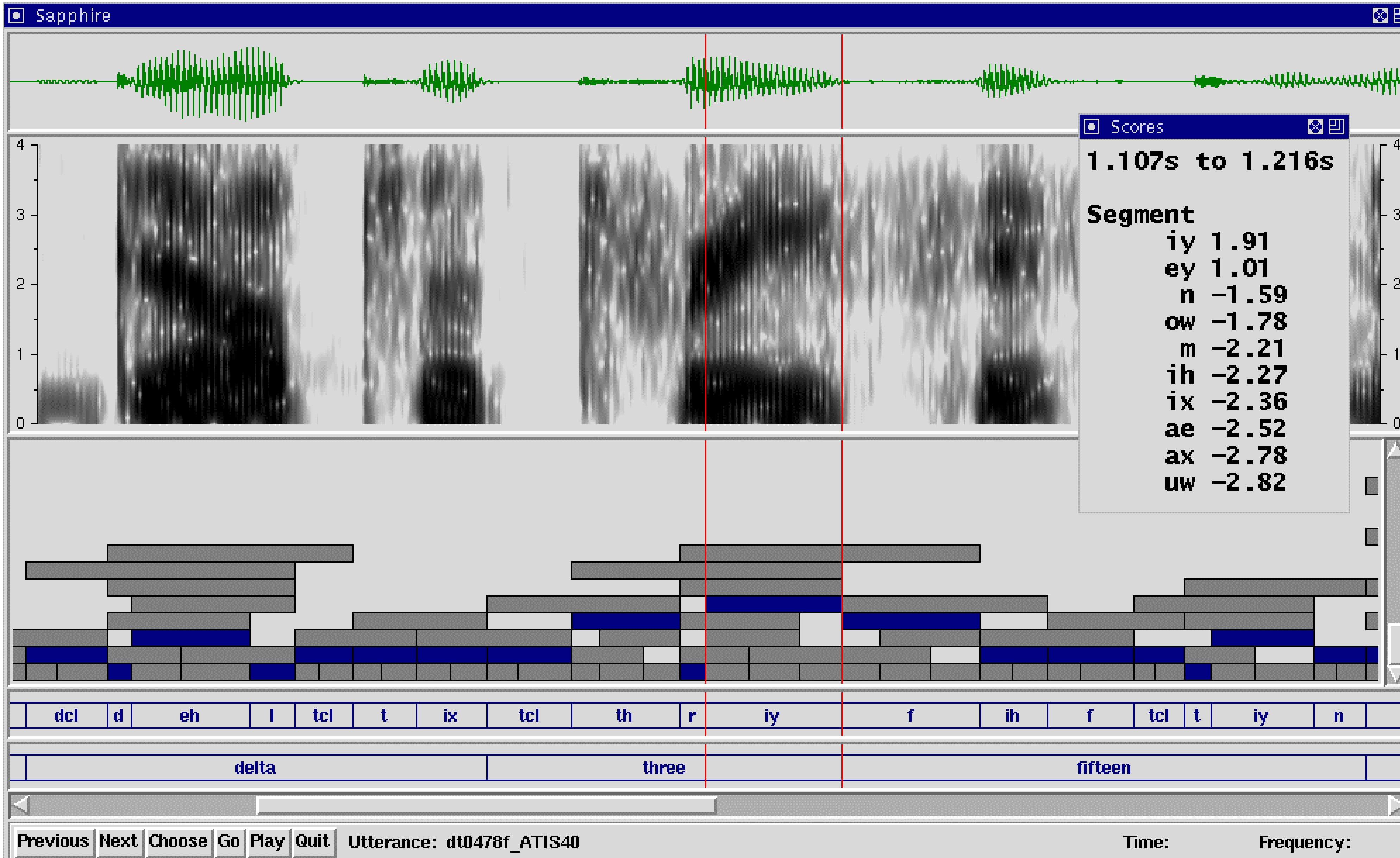
- **Given a particular segmentation,** S**,** A **consists of** X**, the segments associated with** S**, as well as** Y**, the segments not associated with** S**:** $P(A|SU)=P(XY|SU)$
- **Given segmentation** S**, assign feature vectors in** X **to valid units, and all others in** Y **to the anti-phone**
- **Since** $P(XY \mid \bar{\alpha})$ **is a constant,** K**, we can write** $P(XY|SU)$ **assuming independence between** X **and** Y

$$P(XY \mid SU) = P(XY \mid U) = P(X \mid U)P(Y \mid \bar{\alpha})\frac{P(X \mid \bar{\alpha})}{P(X \mid \bar{\alpha})} = K\frac{P(X \mid U)}{P(X \mid \bar{\alpha})}$$

