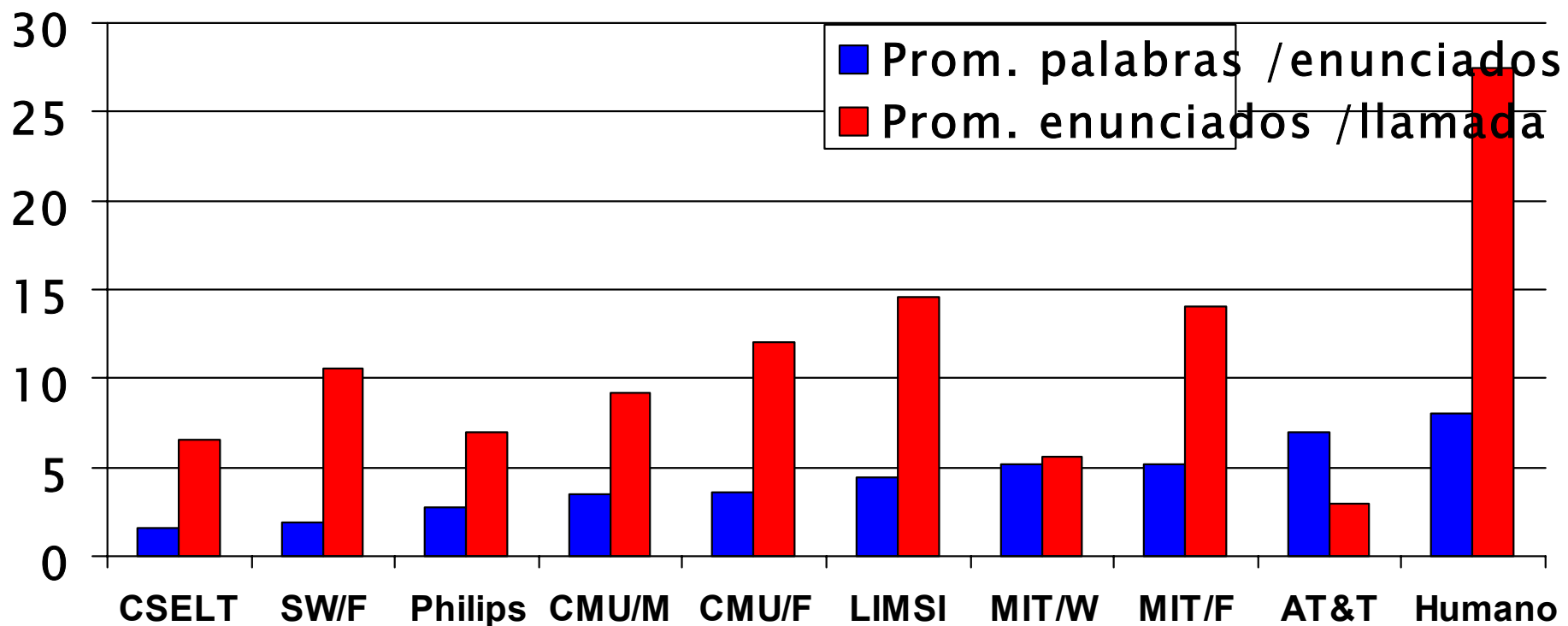


ASR para sistemas de diálogos hablados

- **Introducción**
- **Cuestiones de reconocimiento de voz**
 - Ejemplo del sistema SUMMIT en un entorno de información sobre el tiempo
- **Reducción de la computación**
- **Modelo de agregación**
- **Clasificadores basados en comité**

Ejemplo de sistemas basados en diálogo



- Los vocabularios normalmente poseen miles de palabras
- Los sistemas muy utilizados tienden a ser más conservadores
- Los diálogos dirigidos presentan menos palabras por enunciado
- La media de palabras descende con más confirmaciones
- En las conversaciones entre humanos se emplean más palabras

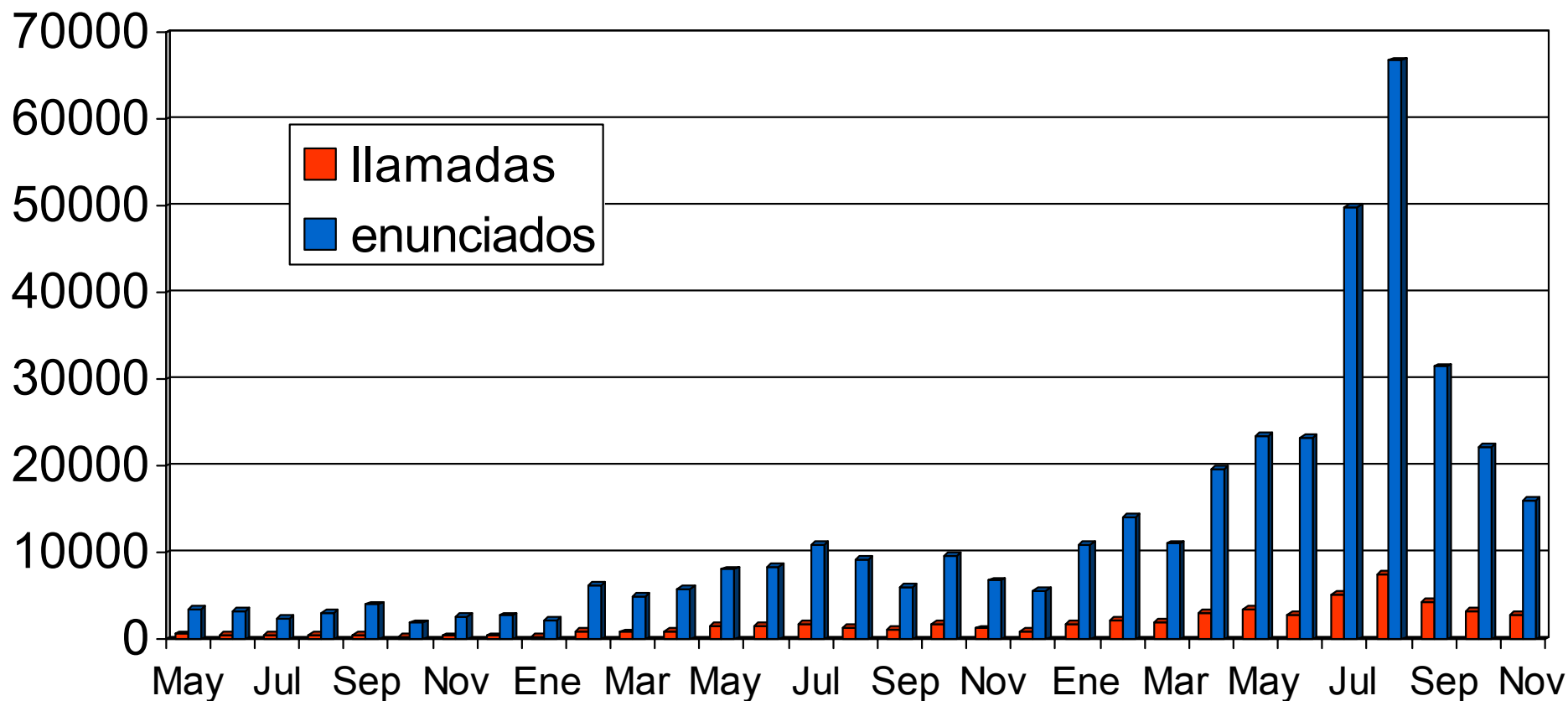
- **Amplitud de banda del teléfono con auriculares variables**
- **Condiciones de fondo con ruidos**
- **Usuarios principiantes con un pequeño número de interacciones**
 - Hombres, mujeres, niños
 - Hablantes nativos y no nativos
 - Consultas auténticas, navegadores, piratas informáticos
- **Efectos del discurso espontáneo**
 - ej., pausas rellenas, palabras parciales, artefactos no discursivos
- **Palabras que no están en el vocabulario y preguntas no pertenecientes al entorno**
- **Vocabulario completo necesario para un entendimiento absoluto**
 - El reconocimiento en contexto de palabras y sintagmas no son estrategias fundamentales
 - Los diálogos en los que ambas partes toman la iniciativa supone poca limitación para el reconocido
- **Decodificación en tiempo real**

