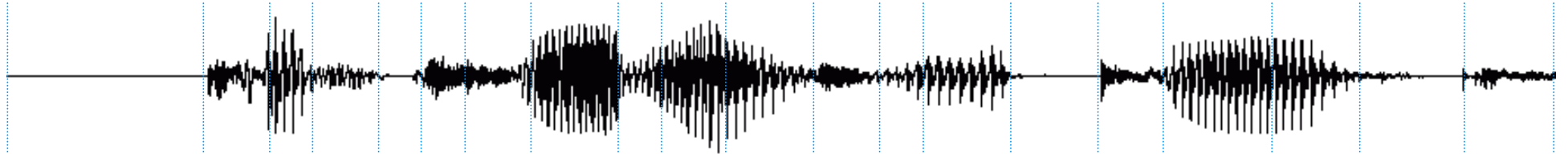
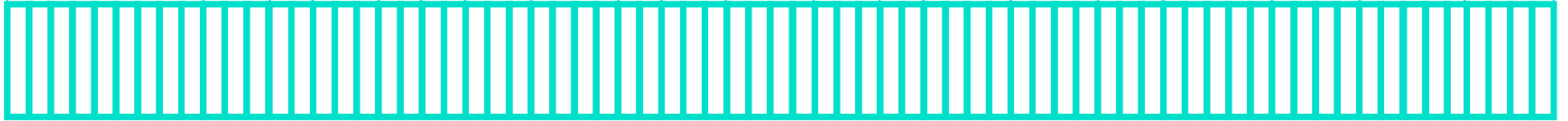
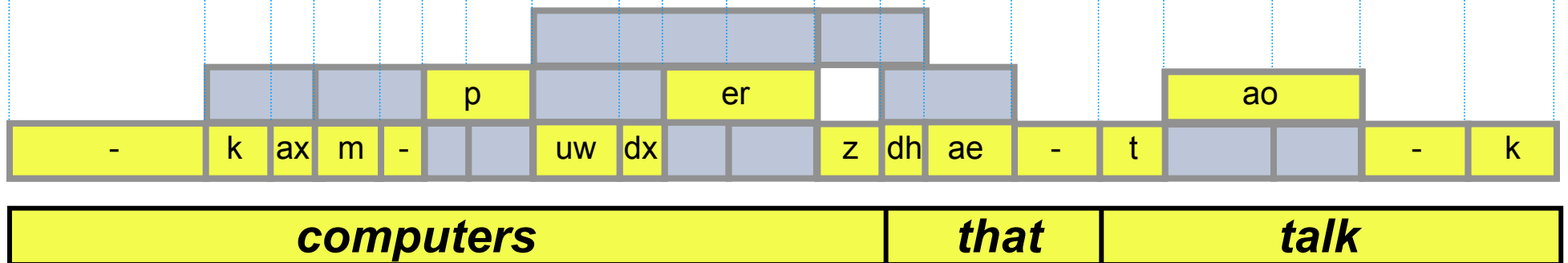


Reconocimiento del habla basado en segmentos

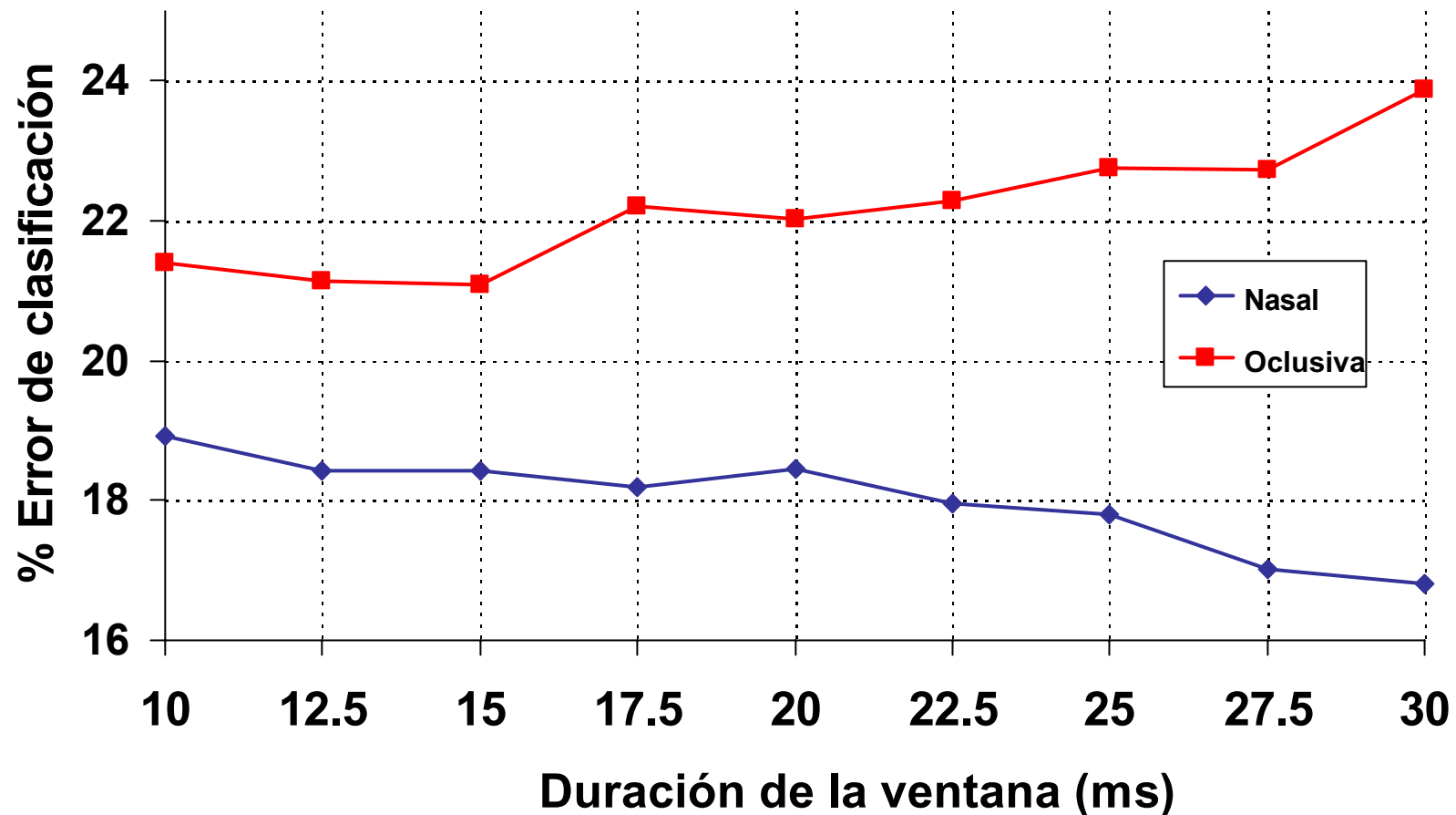
- **Introducción**
- **Búsqueda de espacios de observación basados en grafos**
 - Modelado antifono
 - Modelado próximo al error
- **Modelado de puntos de referencia**
- **Modelado fonológico**