- **We need consider only segments in** S **during search:**

$$W^* = \arg\max_{W,U,S} \prod_i^{N_S} \frac{P(x_i \mid U)}{P(x_i \mid \bar{\alpha})} P(s_i \mid u_i) P(U \mid W) P(W)$$

SUMMIT Segment-based ASR

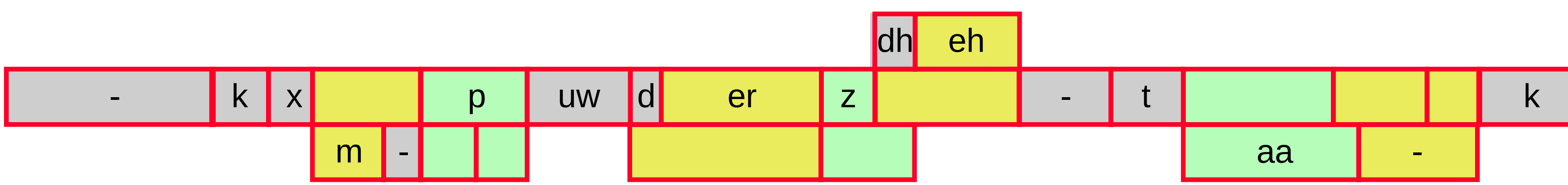


Anti-Phone Framework Properties

- **Models entire observation space, using both positive and negative examples**
- **Log likelihood scores are normalized by the anti-phone**
  - Good scores are positive, bad scores are negative
  - Poor segments all have negative scores
  - Useful for pruning and/or rejection
  - Anti-phone is not used for lexical access
- **No prior or posterior probabilities used during search**
  - Allows computation on demand and/or fastmatch
  - Subsets of data can be used for training
- **Context-independent or -dependent models can be used**
- **Useful for general pattern matching problems with graph-based observation spaces**

Beyond Anti-Phones: Near-Miss Modelling

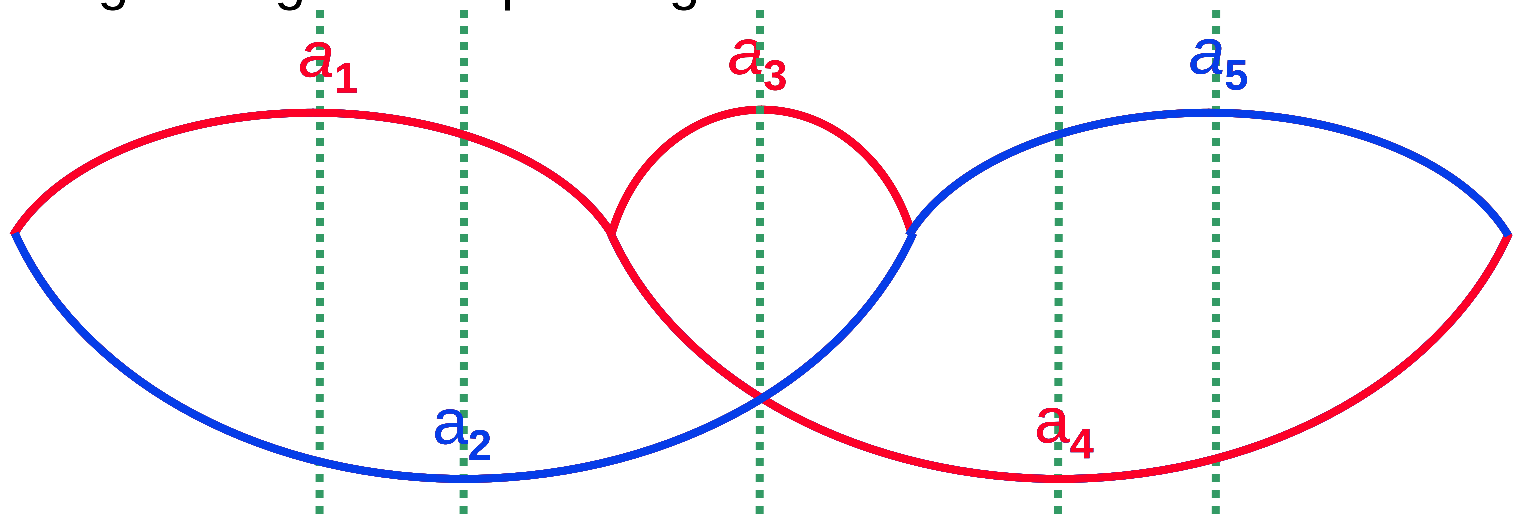
- **Anti-phone modelling partitions the observation space into two parts (i.e., on or not on a hypothesized segmentation)**
- **Near-miss modelling partitions the observation space into a set of mutually exclusive, collectively exhaustive subsets**
  - One near-miss subset pre-computed for each segment in a graph
  - Temporal criterion can guarantee proper near-miss subset generation (e.g., segment A is a near-miss of B iff A's mid-point is spanned by B)