Cuestiones de recopilación de datos

- El desarrollo del sistema es un problema como el del huevo y la gallina
- La recopilación de datos ha evolucionado significativamente
 - Basado en un mago → recopilación de datos basada en el sistema
 - Utilización del laboratorio → uso público
 - Centenas de usuarios → miles → millones
- Los datos de usuarios **reales** solventando problemas **reales** aceleran el desarrollo de la tecnología
 - Considerablemente distinto a un entorno de laboratorio
 - Enfatiza las debilidades, permite la evaluación continua
 - No obstante, precisa de **sistemas** que proporcionen información **real**
- La ampliación de los corpus precisa un entrenamiento no supervisado o la adaptación a datos no etiquetados

Recopilación de datos (Entorno meteorológico)

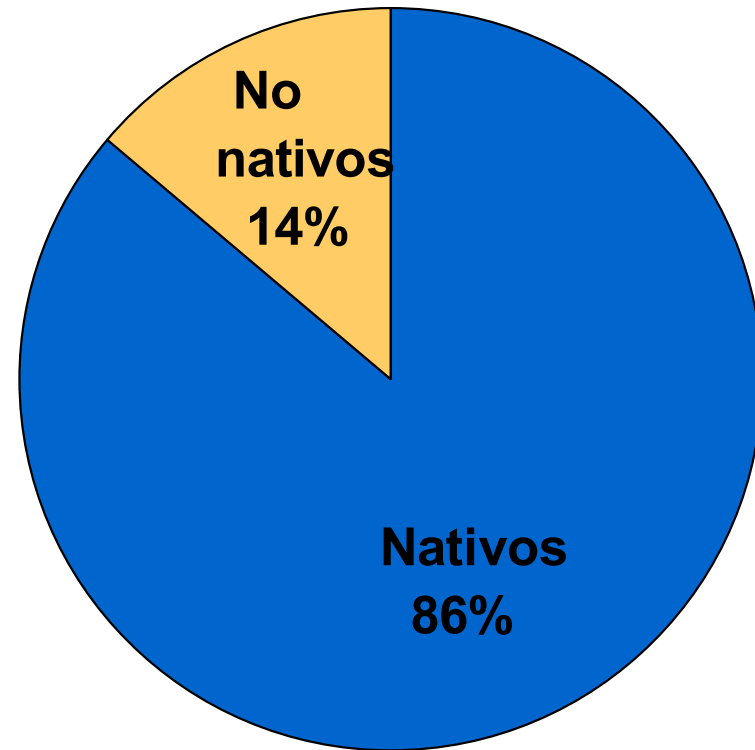
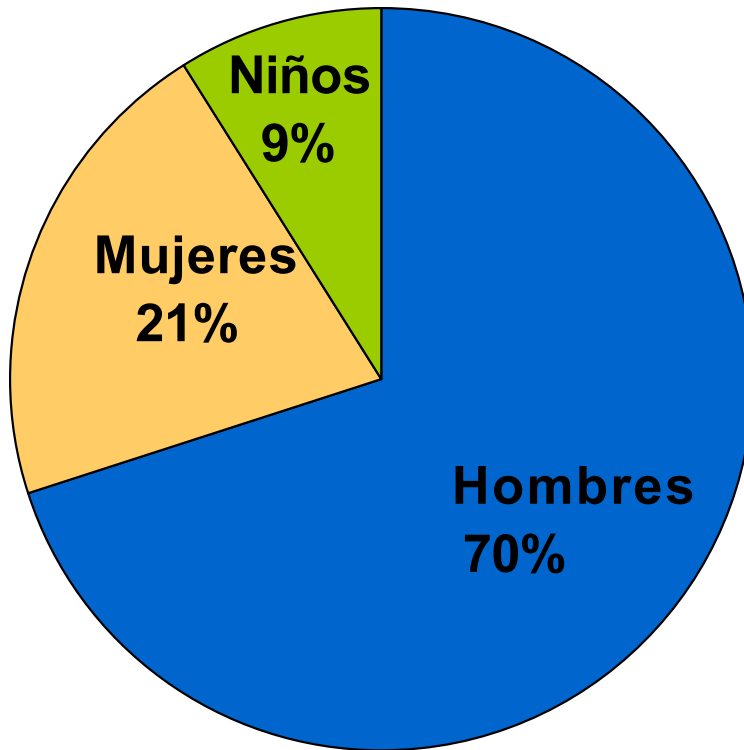
- Recopilación inicial de 3.500 enunciados leídos y 1.000 enunciados del mago



- Más de 756K de enunciados de 112K de llamadas desde mayo, 1997

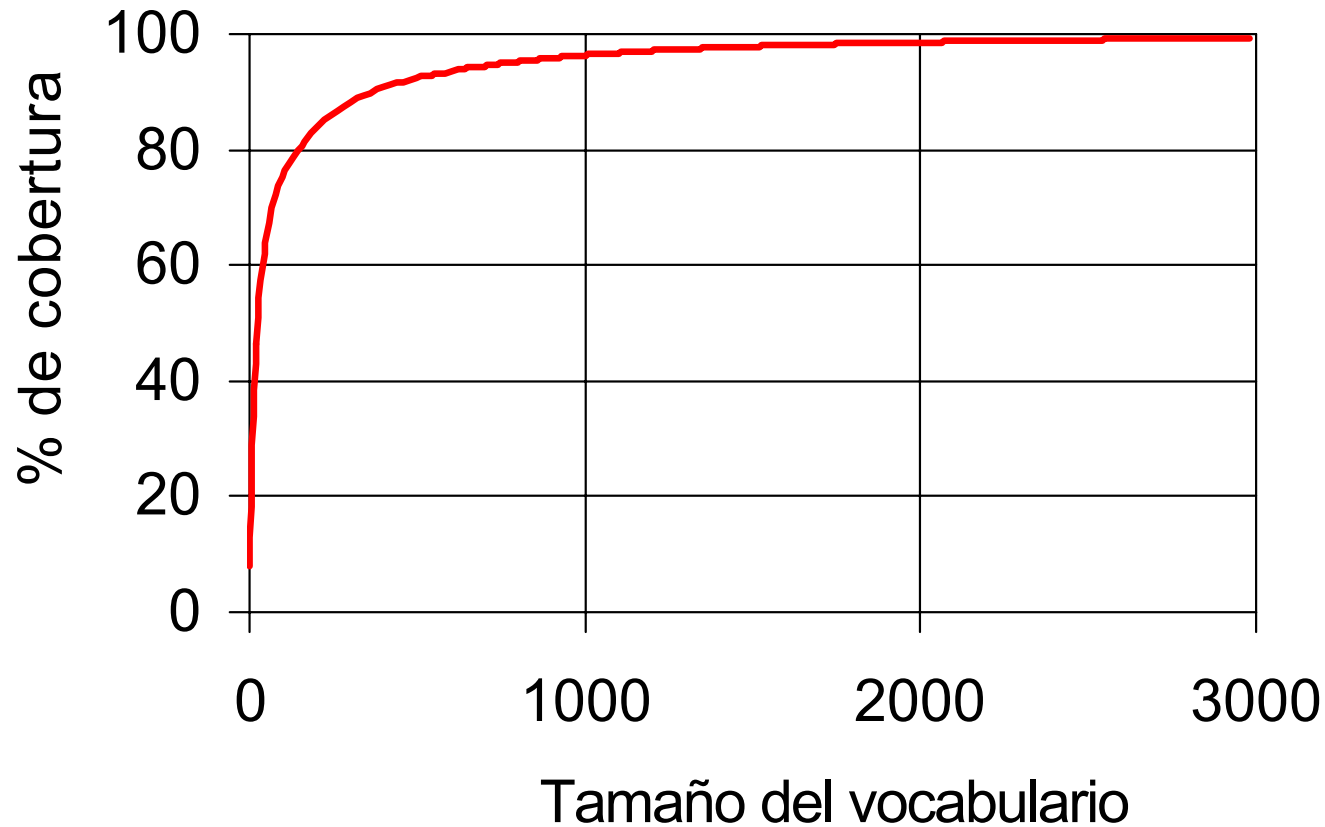
Características de un corpus de información meteorológica