Forma de onda**Mediciones basadas en tramos (cada 5ms)****Red de segmentos creada por la interconexión de límites espectrales**

La búsqueda probabilística encuentra fonos y cadenas de palabras con más posibilidades

- **El modelado acústico se realiza sobre un segmento absoluto**
- **Los segmentos corresponden típicamente a unidades como las fonéticas**
- **Ventajas potenciales:**
 - Modelado de unión de la estructura espectral/temporal mejorado
 - Mediciones acústicas basadas en segmentos o en puntos de referencia
- **Inconvenientes potenciales:**
 - Aumento significativo en la computación de la búsqueda y del modelo
 - Dificultad para conseguir un entrenamiento robusto de los parámetros del modelo

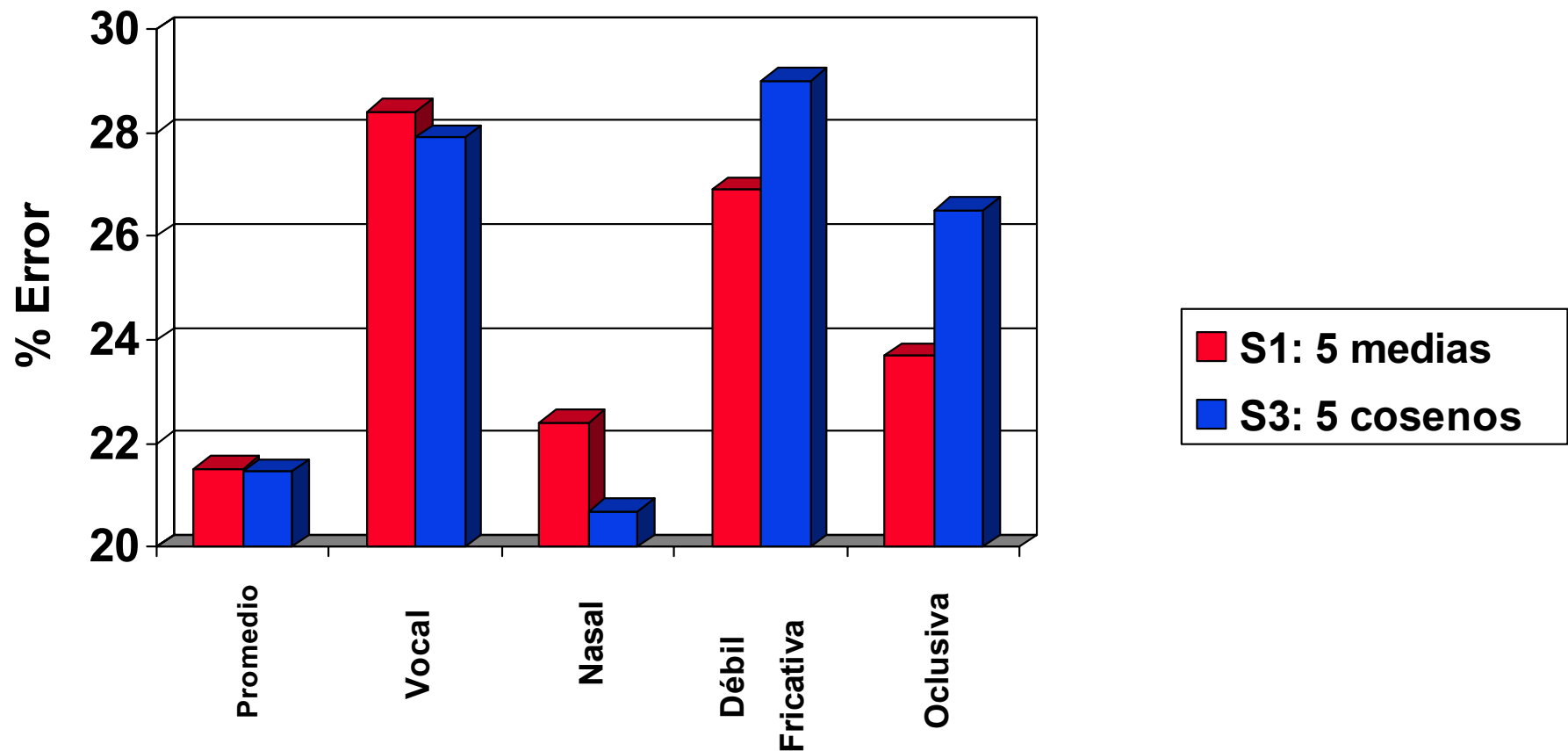
- **Las mediciones homogéneas pueden poner en peligro el rendimiento**
 - Las consonantes nasales son mejor clasificadas con una ventana de análisis más larga
 - Las consonantes oclusivas son mejor clasificadas con una ventana de análisis más corta



- **La extracción de la información propia de la clase puede reducir el error**

Clasificación fonética basada en el comité

- Las modificaciones a nivel temporal afectan a los errores internos a la clase
 - Base coseno fácilmente variable mejor para vocales y nasales
 - Base constante a intervalos mejor para fricativas y oclusivas



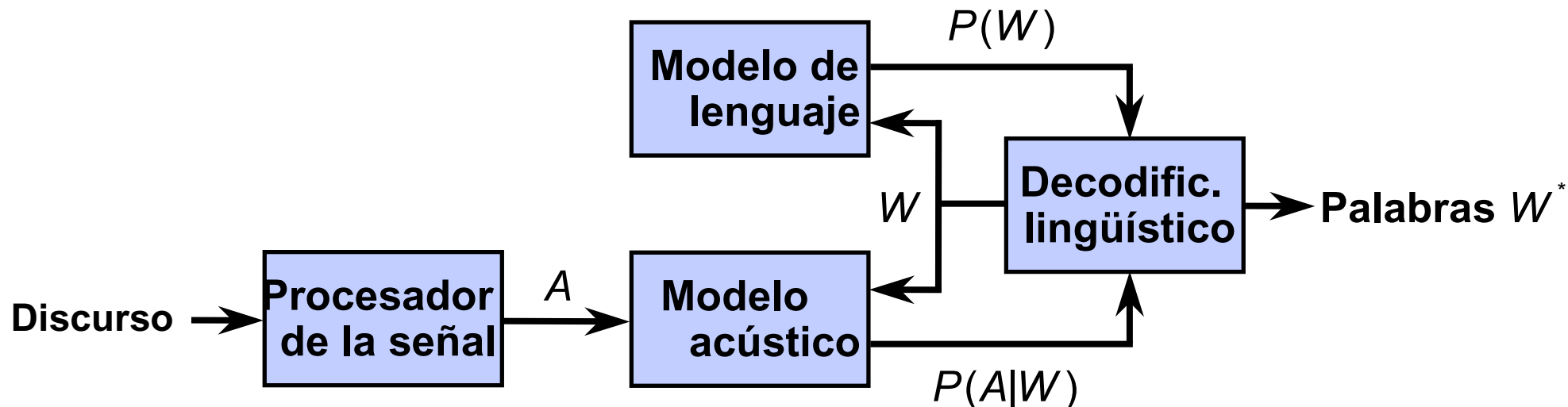
- La combinación de fuentes de información puede reducir el error