- **During recognition, observations in a near-miss subset are mapped to the near-miss model of the hypothesized phone**
- **Near-miss models can be just an anti-phone, but can potentially be more sophisticated (e.g., phone dependent)**

Creating Near-miss Subsets

- **Near-miss subsets, A_i, associated with any segmentation, S, must be mutually exclusive, and exhaustive: $A = \cup A_i \; \forall A_i \in S$**
- **Temporal criterion guarantees proper near-miss subsets**
  - Abutting segments in *S* account for all times exactly once
  - Finding all segments spanning a time creates near-miss subsets



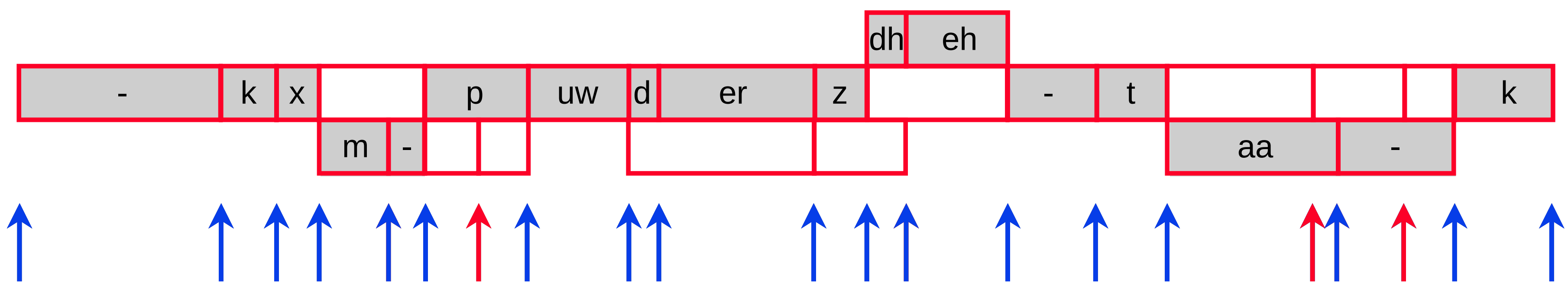
$a_1 \in A_1, A_2$ $a_2 \in A_1, A_2$ $a_3 \in A_2, A_3, A_4$ $a_4 \in A_4, A_5$ $a_5 \in A_4, A_5$

$A_1 = \{a_1 a_2\}$ $A_2 = \{a_1 a_2 a_3\}$ $A_3 = \{a_3\}$ $A_4 = \{a_3 a_4 a_5\}$ $A_5 = \{a_4 a_5\}$

$A = \cup A_i \; \forall S \quad S = \{\{a_1 a_3 a_5\}, \{a_1 a_4\}, \{a_2 a_5\}\}$