- **Corpus dominado por hablantes americanos masculinos**



- **Aproximadamente el 11% de los casos presentaban ruidos considerables**
- **Más del 6% de los datos presentaba resultados de habla espontánea**
- **Al menos un 5% de los datos procedían de hablantes telefónicos**

Selección del vocabulario



- Los entornos restringidos limitan claramente el tamaño del vocabulario
- Un vocabulario de 2000 palabras proporciona buena cobertura para un entorno atmosférico
- menos del 2% del índice de palabras no presentes en el vocabulario está en los grupos de prueba

Vocabulario

- El vocabulario actual consta de casi 2000 palabras
- Basado en las capacidades del sistema y en las preguntas del usuario

Tipo	Tamaño	Ejemplos
Geografía	933	boston, alberta, france, africa
Tiempo	217	temperature, snow, sunny, smog
Básico	815	i, what, january, tomorrow

- Incorporación de palabras comunes reducidas y pares de palabras

Tipo	Ejemplos
Reducción	give_me, going_to, want_to, what_is, i_would
Compuesto	clear_up, heat_wave, pollen_count

- Lexicón basado en el diccionario silábico LDC PRONLEX

Ejemplo de un archivo de vocabulario

Ordenado alfabéticamente

<>*	Marcador de comienzo y fin del enunciado
<pause1>	Pausas a comienzos y fin de enunciados
<pause2>	Modelos de pausa rellenos
<uh>	Los elementos *'d no tienen realización acústica
<um>	
<unknown>*	Modelos de palabra no presentes en el vocabulario
a	Las palabras <>'d no cuentan como error
a_m	Los guiones distinguen las secuencias de letras de las palabras reales
am	
don+t	El símbolo + se utiliza convencionalmente para ' ,
new_york_city	Las minúsculas son una convención común
sixty	
today	Los números se deletrean
today+s	Cada palabra forma una entrada separada

Ejemplo de archivo de formas base

<pause1>

: ⊕

<pause2>

: - +

<uh>

: ah_fp

<um>

: ah_fm

a_m

: ey & eh m

either

: (iy , ay) th er

laptop

: l ae pd t aa pd

new_york

: n uw & y ao r kd

northwest

: n ao r th w eh s td

trenton

: tr r eh n tq en

winter

: w ih nt er

el símbolo anterior
se puede repetir

vocal de pausa con
relleno especial

pronunciaciones
alternas

rotura de palabra
permitiendo pausa

Edición de formas base generadas

- El archivo de formas base automáticamente generado debería comprobarse manualmente para evitar los siguientes problemas:
 - Variantes de pronunciación que faltan y son necesarias
 - Variantes de pronunciación no deseadas que están presentes
 - Palabras del vocabulario que faltan en PRONLEX

going_to	: g ow ix ng & t uw
reading	: (r iy df ix ng , r eh df ix ng)
woburn	: <???



going_to	: g (ow ix ng & t uw , ah n ax)
reading	: r eh df ix ng
woburn	: w (ow , uw) b er n

Aplicación de reglas fonológicas

- Las formas base *fonémicas* son una representación canónica
- Las formas base pueden tener múltiples realizaciones acústicas
- Las realizaciones acústicas son *fonos* o *unidades fonéticas*
- Ejemplo:

batter : b ae t f er

Esto puede realizarse fonéticamente como:

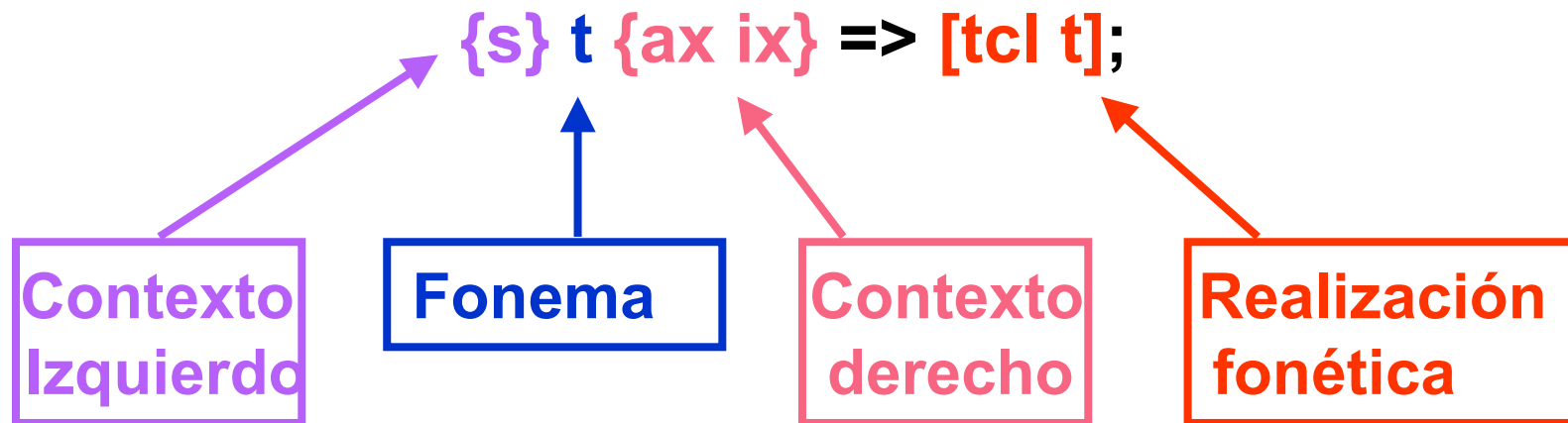
bcl b ae **tcl t** er Standard /t/

o como:

bcl b ae **dx** er Flapped /t/

Ejemplo de reglas fonológicas

- Regla ejemplo para la eliminación de la /t/ (“destination”):



- Regla ejemplo para la palatalización de la /s/ (“miss you”):

$$\{ \} s \{y\} \Rightarrow s \mid sh;$$

Modelado del lenguaje

- Bigramas y trigramas de clase utilizados para generar las 10 mejores salidas
- Datos de entrenamiento aumentados con restricciones de ciudad y estado
- Medida de entropía relativa empleada para ayudar a seleccionar clases

raining, snowing	humidity, temperature
cold, hot, warm	advisories, warnings
extended, general	conditions, forecast, report

- 200 clases de palabras redujeron la perplejidad e índice de error

Tipo	Perplejidad	% Índice de error por palabra
bigrama de palabra	18.4	16.0
+ trigrama de palabra	17.8	15.5
bigrama de clase	17.6	15.6
+ trigrama de clase	16.1	14.9