Experimentos de clasificación fonética (A. Halberstadt, 1998)

- **Corpus TIMIT acústico-fonético**
 - Sólo clasificación independiente del contexto
 - Corpus de entrenamiento de 462 hablantes, conjunto de pruebas básico para 24 hablantes
 - Metodología de evaluación estándar, 39 clases fonéticas comunes
- **Se incorporan varias representaciones acústicas distintas**
 - Varias resoluciones de frecuencia-tiempo (Ventana *hamming* 10-30 ms)
 - Distintas representaciones espectrales (MFCC, PLPCC, etc)
 - Funciones de transformada de coseno frente a funciones de base constante a intervalos
- **Jerarquía MAP y métodos basados en comité**

Método	% Error
Línea de fondo	21.6
Jerarquía MAP	21.0
Comité de 8 clasificadores	18.5*
Comité con jerarquía	18.3

Enfoque estadístico al ASR



- Dadas las observaciones acústicas A , seleccione la secuencia de palabras W^* que maximiza la probabilidad *a posteriori*, $P(W | A)$

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

- La regla de Bayes se emplea típicamente para descomponer $P(W | A)$ en términos acústicos y lingüísticos

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

Consideraciones para la búsqueda en ASR

- Una búsqueda plena tiene en cuenta todas las segmentaciones posibles S junto con las unidades U , para cada secuencia hipotética de palabras W

$$W^* = \operatorname{argmax}_W P(W | A) = \operatorname{argmax}_W \sum_S \sum_U P(WUS | A)$$

- Puede buscar el mejor camino para simplificar la búsqueda, mediante la programación dinámica (ej., Viterbi) o las búsquedas de grafos (ej., A^*)

$$W^*, U^*, S^* \approx \operatorname{arg max}_{W,U,S} P(WUS | A)$$

- La descomposición modificada de Bayes presenta cuatro términos:

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

En los modelos HMM, éstos corresponden a posibilidades o probabilidades **acústicas**, **de estado** y del modelo de **lenguaje**

- **Modelos HMM**

- Velocidad de tramo variable (Ponting et al., 1991, Alwan et al., 2000)
- HMM basado en segmentos (Marcus, 1993)
- HMM segmental (Russell et al., 1993)

- **Modelado de la trayectoria**

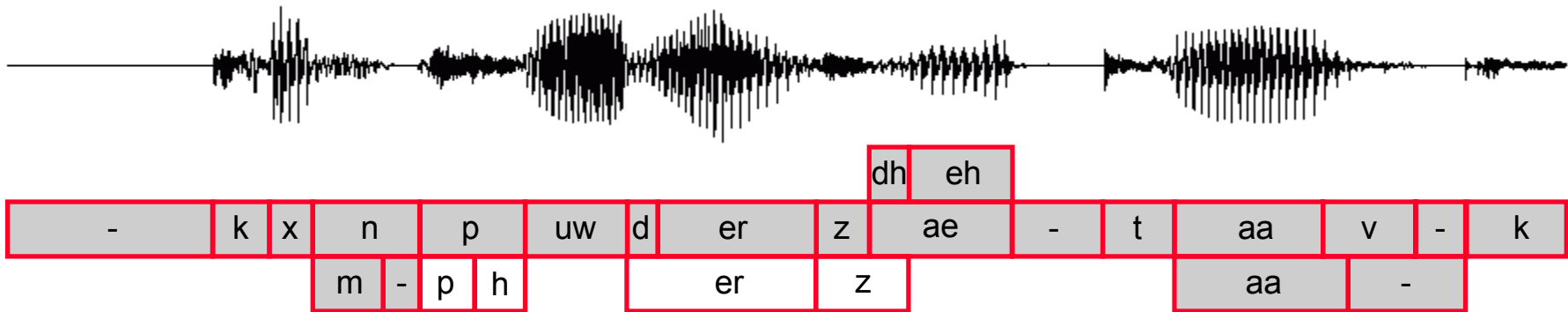
- Modelos de segmentos estocásticos (Ostendorf et al., 1989)
- Modelos de trayectoria paramétricos (Ng, 1993)
- Modelos de trayectoria estadísticos (Goldenthal, 1994)

- **Basados en rasgos**

- FEATURE (Cole et al., 1983)
- SUMMIT (Zue et al., 1989)
- LAFF (Stevens et al., 1992)

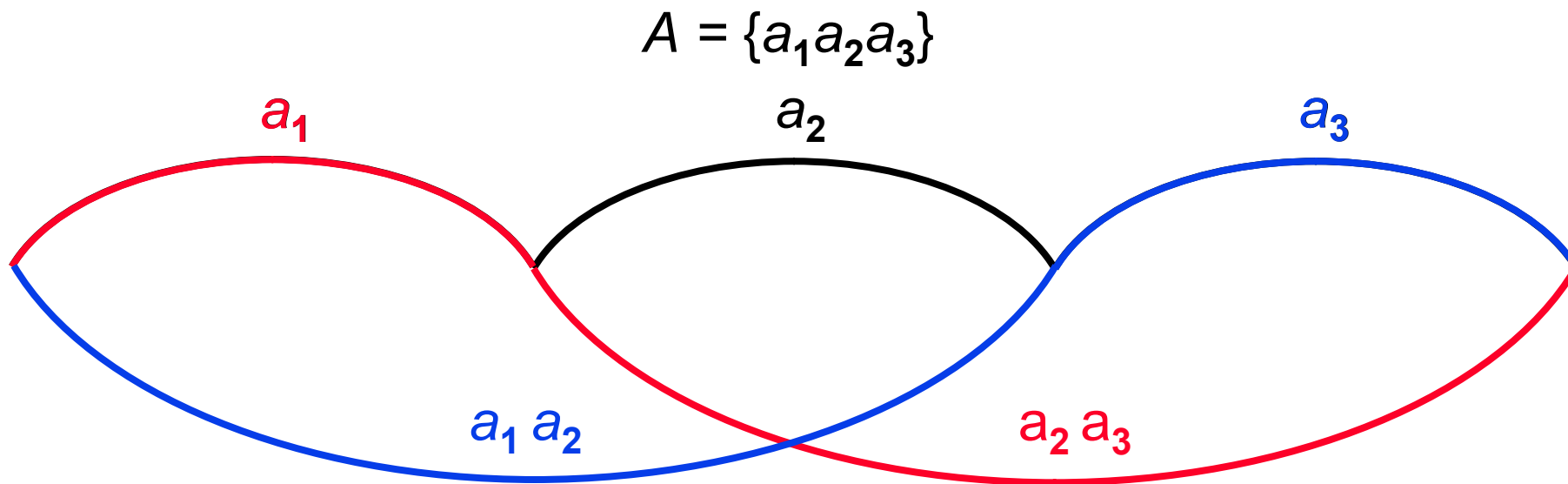
- **El modelado basado en segmentos de línea de fondo incorpora:**
 - Promedios y derivadas de coeficientes espectrales (ej., coeficientes MFCC)
 - Normalización de la dimensionalidad a través del análisis del componente principal
 - Estimación PDF a través de mezclas de gaussianas
- **Ejemplo de investigaciones sobre el modelado fonético-acústico**
 - Clasificadores probabilísticos alternativos (ej., Leung, Meng)
 - Mediciones de rasgos automáticamente aprendidos (ej., Phillips, Muzumdar)
 - Modelos de trayectoria estadísticos (Goldenthal)
 - Rasgos probabilísticos jerárquicos (ej., Chun, Halberstadt)
 - Modelado próximo al error (Chang)
 - Segmentación probabilística (Chang, Lee)
 - Clasificadores basados en comité (Halberstadt)