Modelling Landmarks

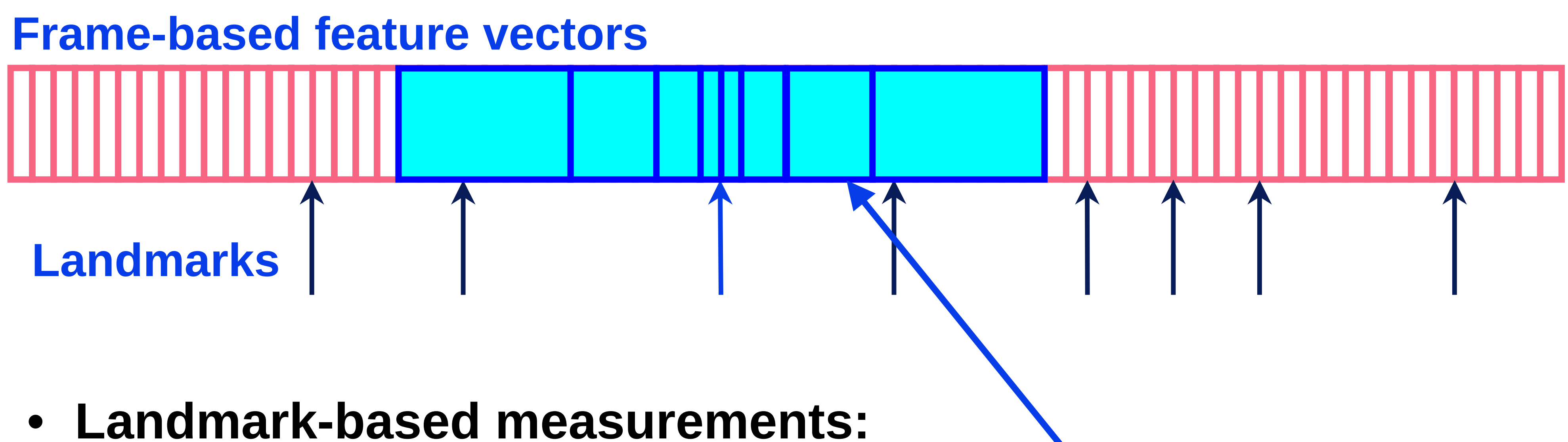
- **We can also incorporate additional feature vectors computed at hypothesized landmarks or phone boundaries**



- **Every segmentation accounts for every landmark**
  – Some landmarks will be transitions between lexical-units
  – Other landmarks will be considered internal to a unit
- **Both context-independent or dependent units are possible**
- **Effectively model transitions between phones (i.e., diphones)**
- **Frame-based models can be used to generate segment graph**