Definiendo clases de palabra n-grama

CITY ==> boston

CITY ==> chicago

CITY ==> seattle

<U>_DIGIT ==> one

<U>_DIGIT ==> two

<U>_DIGIT ==> three

DAY ==> today | tomorrow

Las definiciones de clase presentan el nombre de la clase a la izquierda y la palabra a la derecha

Los nombres de clase con “<U>_” imponen la misma probabilidad a todas las palabras

Las palabras alternas de una clase pueden situarse en la misma línea con un separador “|”

El archivo de la oración de entrenamiento

- Un modelo n -grama se calcula a partir de datos de entrenamiento
- El archivo de entrenamiento contiene un enunciado por línea
- Las palabras del archivo de entrenamiento deben tener el mismo caso y forma que las palabras del archivo del vocabulario
- El archivo de entrenamiento emplea las siguientes convenciones:
 - Cada enunciado limpio comienza con **<pause1>** y acaba con **<pause2>**
 - Los guiones de las palabras compuestas se eliminan antes del entrenamiento
 - Los guiones se vuelven a insertar automáticamente durante el entrenamiento debido a las palabras compuestas presentes en el archivo del vocabulario
- Pueden utilizarse unidades artefacto especiales para los ruidos y otros eventos significativos no discursivos:
 - **<clipped1>**, **<clipped2>**, **<hangup>**, **<cough>**, **<laugh>**

Ejemplo de archivo de la oración de entrenamiento

<pause1> when is the next flight to chicago <pause2>

<pause> to san <partial> san francisco <pause2>

<pause1> <um> boston <pause2>

palabra parcial,
ej., san die(go)

<clipped1> it be in time <pause2>

palabra cortada,
ej., ~(w)ill it

<pause1> good bye <hangup> (cuelga)

<pause1> united flight two oh four <pause2>

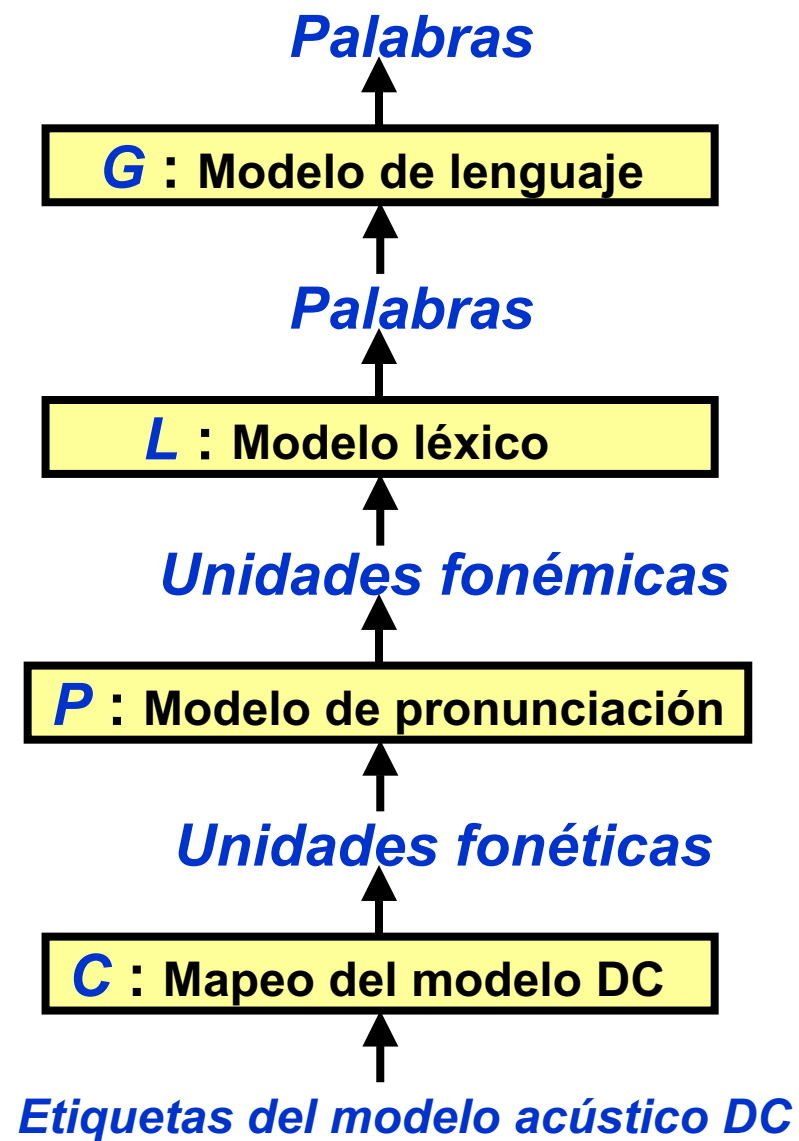
<pause1> <cough>(tos) excuse me <laugh>(risa) <pause2>

