

Evaluación Albayzín 2006

Red Temática en Tecnologías del Habla

(Noviembre 2006)



CATEGORÍAS

- Reconocimiento de habla:
 - Reconocimiento de habla con voz patológica
- Síntesis de voz:
 - Mimetismo de voces con el fin de engañar a un sistema de control de acceso biométrico
- Segmentación/Identificación de voz:
 - Localizar los segmentos en los que habla un locutor e identificar al locutor
- Traducción:
 - Traducción de texto a lengua de signos



PROCEDIMIENTO

- El 21 de Julio: planes de evaluación
- El 16 de Agosto: material de entrenamiento y desarrollo.
- El 18 de Octubre de 2006: material de evaluación.
- El 27 de Octubre de 2006 a las 24:00 fue la fecha límite para recibir los resultados.
- El 8 de Noviembre de 2006 presentación de resultados.



RECONOCIMIENTO DE HABLA (I)

- Entrenamiento:
 - Para la evaluación dependiente del locutor, el material de entrenamiento contendrá al menos 2 repeticiones de cada palabra de cada locutor.
 - Para la evaluación independiente del locutor, el material de entrenamiento contendrá mas de 500 locuciones de habla no patológica y mas de 1000 de habla patológica.
- Test:
 - Dependiente del locutor, más de 40 locutores.
 - Independiente del locutor, más de 20 locutores.
 - Medida: Tasa de reconocimiento.



RECONOCIMIENTO DE HABLA (II)

- NO HAY SISTEMAS

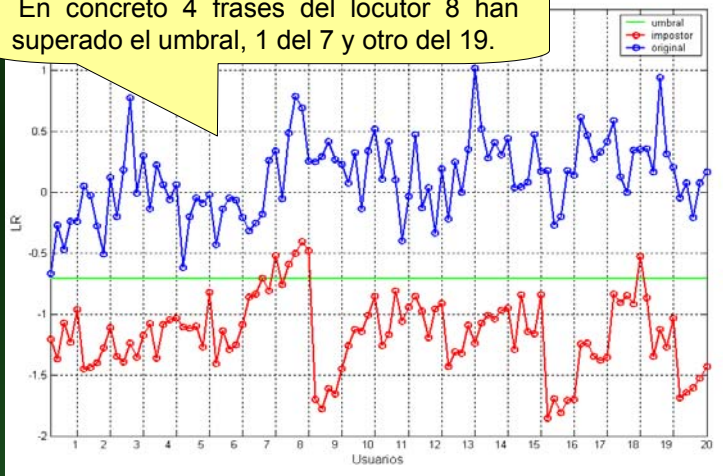


SÍNTESIS DE HABLA (I)

- Entrenamiento:
 - Un fichero de audio con una frase para cada hablante registrado en el sistema de control de accesos
 - Sistema Biométrico: Mod universales de 512 gaussianas entrenados por ML con 1640 ficheros de Albayzín. Modelos de locutor = adaptacion Bayesiana de medias, varianzas y pesos
- Test:
 - Dependiente del texto
 - Independiente del texto
 - Número de aceptaciones del sistema biométrico que se consigan

• RESULTADOS

En concreto 4 frases del locutor 8 han superado el umbral, 1 del 7 y otro del 19.

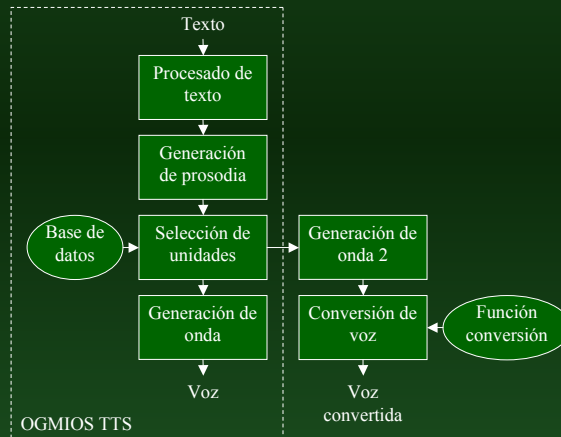


• RESULTADOS

- Valor medio de la puntuación de los locutores impostores utilizando la base de datos Albayzín = **-1.73**
- Valor medio de la puntuación de los locutores para las señales sintetizadas = **-1.16**
- La puntuación media para la voz de los usuarios registrados es de **0.12**

- SISTEMA OGMIOS Jr (Daniel, UPC):

- Descripción del sistema



- SISTEMA OGMIOS Jr (UPC):

- Generación de onda

- Armónicos (f_0 , A , φ) + ruido (LPC)
- Tramas de longitud constante
- Métodos de manipulación prosódica y concatenación de unidades