- **El sistema SUMMIT para el reconocimiento de voz está basado en segmentos fonéticos**
 - Los tiempos de comienzo y finalización explícitos del fono se plantean como hipótesis durante la búsqueda
 - Difiere de los métodos convencionales basados en tramos (ej., modelos HMM)
 - Permite el modelado fonético acústico basado en segmentos
 - Las mediciones pueden extraerse sobre límites y segmentos



- **El reconocimiento se consigue buscando un grafo fonético**
 - El grafo puede computarse mediante un criterio acústico o modelos probabilísticos
 - Las segmentaciones rivales utilizan distintos espacios de observación
 - La decodificación probabilística debe dar cuenta del espacio de observación basado en grafos

- El espacio de observación A corresponde a la secuencia temporal de tramos acústicos (ej., zonas espectrales)

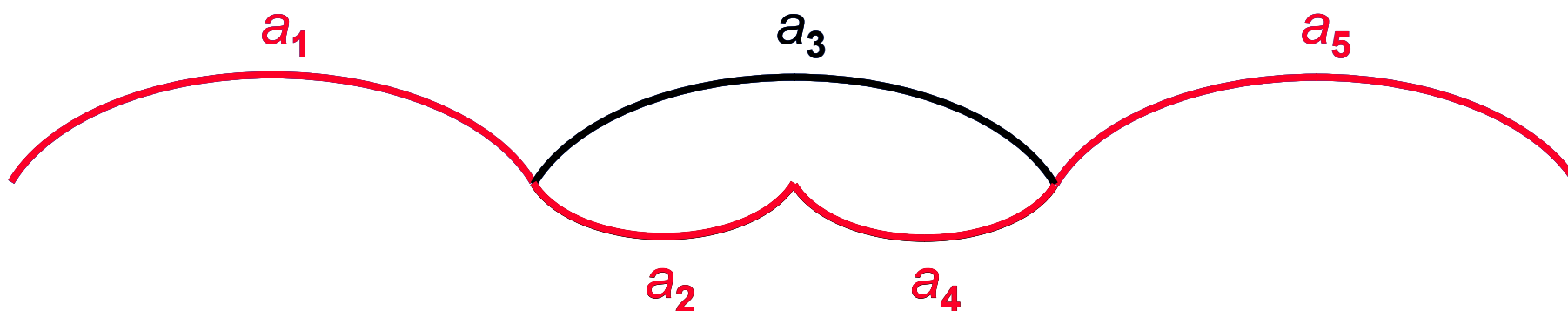


- Cada segmento hipotético s_i , está representado por las series de tramos computados entre los tiempos de comienzo y fin del segmento
- La probabilidad acústica $P(A|SW)$, se deriva del mismo espacio de observación para todas las hipótesis de la palabra

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

- Cada segmento S_i , está representado por un vector característico simple, a_i

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



$$X = \{a_1 a_3 a_5\}$$

$$Y = \{a_2 a_4\}$$

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

$$Y = \{a_3\}$$

- Dada una segmentación concreta S , A consta de X , que son los vectores característicos asociados con S , además de Y , que son los vectores característicos asociados con segmentos que **no** están en S :

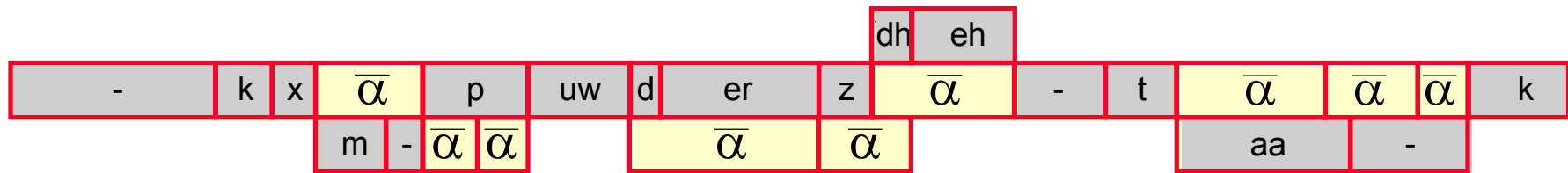
$$A = Y \cup X$$

- Para comparar las distintas segmentaciones, es necesario predecir la probabilidad de X e Y : $P(A|SW) = P(XY|SW)$

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

Búsqueda de espacios de observación basados en grafos: El modelo antifono

- Crear una unidad $\bar{\alpha}$, para modelar los segmentos que no sean fonos
- Para una segmentación S , asignar antifonos a segmentos
 - Todos los segmentos quedan explicados en el grafo fonético
 - Los caminos alternativos a través del grafo pueden compararse legítimamente



- Las probabilidades del camino se pueden descomponer en dos términos:
 - 1 La probabilidad de todos los segmentos producida por el antifono (una constante)
 - 2 La proporción de las probabilidades entre el fono y el antifono para todos los segmentos del camino
- Formulación MAP para la secuencia de palabras más probable W , determinada por:

$$W^* = \operatorname{argmax}_{W,S} \prod_i^{N_S} \frac{P(x_j | u_j)}{P(x_j | \bar{\alpha})} P(s_j | u_j) P(U | W) P(W)$$