todos los sonidos significativos se transcriben

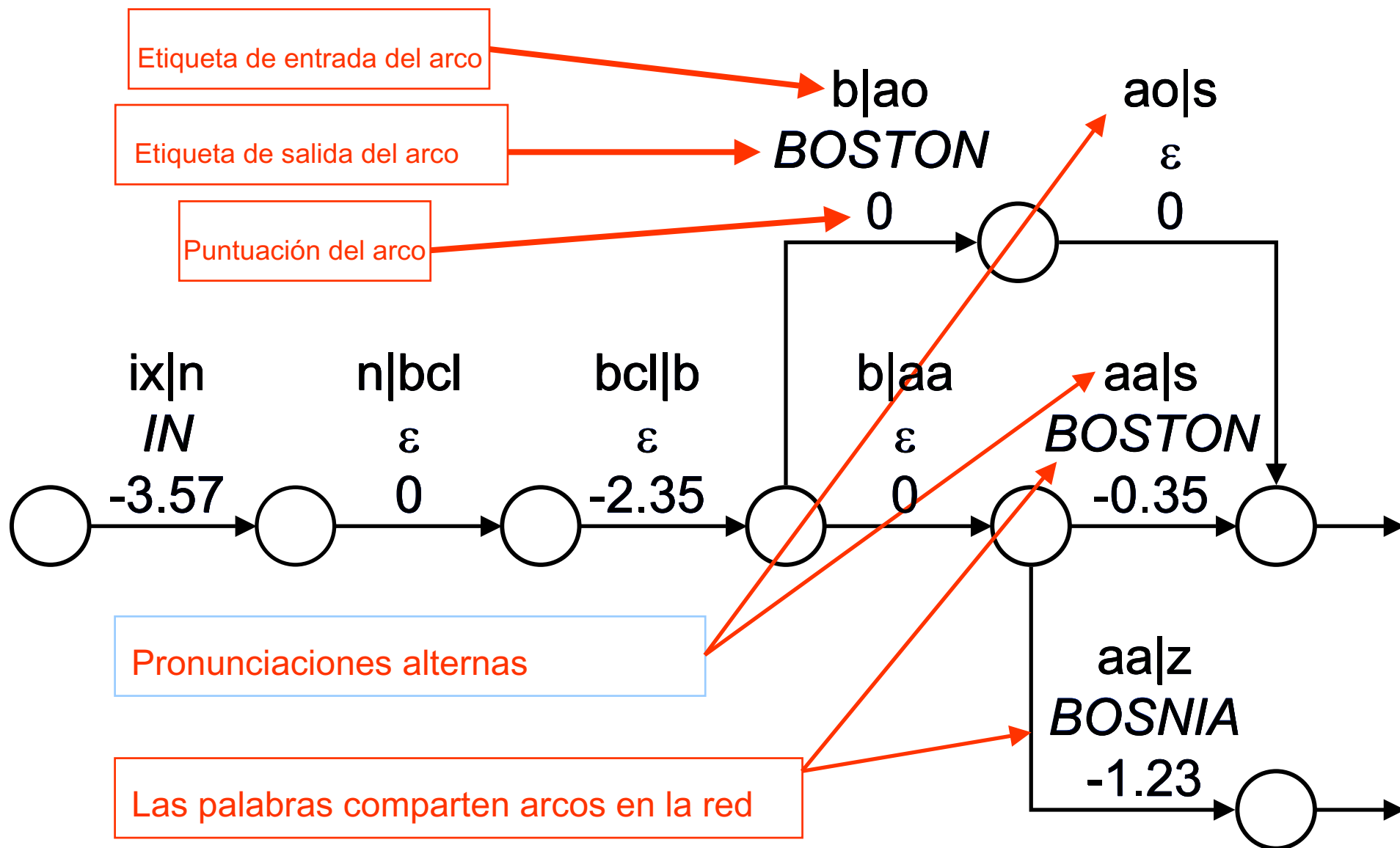
Composición de redes léxicas de FST

- Cuatro redes de FST se componen para formar una red de búsqueda plena
 - **G** : Modelo de lenguaje
 - **L** : Modelo léxico
 - **P** : Modelo de pronunciación
 - **C** : Mapeo del modelo acústico dependiente del contexto
- Matemático compuesto empleando la expresión:

CoPoLoG



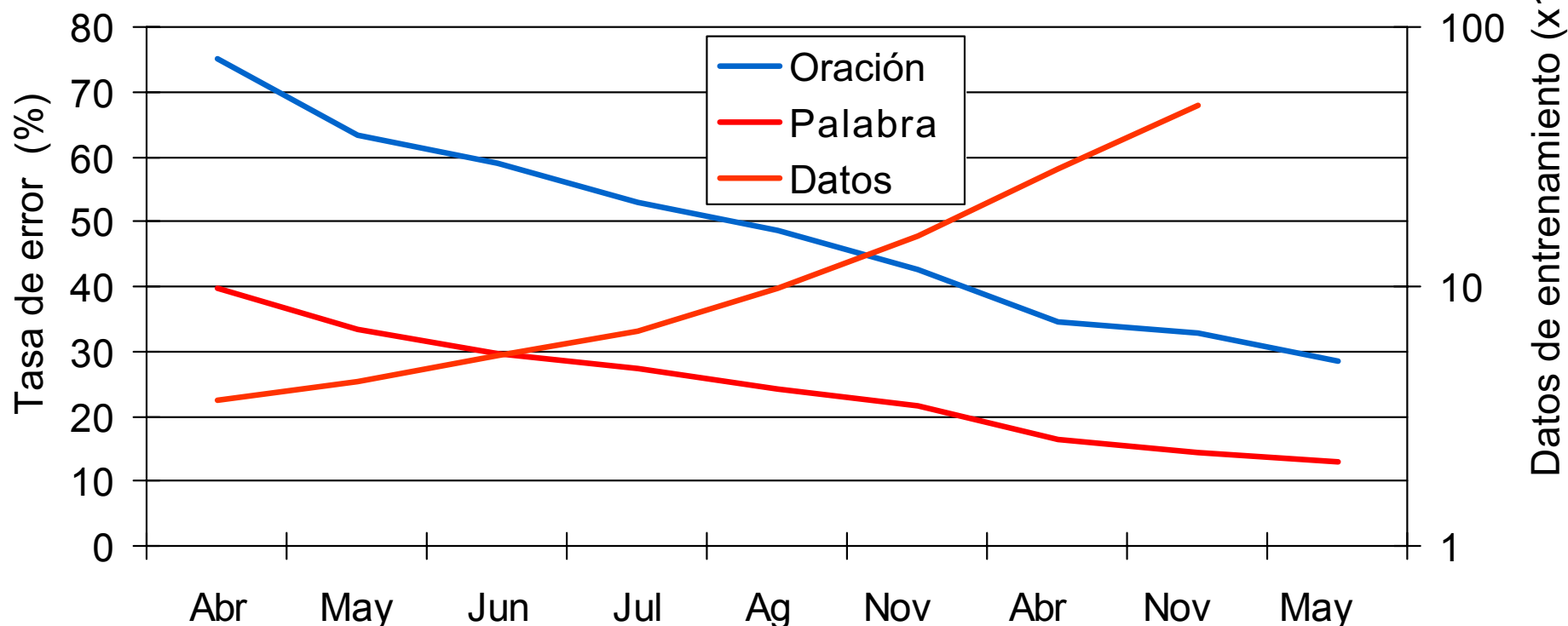
Ejemplo de FST



- **Los modelos se pueden construir para segmentos y límites**
 - Se puede conseguir la mejor exactitud cuando ambos se utilizan
 - El reconocimiento actual en *Tiempo-real* emplea sólo modelos límite
- **Etiquetas límite combinadas en las clases**
 - Clases determinadas mediante el agrupamiento del árbol de decisión
 - Un modelo de mezcla de gaussiana entrenado por clase
 - Un vector característico de dimensión 112 reducido a 50 dimensiones a través de PCA
 - 1 componente gaussiano para cada 50 muestras de entrenamiento (basado en el número de dimensiones)
- **Modelos entrenados durante más de 100 horas de discurso espontáneo por teléfono, recopilado de varios entornos**