- Conversión de voz

- Normalización de $\log-f_0$
- Conversión de tracto vocal (14 coef. LSF) por medio de GMM (8 gaussianas)
- Estimación de envolvente de fase a partir del tracto convertido
- Predicción de componente ruidoso a partir del tracto convertido, sólo en tramas sonoras



SÍNTESIS DE HABLA (VI)

- SISTEMA OGMIOS Jr (UPC):
 - Método de trabajo
 - Imitación de las 5 frases de los 20 hablantes objetivo con el sintetizador: corpus “paralelo” de entrenamiento
 - Alineamiento por medio de segmentación HMM + corrección manual (muy pocos datos!!)
 - Entrenamiento de las funciones de conversión
 - Generación de las frases finales con el sintetizador, y aplicación de las funciones de conversión acústica



SEGMENTACIÓN (I)

- Entrenamiento:
 - 5 ficheros de 5 locutores en canales diferentes.
- Test:
 - 20 ficheros de varios minutos con intervenciones de los 5 locutores.
 - Diarization Error (DER): porcentaje de habla mal asignada a un locutor respecto el tiempo total de voz.
 - Diarization Error Modified (DERM): DER considerando la identificación correcta del locutor etiquetado.



SEGMENTACIÓN (II)

- RESULTADOS

SISTEMA	DER	DERM
EHU	15.20	19.50
UPC1	17.44	17.44
UPC2	18.03	18.03



SEGMENTACIÓN (III)

- SISTEMA Sautrela (Mikel, EHU):

- Condiciones de evaluación

- Entrenamiento de modelos de locutor:
- Ajuste de los parámetros de segmentación/identificación:

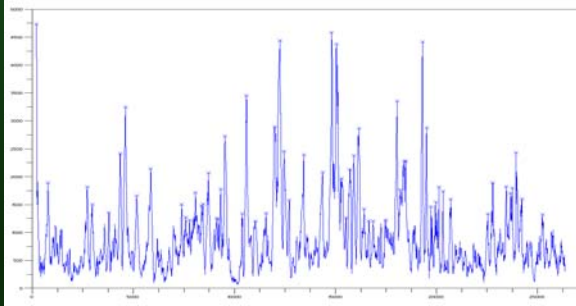
LOC1 (17.35sg)
LOC2 (12.51sg)
LOC3 (13.12sg)
LOC4 (11.17sg)
LOC5 (10.04sg)

fragmento09
(265.65sg)

- Aproximación algorítmica

- Parametrización: 12 cepstrales
- Segmentación:
 - Ventana deslizante dividida en dos partes iguales
 - Modelo con una única gaussiana: $N(izq)$, $N(der)$
 - Se localizan los instantes de máxima *disimilitud*:

$$d(izq, der) = \logprob(izq | N(izq)) - \logprob(izq | N(der)) + \logprob(der | N(der)) - \logprob(der | N(izq))$$



Es máximo todo aquel punto que supere un umbral y no tenga un valor superior en su entorno.

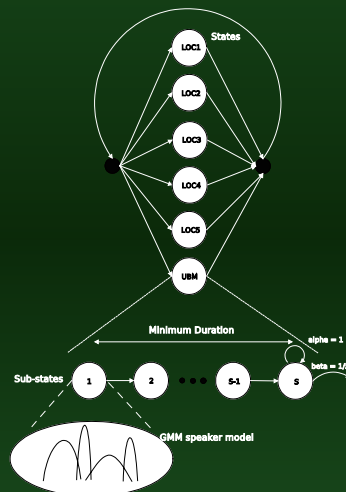
Figura 1. Deslizando la ventana, se obtiene una función de disimilitud

• Identificación

- Un modelo de una sola gaussiana por locutor: $N(LOC_i)$
- Un modelo de una sola gaussiana dado el segmento: $N(seg)$
- Locutor de máxima verosimilitud:
- Función de etiquetado: $LOC_{max} = \arg \max_{LOC_i} \{prob(seg | N(LOC_i))\}$

$$eti(seg) = \begin{cases} LOC_{max} & , \logprob(seg | N(LOC_{max})) \geq \logprob(seg | N(seg)) - \text{margen} \\ OTROS & , \logprob(seg | N(LOC_{max})) < \logprob(seg | N(seg)) - \text{margen} \end{cases}$$