Modelado de unidades no-léxicas: El antifono

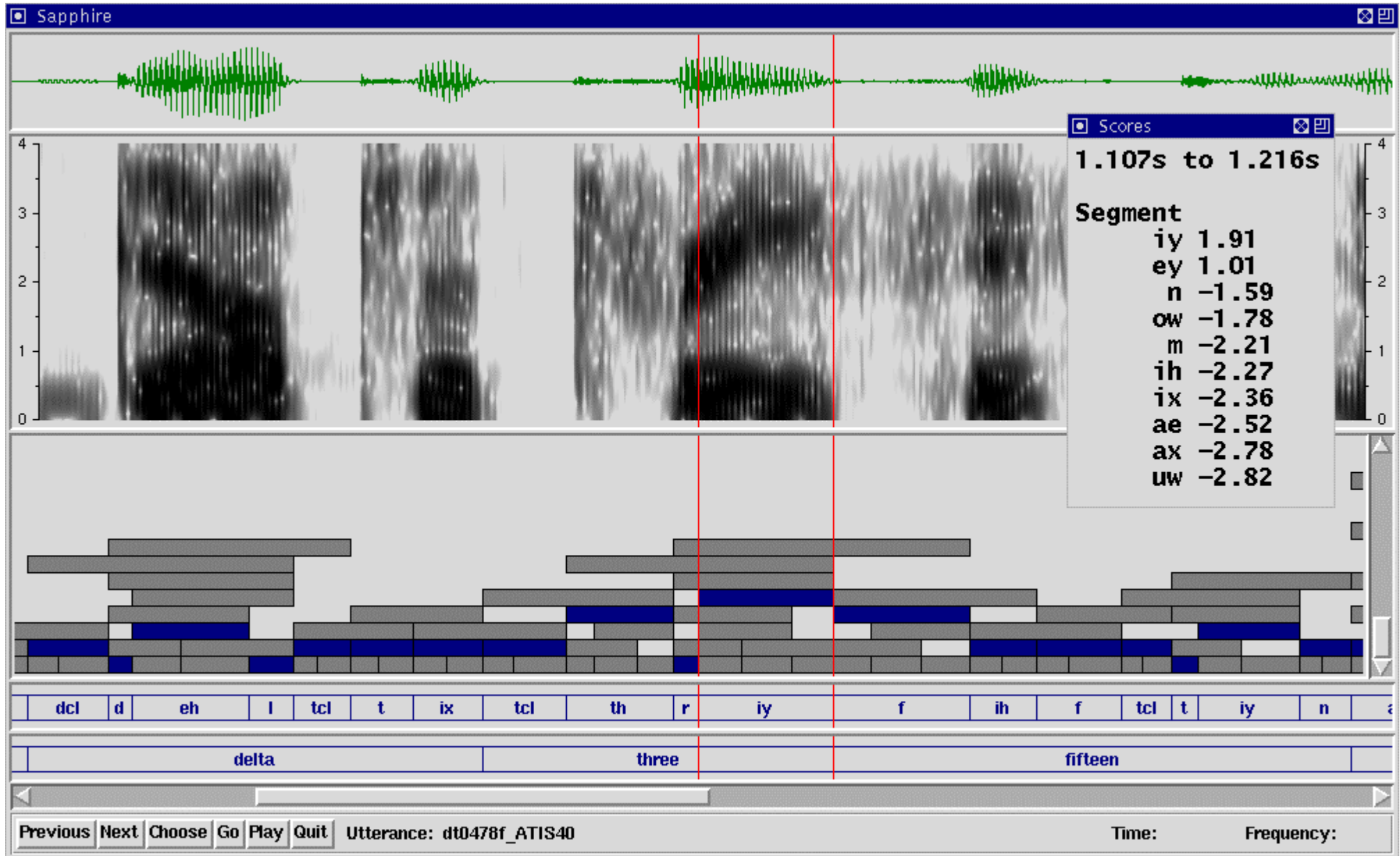
- Dada una segmentación especial S , A consta de X , que son los segmentos asociados con S , además de Y , que son los segmentos **no** asociados con S : $P(A|SU)=P(XY|SU)$
- Dada la segmentación S , asignar vectores característicos en X para validar las unidades, y el resto en Y para el antifono
- Siendo $P(XY | \bar{\alpha})$ una constante K , podemos escribir $P(XY|SU)$, suponiendo la independencia entre X e Y

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

- Es necesario que durante la búsqueda consideremos sólo segmentos en S :

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

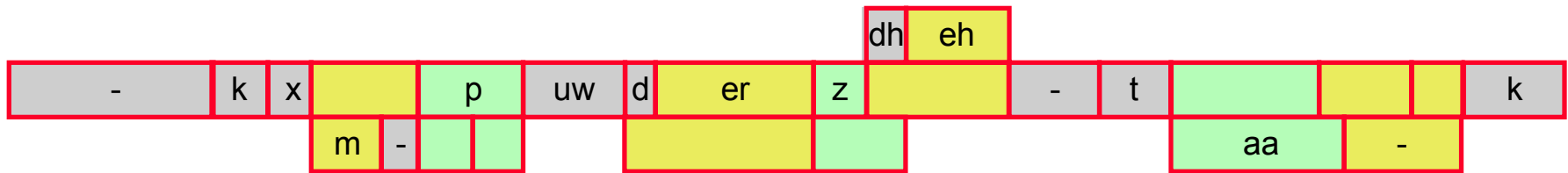
Sistema SUMMIT para el ASR basado en segmentos



- **Modelos de espacios de observación absoluta, que emplean ejemplos positivos y negativos**
- **La puntuaciones de la probabilidad logarítmica se normalizan por el antifono**
 - Las buenas puntuaciones son positivas, las malas son negativas
 - Todos los segmentos pobres presentan puntuaciones negativas
 - Útil para el recorte y/o el rechazo
 - El antifono no se emplea para el acceso léxico
- **Durante la búsqueda no se utilizan probabilidades anteriores o posteriores**
 - Permite la computación sobre la demanda y/o la correspondencia rápida
 - Los subconjuntos de datos pueden utilizarse para el entrenamiento
- **Se pueden utilizar modelos dependientes o independientes del contexto**
- **Útil para problemas generales de correspondencia de patrones con espacios de observación basados en grafos**

Más allá de los antifonos: Modelado casi error

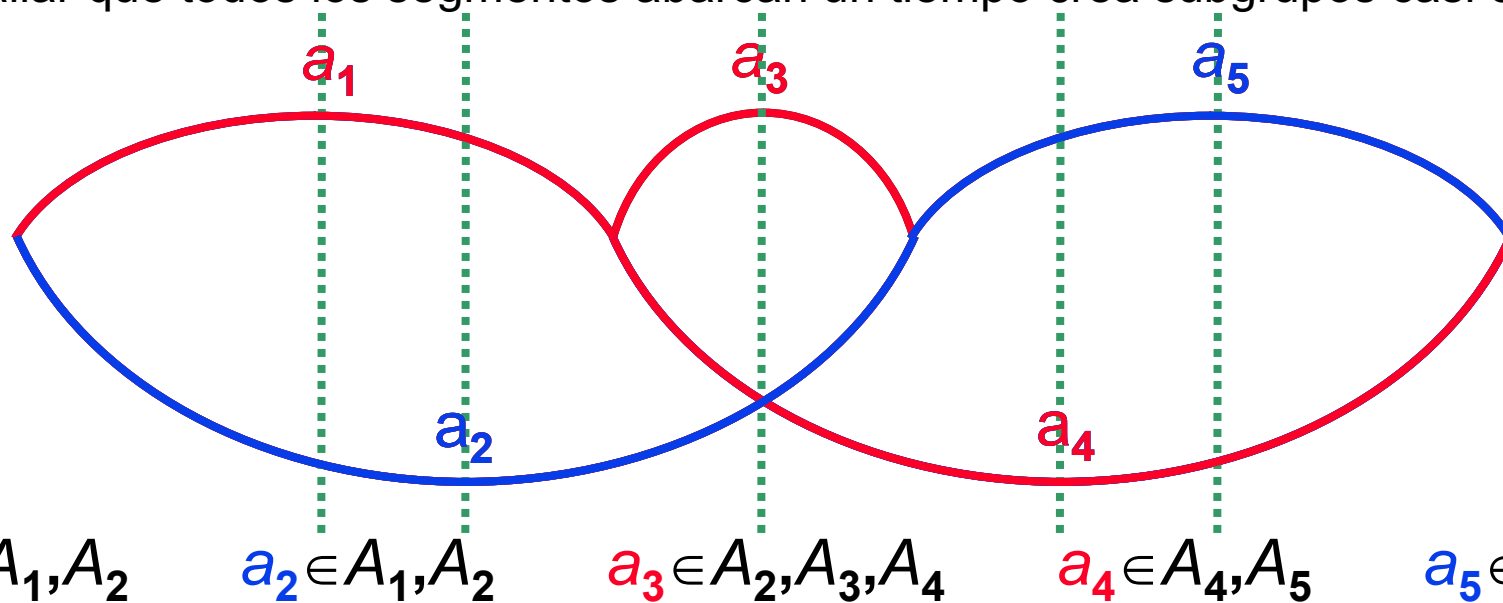
- El modelado antifono divide el espacio de observación en dos partes (ej., sobre una segmentación hipotética o no)
- El modelado casi error divide el espacio de observación en un grupo de subgrupos mutuamente exclusivos y conjuntamente exhaustivos
 - Un subgrupo casi error precomputado para cada segmento del grafo
 - El criterio temporal puede garantizar una generación adecuada de subgrupo casi error (ej., el segmento A es un casi error de B si y sólo si B abarca el medio punto de A)