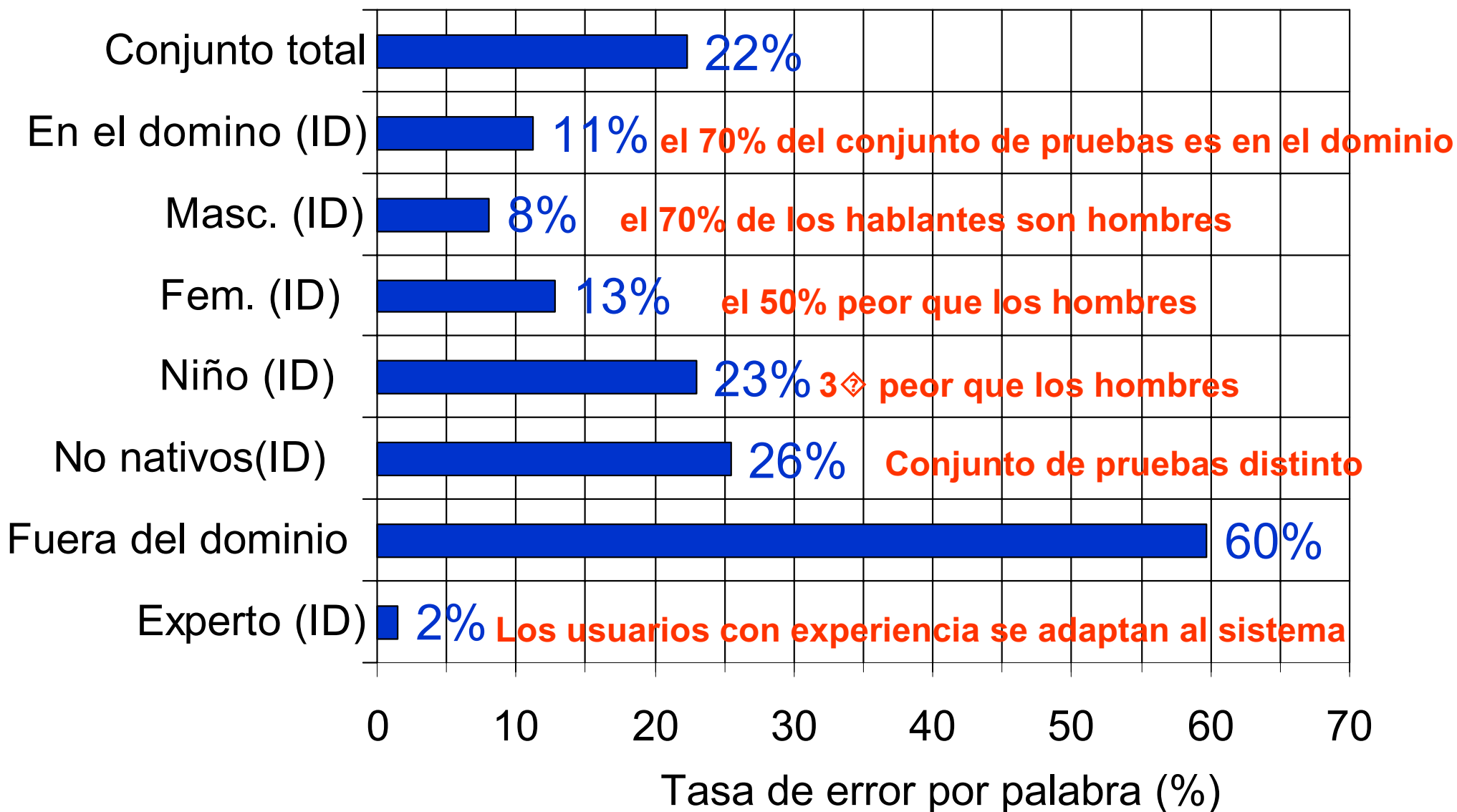
- **La búsqueda emplea pases hacia delante y hacia atrás:**
 - La búsqueda de Viterbi hacia delante utiliza un bigrama
 - La búsqueda A* hacia atrás utiliza el bigrama para crear un grafo de palabra
 - Volver a puntuar el grafo de la palabra con el trigrama (ej., sustraer las puntuaciones del bigrama)
 - La búsqueda A* emplea el trigrama para crear las salidas *N*-mejores
- **La búsqueda se basa en dos tipos de recorte:**
 - Recorte basado en la puntuación de probabilidad relativa
 - Recorte basado en el número máximo de hipótesis
 - El recorte proporciona un equilibrio entre velocidad y exactitud
- **La búsqueda puede controlar el equilibrio entre inserciones y eliminaciones**
 - Modelo de lenguaje predispuesto a favor de las oraciones cortas
 - Heurística del peso de transición de palabra (*wtw*) ajustada para eliminar el margen de error

Experimentos de reconocimiento

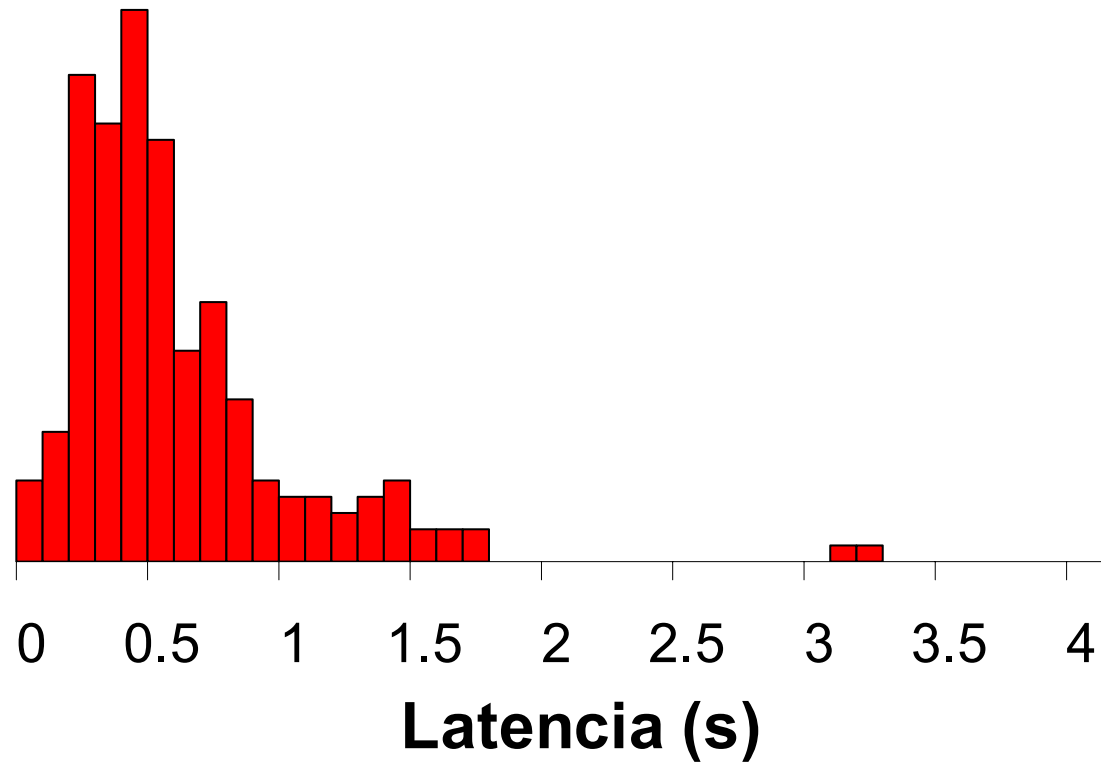


- **La recopilación de datos reales mejora el rendimiento:**
 - Permite un aumento en la complejidad y una mejora en la robustez para modelos acústicos y del lenguaje
 - Mejor opción que las condiciones de grabación en el laboratorio

Análisis del error (Conjunto de pruebas de 2506 enunciados)