Modelling Landmarks

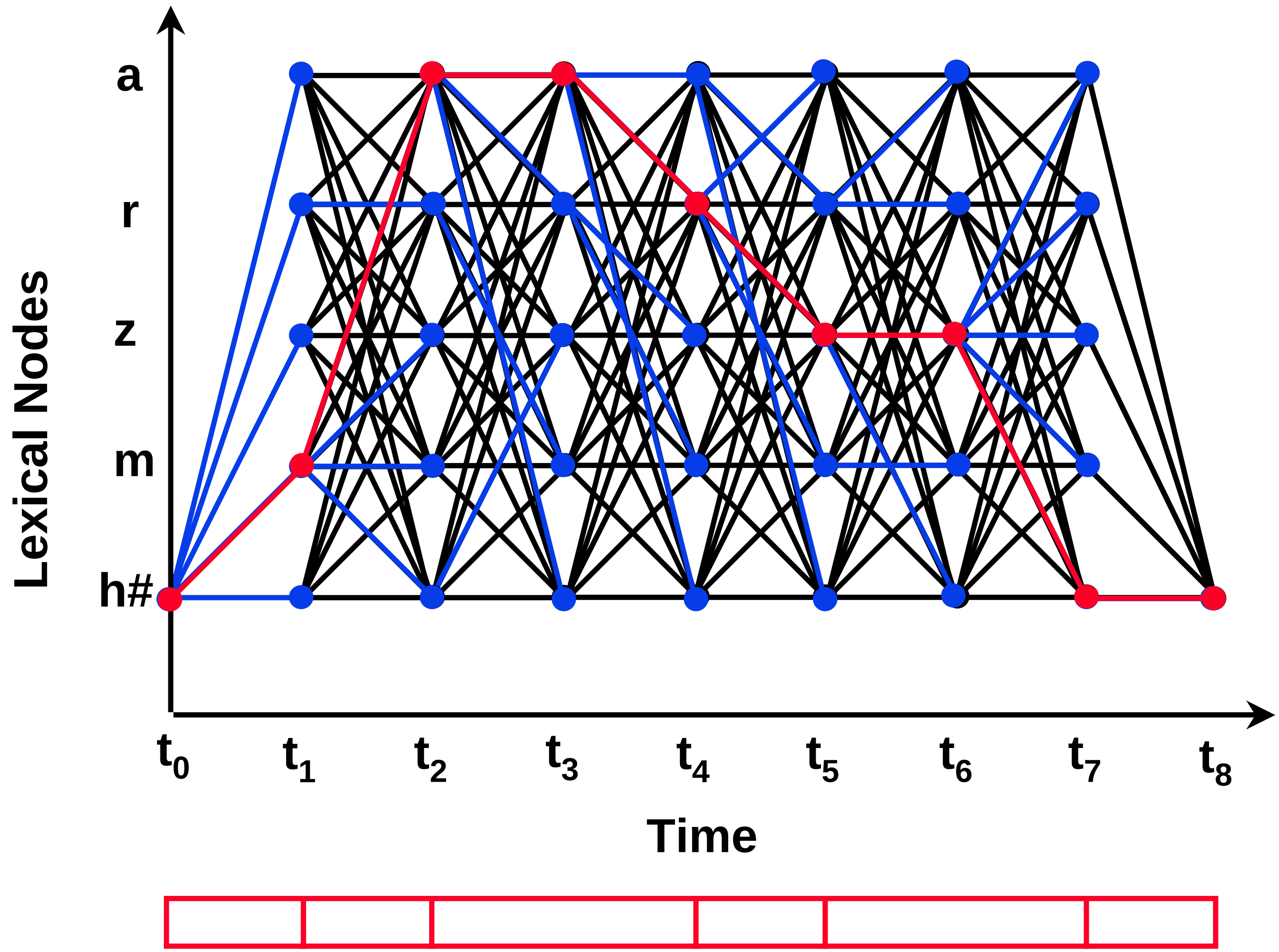
- **Frame-based measurements:**
  - Computed every 5 milliseconds
  - Feature vector of 14 Mel-Scale Cepstral Coefficients (MFCCs)



- **Landmark-based measurements:**
  - Compute average of MFCCs over 8 regions around landmark
  - 8 regions X 14 MFCC averages = 112 dimension vector
  - 112 dims. reduced to 50 using principal component analysis