- Durante el reconocimiento, las observaciones de un subgrupo casi error corresponden al modelo casi error del fono hipotético
- Los modelos casi error pueden ser simplemente un antifono, pero pueden llegar a ser más sofisticados (ej., dependiente del fono)

Creación de subgrupos casi error

- Los subgrupos casi error A_i , asociados con cualquier segmentación S , deben ser mutuamente exclusivos y exhaustivos: $A = \cup A_i \quad \forall A_i \in S$
- El criterio temporal garantiza los subgrupos casi error adecuados
 - Los segmentos contiguos en S explican todos los tiempos exactamente una vez
 - Hallar que todos los segmentos abarcan un tiempo crea subgrupos casi error

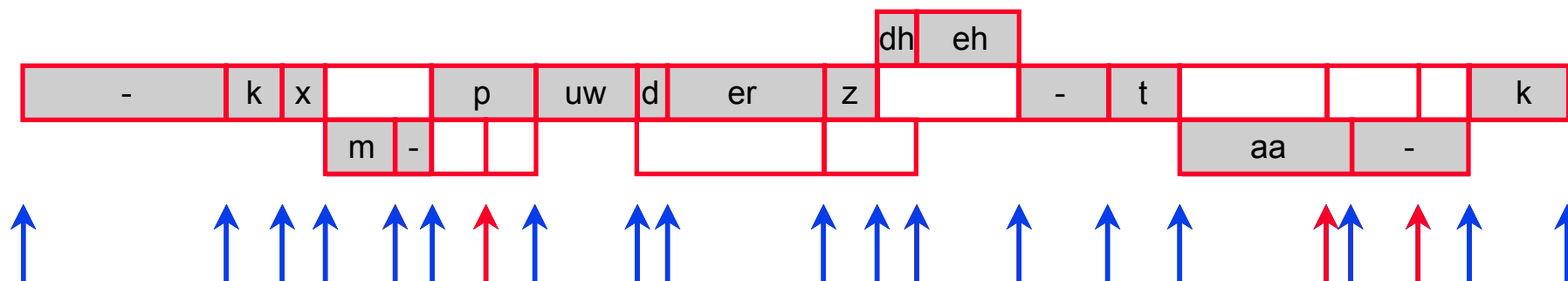


$$A_1 = \{a_1 a_2\} \quad A_2 = \{a_1 a_2 a_3\} \quad A_3 = \{a_3\} \quad A_4 = \{a_3 a_4 a_5\} \quad A_5 = \{a_4 a_5\}$$

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

Modelado de límites

- También podemos incorporar vectores característicos adicionales computados en límites hipotéticos o fronteras de fonos

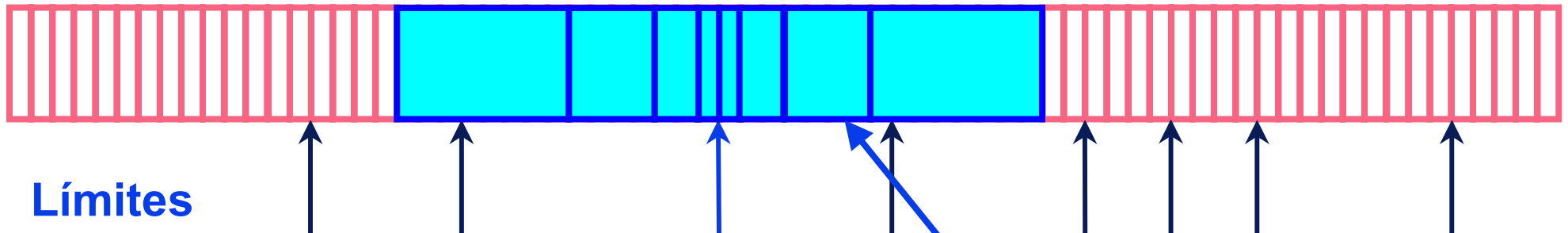


- Cada segmentación da cuenta de cada límite
 - Algunos límites serán **transiciones** entre unidades léxicas
 - Otros límites se considerarán **internos** a la unidad
- Son posibles tanto las unidades dependientes de contexto como las independientes
- Se modelan eficazmente las transiciones entre fonos (ej., **difonos**)
- Los modelos basados en tramos pueden utilizarse para generar segmentos del grafo

Modelado de límites

- **Mediciones basadas en tramos:**
 - Computadas cada 5 milisegundos
 - Vector característico de 14 coeficientes cepstrales de escala Mel (MFCC)