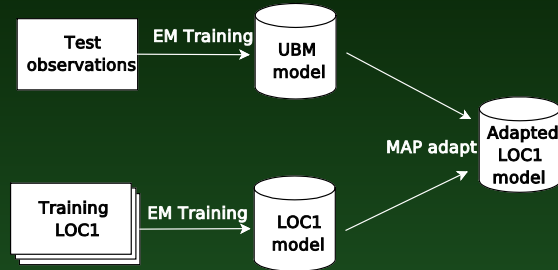
- SISTEMA TALP (Jordi, UPC):
 - Ergodic HMM with 6 states, each one corresponds to each “LOC” and to “OTROS”
 - Each state contains a set of S sub-states imposing a minimum duration
 - Each sub-state has a pdf modeled by a GMM model of size 64 Gauss, tied across all sub-states in a speaker





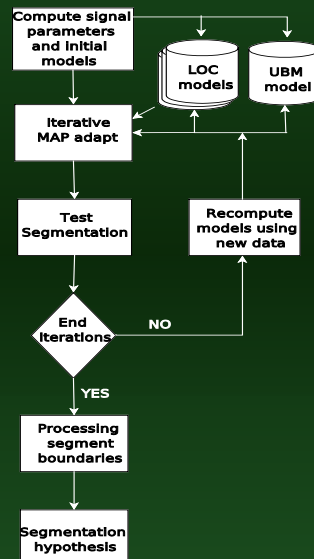
SEGMENTACIÓN (VI)

- The signal spectrum is estimated with 12 MFCC obtained every 10 ms
- Each speaker model is MAP adapted (mean + weight) from a UBM model
- UBM model is EM trained with all the show data



SEGMENTACIÓN (VII)

- Iterative segmentation to find new data and MAP re-adaptation of the models
- LLR at frame-level using UBM and threshold tuned with development data
- Gaussian pruning
- Frame purification
- Post-processing of the short-segment boundaries





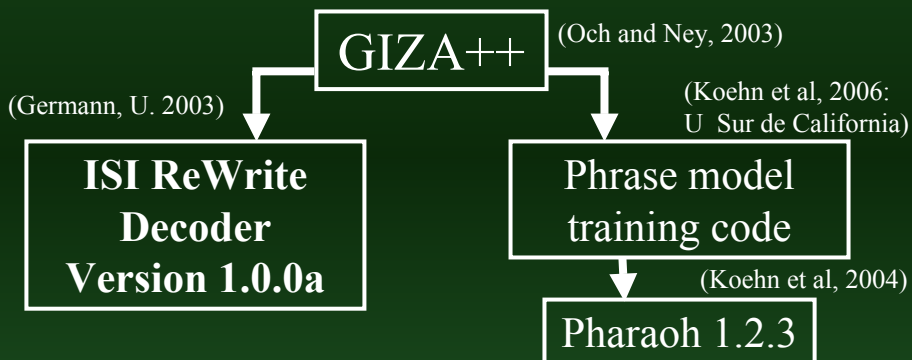
TRADUCCIÓN (I)

- Entrenamiento:
 - 200 frases con su traducción en Lengua de Signos (dominio: petición o renovación del DNI).
- Test:
 - 81 frases de texto
 - WER, PIWER, BLEU, NIST.




TRADUCCIÓN (II)

- SISTEMA REF: sistemas de referencia



- RESULTADOS

SISTEMA	WER	PIWER	BLEU	NIST
 UPV-EHU	28.90	20.93	0.62	7.06
UPV-UCLM	35.33	23.79	0.54	6.48
Phrase Model	37.09	27.97	0.50	6.12
ReWrite Dec.	53.59	37.67	0.21	5.23

- SISTEMA 1 (Alicia, UPV-EHU): transductor de estados finitos

Sean:

- s cadena en la lengua origen
- Δ vocabulario en la lengua destino

La traducción esperada:

$$\hat{t} = \arg \max_{\forall t \in \Delta^*} P(s, t)$$

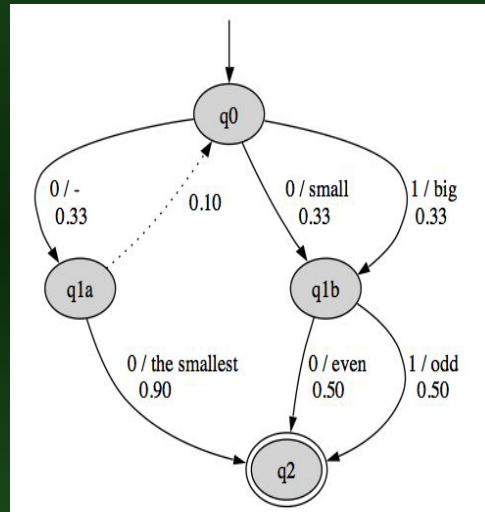
APRENDIZAJE

en base a un conjunto de ejemplos bilingües

0 0 - the smallest

0 1 - small odd

1 0 - big even



BÚSQUEDA

- Generaliza eventos no-vistos en los ejemplos:

1 1 - big odd

- En caso de ambigüedad, la cadena más probable.