Probabilistic Segmentation

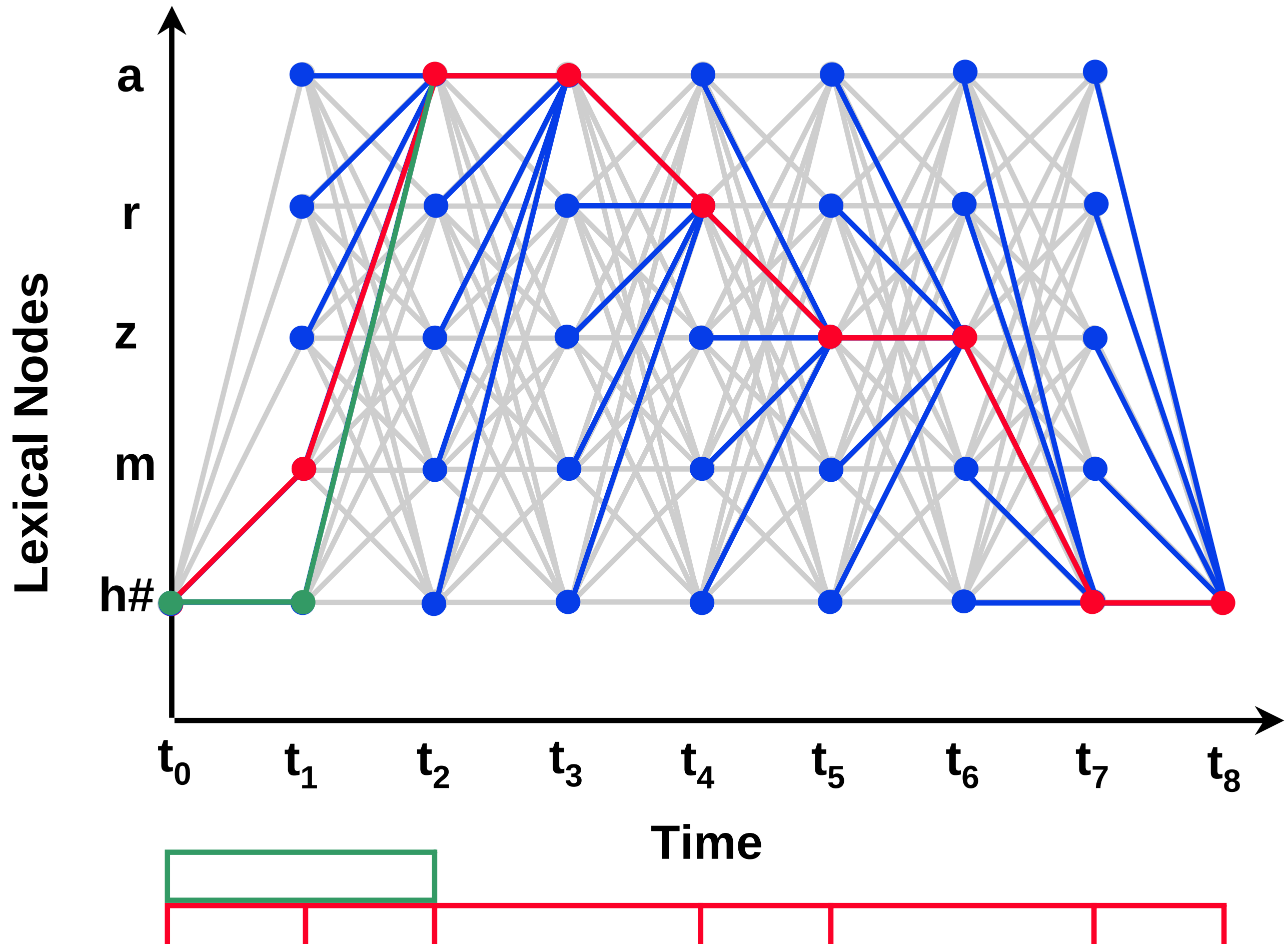
- Uses forward Viterbi search in first-pass to find best path



- Relative and absolute thresholds used to speed-up search

Probabilistic Segmentation (con't)

- Second pass uses backwards *A** search to find *N*-best paths
- Viterbi backtrace is used as future estimate for path scores



- Block processing enables pipelined computation

Phonetic Recognition Experiments

- **TIMIT acoustic-phonetic corpus**
  - 462 speaker training corpus, 24 speaker core test set
  - Standard evaluation methodology, 39 common phonetic classes
- **Segment and landmark representations based on averages and derivatives of 14 MFCCs, energy and duration**
- **PCA used for data normalization and reduction**
- **Acoustic models based on aggregated Gaussian mixtures**
- **Language model based on phone bigram**
- **Probabilistic segmentation computed from diphone models**