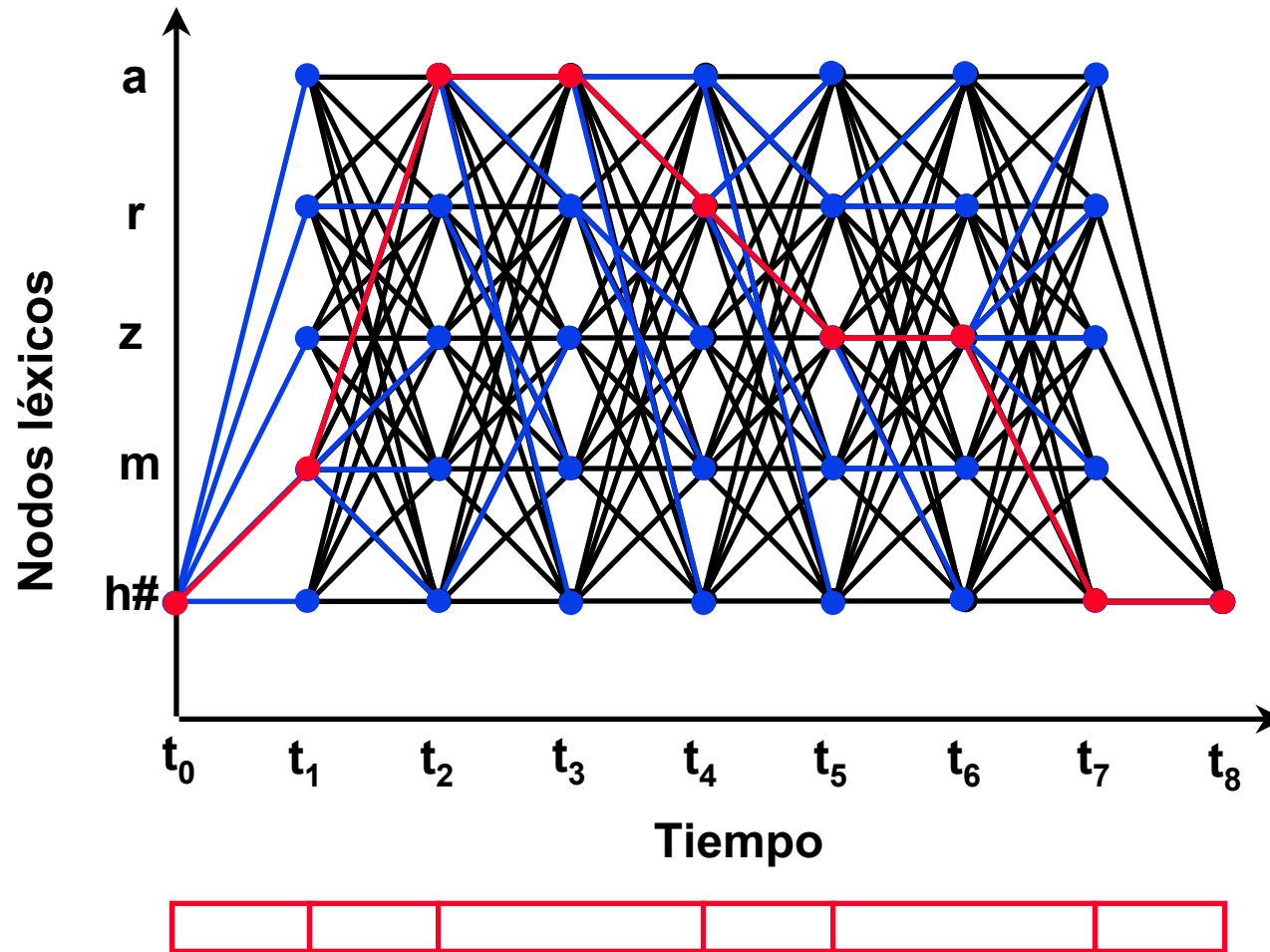
Vectores característicos basados en tramos



- **Mediciones basadas en límites:**
 - Computar el promedio de coeficientes MFCCs sobre 8 regiones alrededor del límite
 - 8 regiones X 14 promedios de MFCC = 112 vectores de dimensión
 - 112 dimensiones reducidas a 50 empleando el análisis del componente principal

Segmentación probabilística

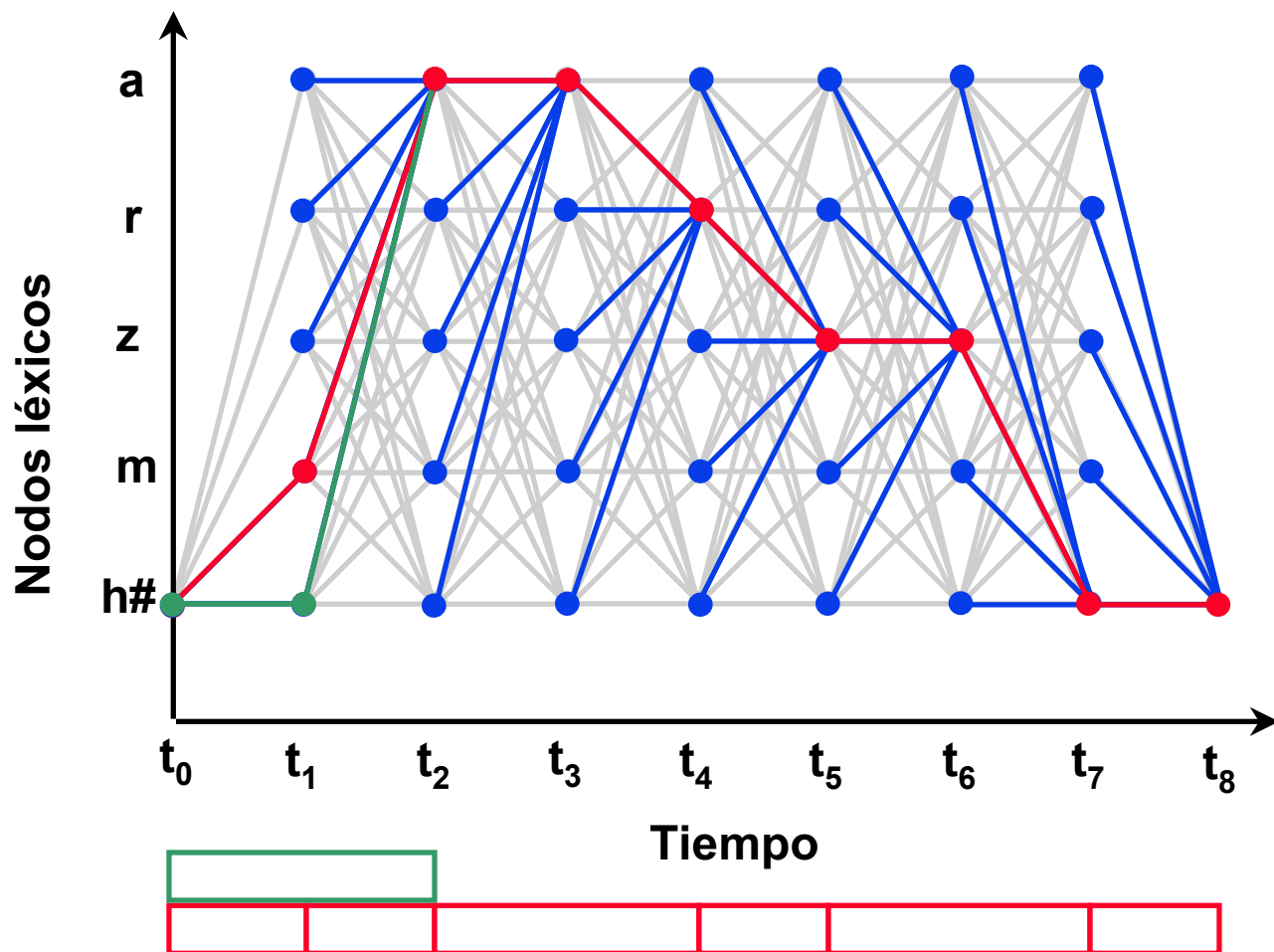
- Emplea la búsqueda de Viterbi hacia adelante como primer paso para hallar el mejor camino