0 0 - the smallest - $P \sim 0.30$

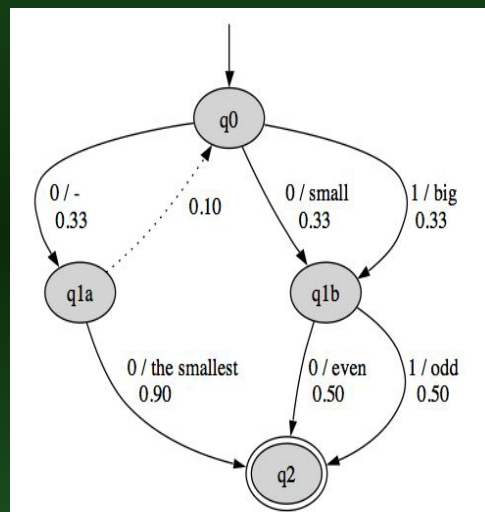
0 0 - small even - $P \sim 0.17$

- Palabras desconocidas: explora todo y selecciona la más probable.

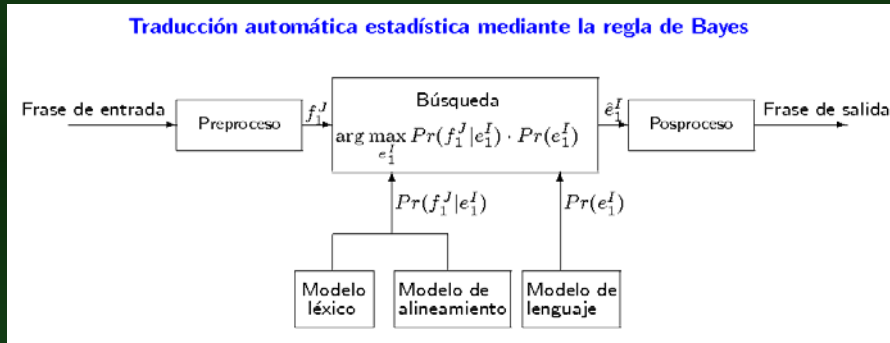
A 1 - big - $P \sim 0.01$

A 1 - big odd - $P \sim 0.17$

A 1 - small odd - $P \sim 0.17$



- SISTEMA 2 (Daniel, UPV-UCLM):



Modelos de secuencias de palabras

- Historia generativa:
 1. La frase e_1^I se divide en K secuencias (\tilde{e}_k^K)
 2. Cada secuencia \tilde{e}_k se traduce por una secuencia \tilde{f}
 3. Finalmente las secuencias \tilde{f} se reordenan para componer la frase $\tilde{f}_1^K = f_1^J$
- Todas las decisiones que se van haciendo pueden resumirse en la variable oculta $\tilde{a} = \tilde{a}_1^K$

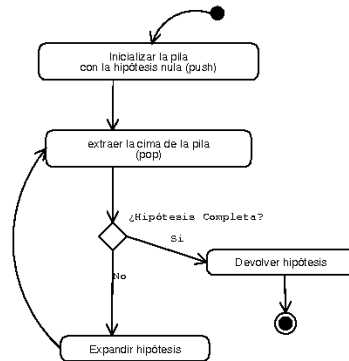
$$Pr(\tilde{f}_1^J | \tilde{e}_1^I) = \sum_{\tilde{a}} Pr(\tilde{a}, \tilde{f}_1^J | \tilde{e}_1^I) = \sum_{\tilde{a}} Pr(\tilde{a} | \tilde{e}_1^I) Pr(\tilde{f}_1^J | \tilde{a}, \tilde{e}_1^I)$$

- Típicamente, se hacen una serie de asunciones que convierten los modelos en diccionarios estadísticos de secuencias:

$$Pr(f_1^J | e_1^I) = \alpha(e_1^I) \sum_{\tilde{a}} \prod_{k=1}^K p(\tilde{f}_k | \tilde{e}_{a_k})$$

Algoritmos de pila

- Los algoritmos de pila ejecutan un proceso iterativo:



- Las hipótesis se almacenan en orden ascendente según su probabilidad
- Expandir una hipótesis consiste en aplicarle un conjunto de operadores

- Síntesis (mimetización de voces):
 - Propuesta de un sistema biométrico como medidor de calidad en mimetización de voces.
- Segmentación/Identificación:
 - Base de datos + Problema (seg./iden.).
 - Evaluación y descripción de Técnicas.
- Traducción:
 - Base de datos.
 - Evaluación y descripción de Técnicas.



AGRADECIMIENTOS

**!!!!!! MUCHAS GRACIAS A TODOS
LOS PARTICIPANTES Y A LOS
COLABORADORES !!!!!!**