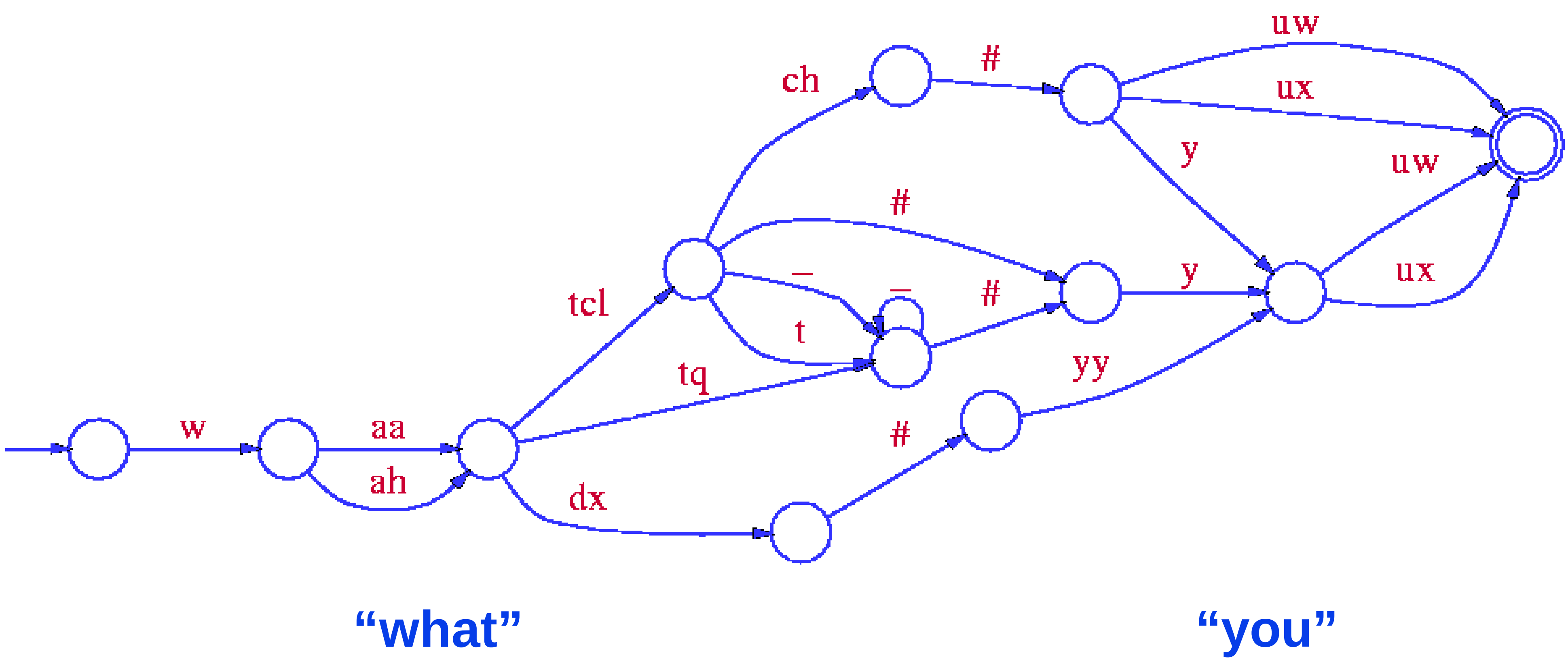
Method	% Error
Triphone CDHMM	27.1
Recurrent Neural Network	26.1
Bayesian Triphone HMM	25.6
Anti-phone, Heterogeneous classifiers	24.4

Phonological Modelling

- **Words described by phonemic baseforms**
- **Phonological rules expand baseforms into graph, e.g.,**
  - Deletion of stop bursts in syllable coda (e.g., *laptop*)
  - Deletion of /t/ in various environments (e.g., *intersection*, *destination, crafts)*
  - Gemination of fricatives and nasals (e.g., *this side, in nome*)
  - Place assimilation (e.g., *did you* (/d ih jh uw/))
- **Arc probabilities,** *P(U|W),* **can be trained**
- **Most HMMs do not have a phonological component**

Phonological Example

- **Example of "what you" expanded in SUMMIT recognizer**
  - Final /t/ in "what" can be realized as released, unreleased, palatalized, or glottal stop, or flap



Word Recognition Experiments

- **Jupiter telephone-based, weather-queries corpus**
  - 50,000 utterance training set, 1806 “in-domain” utterance test set
- **Acoustic models based on Gaussian mixtures**
  - Segment and landmark representations based on averages and derivatives of 14 MFCCs, energy and duration
  - PCA used for data normalization and reduction
  - 715 context-dependent boundary classes
  - 935 triphone, 1160 diphone context-dependent segment classes
- **Pronunciation graph incorporates pronunciation probabilities**
- **Language model based on class bigram and trigram**
- **Best performance achieved by combining models**

Method	% Error
Boundary models	7.6
Segment models	9.6
Combined	6.1

Summary

- **Some segment-based speech recognition techniques transform the observation space from frames to graphs**
- **Graph-based observation spaces allow for a wide-variety of alternative modelling methods to frame-based approaches**
- **Anti-phone and near-miss modelling frameworks provide a mechanism for searching graph-based observation spaces**
- **Good results have been achieved for phonetic recognition**
- **Much work remains to be done!**

References

- J. Glass, “A Probabilistic Framework for Segment-Based Speech Recognition,” to appear in *Computer, Speech & Language*, 2003.
- D. Halberstadt, “Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition,” Ph.D. Thesis, MIT, 1998.
- M. Ostendorf, et al., “From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition,” *Trans. Speech & Audio Proc.,* 4(5), 1996.