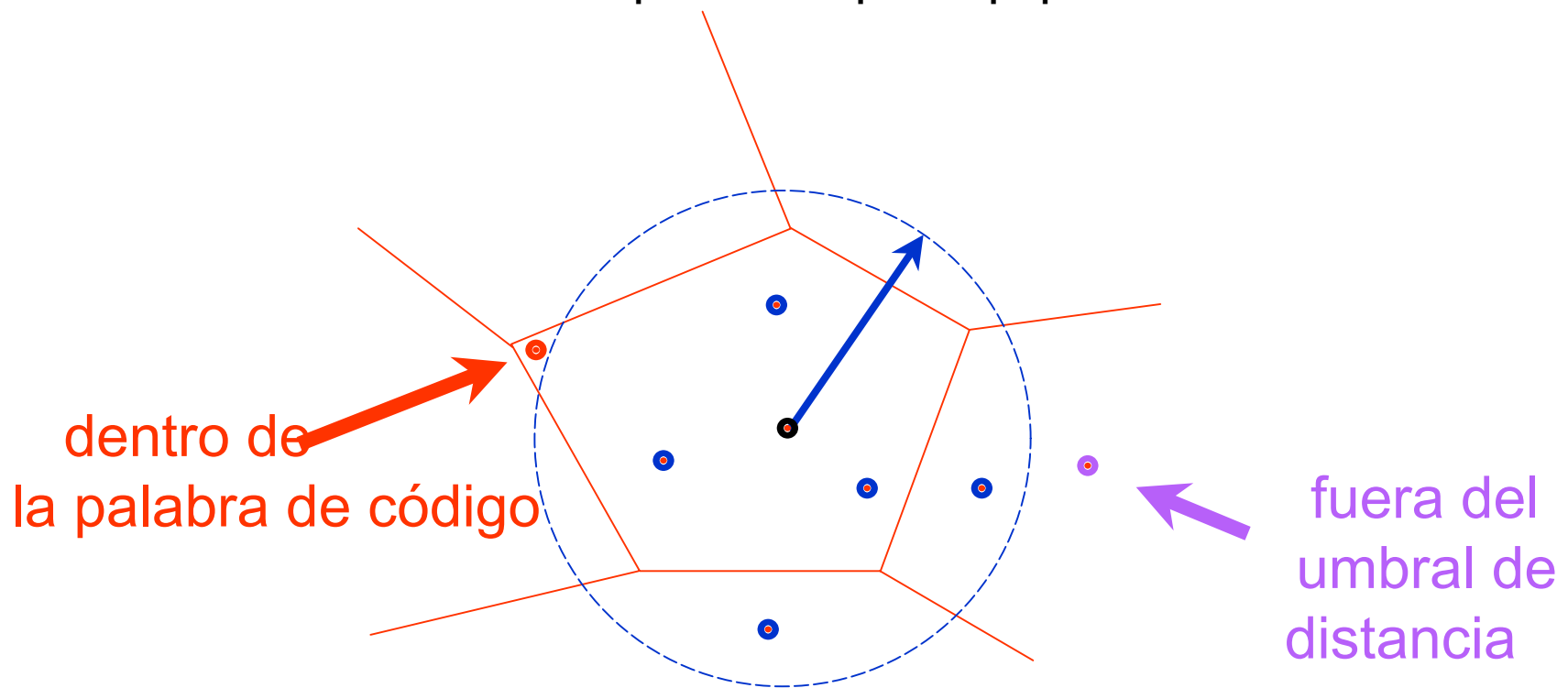
Latencia de la búsqueda A*



- Promedio de latencia ⤴ .62 segundos
- 85% < 1 segundo ; 99% < 2 segundos
- La latencia no depende de la longitud del enunciado

Selección gaussiana

- El ~ 50% de la computación total es la evaluación de las densidades gaussianas
- Puede utilizar VQ (cuantización vectorial) binaria para seleccionar los componentes mezcla de evaluación
- Criterios de selección del componente para cada palabra de código de la VQ:
 - Aquellos dentro del umbral de distancia
 - Aquellos dentro de la palabra de código (ej., cada componente utilizado al menos una vez)
 - Al menos un componente / modelo por palabra código (ej., sólo si es necesario)
- Puede reducir considerablemente la computación con pérdida pequeña de error



Experimentos de agregación

- Combinar distintas pasadas de entrenamiento puede aumentar el rendimiento
- Tres sistemas experimentales: clasificación fonética, reconocimiento fonético (TIMIT) y reconocimiento de voz (RM)
- Modelos acústicos:
 - Densidades de mezcla de gaussianas, K -medias inicializado aleatoriamente
 - 24 pruebas distintas de entrenamiento
- Medida de rendimiento medio de los únicos modelos agregados M multiplicados por N (comenzando a partir de 24 modelos separados)

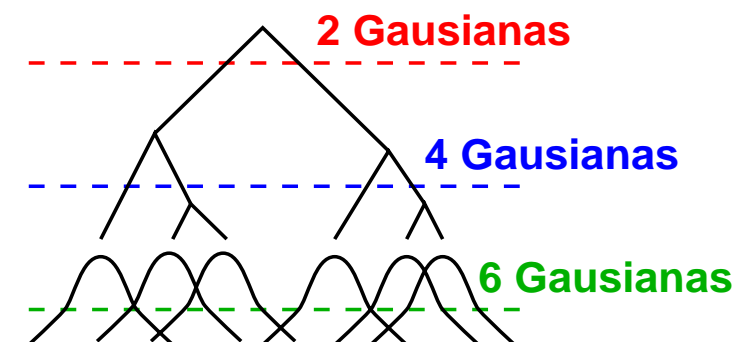
% Error	Clasificación de fono	Reconocimiento de fono	Rec. de palabra
M=24 N=1	22.1	29.3	4.5
M=6 N=4	20.7	28.4	4.2
M=1 N=24	20.2	28.1	4.0
% de reducción	8.3	4.0	12.0

Modelo de agregación

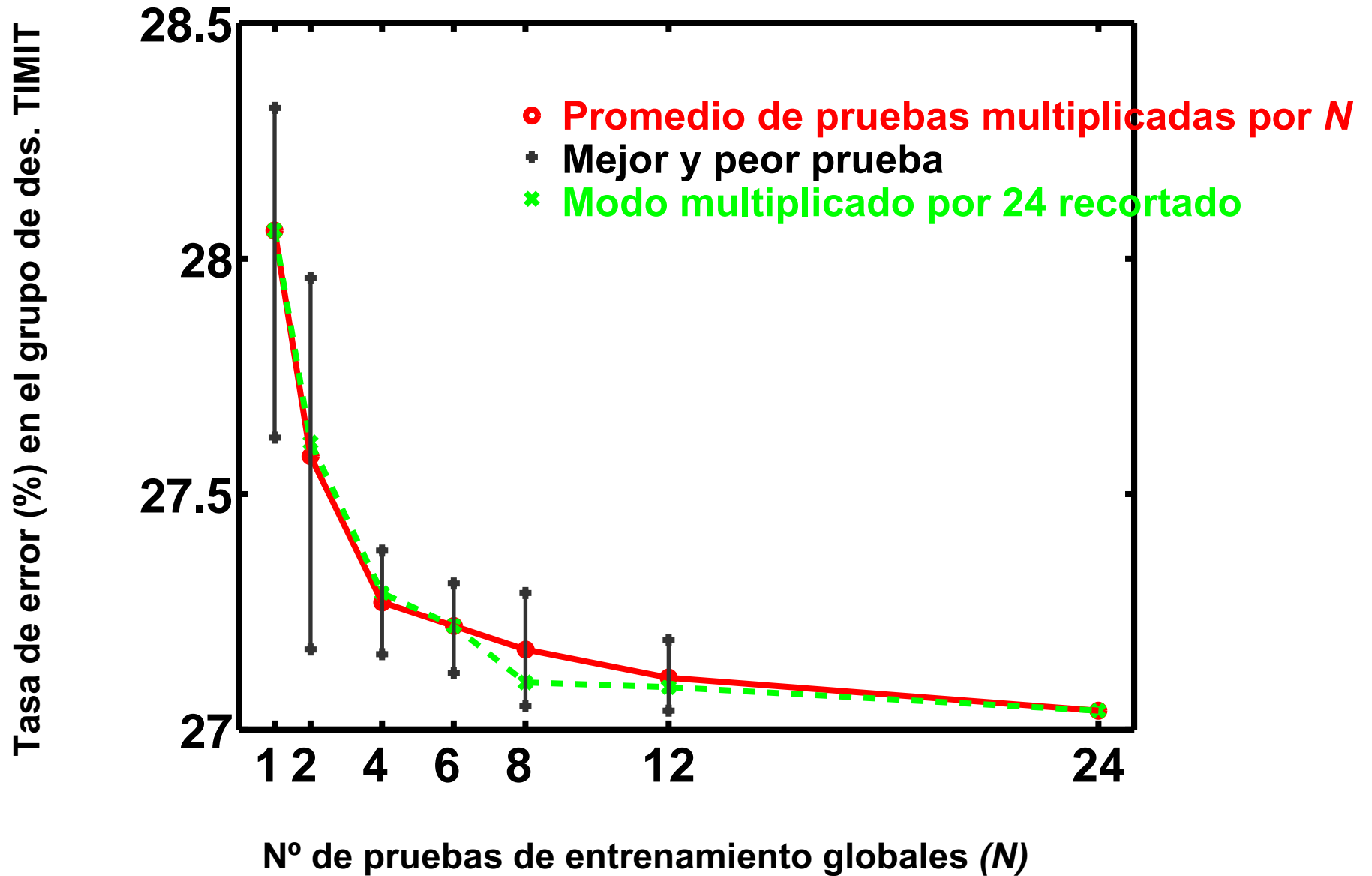
- La agregación combina clasificadores N con igual peso para formar un clasificador global

$$\varphi_A(\vec{X}) = \frac{1}{N} \sum_{n=1}^N \varphi_n(\vec{X})$$

- El error esperado de un clasificador global es menor que el error esperado de cualquier constituyente seleccionado aleatoriamente
- El clasificador global multiplicado por N posee N veces más computación
- Los kernel gaussianos del modelo global pueden ser agrupados jerárquicamente y recortados selectivamente
 - Experimento: Recortar el modelo multiplicado por 24 hasta el tamaño de modelos más pequeños multiplicados por N