- Se utilizan umbrales relativos y absolutos para acelerar la búsqueda

Segmentación probabilística (continuación)

- El segundo paso emplea la búsqueda A^* hacia atrás para hallar los caminos N -mejores
- El rastreo hacia atrás de Viterbi se utiliza como una estimación futura para las puntuaciones de los caminos



- El procesamiento de bloqueo permite la computación mediante algoritmos de conductos

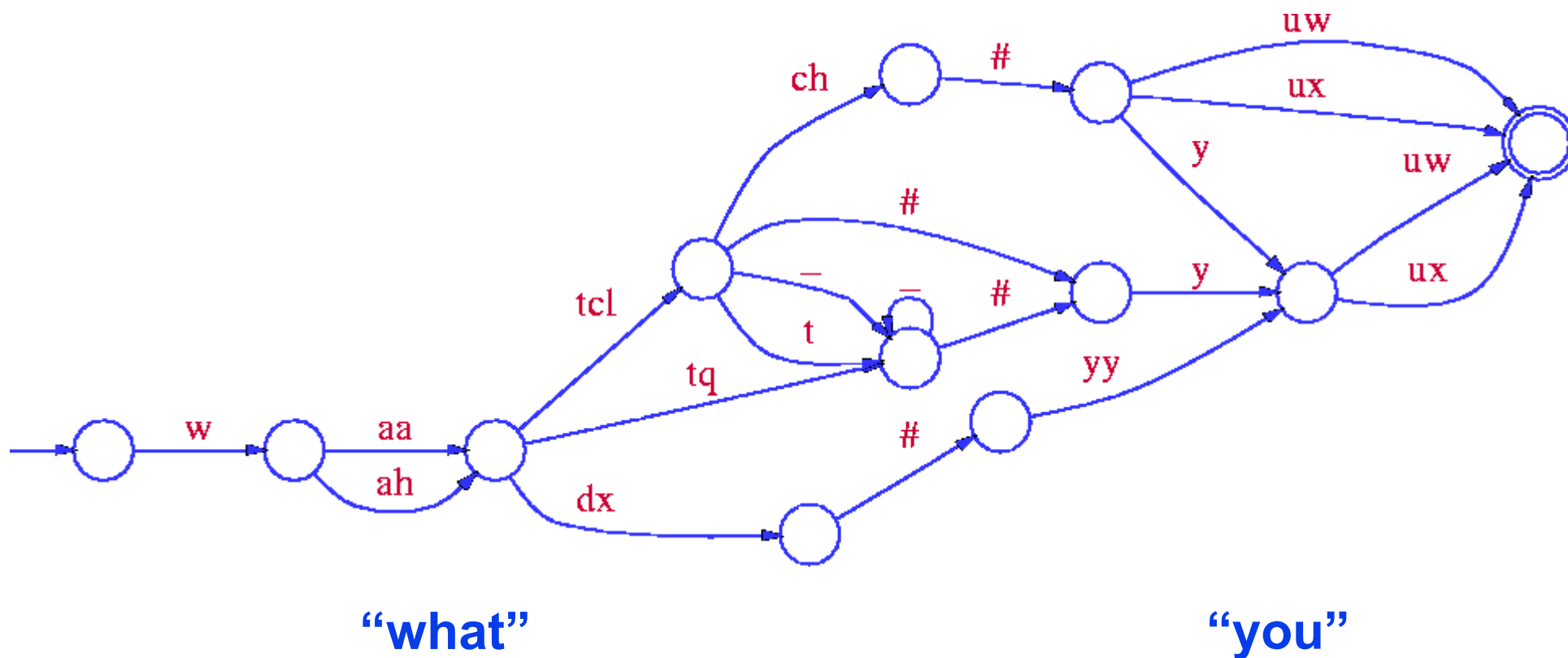
- **Corpus TIMIT acústico-fonético**
 - Corpus de entrenamiento de 462 hablantes, conjunto de pruebas básico para 24 hablantes
 - Metodología de evaluación estándar, 39 clases fonéticas comunes
- **Representaciones del segmento y límites basadas en el promedio y en las derivadas de 14 coeficientes MFCC, energía y duración**
- **PCA utilizado para la normalización de datos y para la reducción**
- **Modelos acústicos basados en mezclas de gaussianas agregadas**
- **Modelo de lenguaje basado en bigramas de fonos**
- **Segmentación probabilística computada a partir de modelos trifonos**

Método	% Error
Trifono CDHMM	27.1
Red neural recurrente	26.1
HMM trifono bayesiano	25.6
Antifono, clasificadores heterogéneos	24.4

- **Palabras descritas mediante formas bases fonémicas**
- **Las reglas fonológicas expanden formas bases en el grafo, ej.,**
 - La eliminación de la oclusiva estalla en la coda de la sílaba (ej., *laptop* (ordenador portátil))
 - Eliminación de /t/ en varios entornos (ej., *intersection, destination, crafts)*
 - Geminación de fricativas y nasales (ej., *this side, in nome*)
 - Asimilación del lugar (ej. ., *did you* (/d ih jh uw/))
- **Las probabilidades del arco $P(U|W)$, pueden ser entrenadas**
- **La mayoría de los modelos HMM no presentan un componente fonológico**

Ejemplo fonológico

- **Ejemplo de “what you” expandido en un reconocedor SUMMIT**
 - La /t/ final de “what” puede producirse como liberada, no liberada, palatalizada u oclusiva glotal, o como un flap



Experimentos de reconocimiento de voz

- **Corpus Jupiter, en entorno telefónico, consultas sobre el tiempo**
 - Conjunto de entrenamiento de 50.000 enunciados, conjunto de pruebas de 1806 enunciados "del dominio"
- **Modelos acústicos basados en mezclas de gaussianas**
 - Representaciones del segmento y de los límites basadas en promedios y derivadas de 14 coeficientes MFCC, energía y duración
 - PCA utilizado para la normalización de datos y para la reducción
 - 715 clases frontera dependientes del contexto
 - 935 trifonos, 1160 clases de segmentos dependientes del contexto difono
- **El grafo de pronunciación incorpora probabilidades de pronunciación**
- **Modelo de lenguaje basado en bigrama y trigrama de clase**
- **El mejor rendimiento se consigue combinando modelos**

Método	% Error
Modelos frontera	7.6
Modelos del segmento	9.6
Combinado	6.1

- **Algunas técnicas de reconocimiento de voz basadas en segmentos transforman el espacio de observación de tramos en grafos**
- **Los espacios de observación basados en grafos permiten una amplia variedad que va desde métodos de modelado alternativo hasta enfoques basados en tramos**
- **Los marcos de modelado antifono y casi error facilitan el mecanismo de búsqueda de espacios de observación basados en grafos**
- **Se han conseguido buenos resultados para el reconocimiento fonético**
- **Aún queda mucho trabajo por realizar**

- **J. Glass, “A Probabilistic Framework for Segment-Based Speech Recognition,” próximamente en *Computer, Speech & Language*, 2003.**
- **D. Halberstadt, “Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition,” Tesis doctoral 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.**