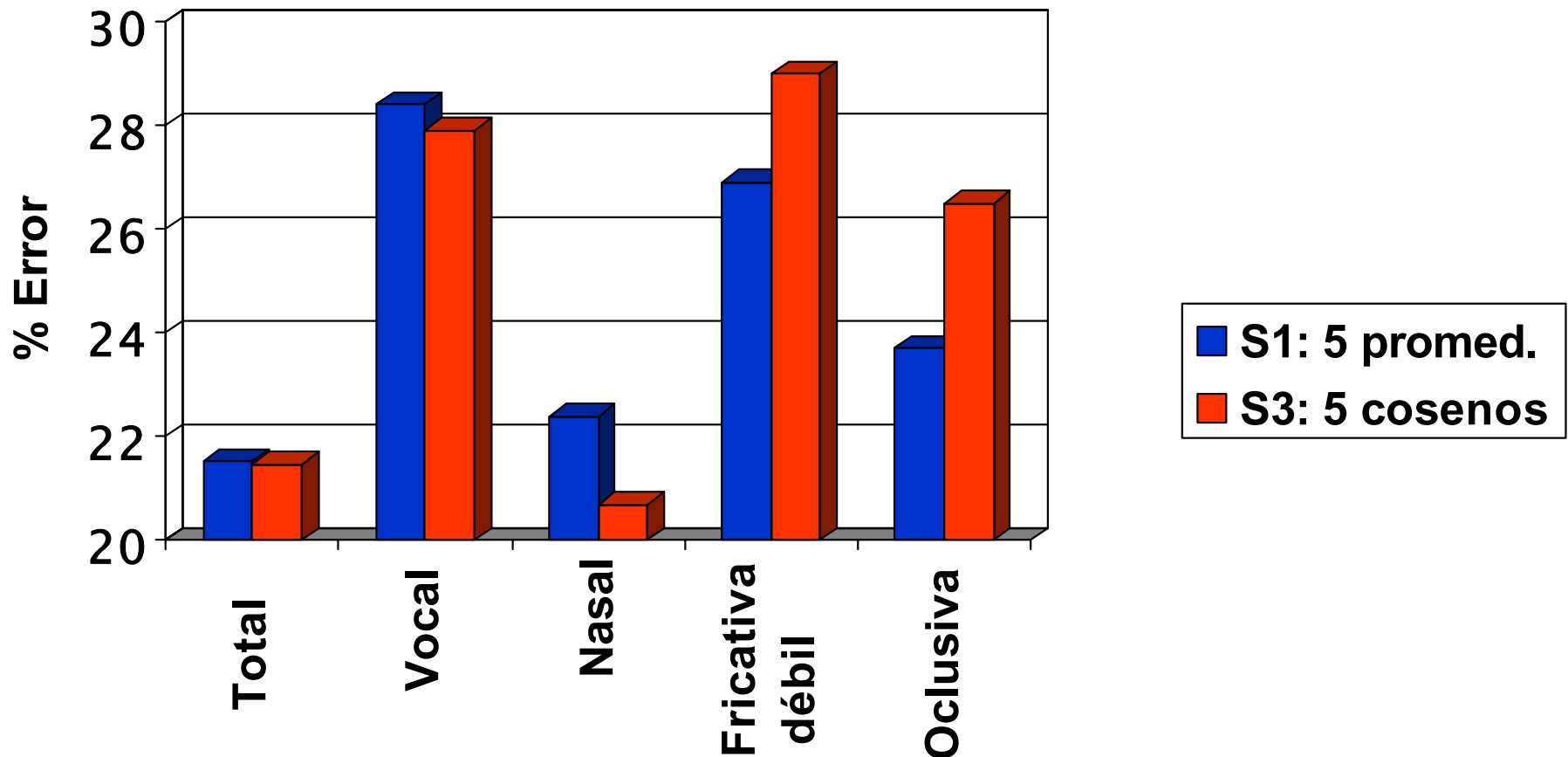


Experimentos de agregación



Clasificación basada en comité

- El cambio de base temporal afecta al error dentro de la clase
 - Base del coseno ligeramente variable mejor para vocales y nasales
 - Base constante a intervalos mejor para fricativas y oclusivas



- Combinar las fuentes de información puede reducir el error

- Utiliza vectores característicos acústicos múltiples y clasificadores para incorporar distintas fuentes de información
- 3 métodos de combinación explorados (ej., votación, lineal, indep.)
- Obtiene la clasificación fonética del estado del arte y los resultados del reconocimiento (TIMIT)
- Combinación de 3 modelos límite en el dominio atmosférico Júpiter
 - Reducción relativa del 10-16 % de la tasa de error por palabra por encima de la línea de fondo
 - Reducción relativa del 14-20 % de la tasa de error por sustitución por encima de la línea de fondo

Mediciones acústicas	% Error	% Sus.
B1 (30 ms, 12 MFCC, prom. abreviado)	11.3	6.4
B2 (30 ms, 12 MFCC+ZC+E+LFE, 4 cos \pm 50ms)	12.0	6.7
B3 (10ms, 12 MFCC, 5 cos \pm 75ms)	12.1	6.9
B1 + B2 + B3	10.1	5.5

- **Sistema ROVER desarrollado en NIST [Fiscus, 1997]**
 - Prueba Benchmark Hub-5E LVCSR en 1997
 - “La salida del reconocedor vota la reducción del error”
 - Combina la salida del reconocimiento de palabras etiquetadas confidencialmente a partir de múltiples reconocedores
 - Producción del 12.5% de reducción relativa en WER (tasa de error por palabra)
- **Noción de combinación de fuentes de información múltiples**
 - Basado en sílabas y en palabras [Wu, Morgan et al, 1998]
 - Inventarios fonéticos distintos [AT&T]
 - 80, 100 o 125 tramos por segundo [BBN]
 - Trifono y quintuple fono [HTK]
 - Reconocimiento del discurso basado en subbandas [Bourland, Dupont]

- **E. Bocchieri; Vector quantization for the efficient computation of continuous density likelihoods. *Proc. ICASSP*, 1993.**
- **T. Hazen y A. Halberstadt; Using aggregation to improve the performance of mixture Gaussian acoustic models. *Proc. ICASSP*, 1998.**
- **J. Glass, T. Hazen y L. Hetherington; Real-time telephone-based speech recognition in the Jupiter domain. *Proc. ICASSP*, 1999.**
- **A. Halberstadt; Heterogeneous acoustic measurements and multiple classifiers for speech recognition. Tesis doctoral, MIT, 1998.**
- **T. Watanabe et al.; Speech recognition using tree-structured probability density function. *Proc. ICSLP*, 1994.**