

INFORME FINAL DE LA EVALUACIÓN PARA SEGMENTACIÓN E IDENTIFICACIÓN DE LOCUTORES

1.- Introducción

El presente Plan de Evaluación para Segmentación e Identificación de Hablantes proporciona las bases para la evaluación de aplicaciones en este campo de Tecnologías del Habla.

El objetivo de esta evaluación es fomentar los trabajos de investigación en relación con la segmentación de la voz en conversaciones con varios locutores, y además, la identificación parcial de algunos de éstos. Con esta finalidad se plantea un plan de evaluación que consiste en la segmentación e identificación de las intervenciones de hasta 5 locutores conocidos, en archivos donde podrán existir intervenciones de otros locutores. Ambos aspectos, segmentación e identificación, son importantes: una buena segmentación pero una mala identificación será considerada como una segmentación errónea. Durante la fase de entrenamiento o desarrollo se ofrecerán grabaciones cortas correspondientes a los 5 locutores a segmentar e identificar.

Los participantes se comprometen a presentar los resultados de la evaluación en una sesión especial que tendrá lugar durante las IV Jornadas en Tecnología del Habla. La participación se realiza a modo individual o en equipo formado por un máximo de 3 investigadores donde al menos uno de ellos debe ser estudiante.

2.- Medición de prestaciones

Para la evaluación se utilizarán la herramienta del NIST “md-eval-v21.pl” (The diarization evaluation tool) disponible en <http://www.nist.gov/speech/tests/rt/rt2006/spring/>¹

El modo de funcionamiento de esta herramienta se ha tomado como referencia para la definición del formato de los archivos de resultados (formato RTTM) que se presenta en el punto 4.

3.- Condiciones de evaluación

El material de entrenamiento consistirá en 5 archivos de entre 1 y 5 segundos, cada uno de ellos con la voz de un locutor.

¹ Gracias a Xavier Anguera por las gestiones para disponer de esta herramienta de evaluación

Todos los archivos contendrán muestras digitalizadas de audio con el siguiente formato:

- Frecuencia de muestreo: 16 KHz
- Canales de audio: mono
- Tamaño de muestra: 16 bits
- Alineamiento de octeto: LSB primero (little endian).
- Ficheros PCM sin cabecera ni compresión.

Los archivos de audio tendrán la extensión **.pcm**.

4.- Formato de los archivos de etiquetas a generar

Los resultados de la segmentación se deberán presentar en archivos de texto (hipótesis), con el mismo nombre correspondiente al archivo de audio de partida, pero con extensión **.hyp**, y cuyas líneas contendrán la información temporal del punto de comienzo y duración de cada locutor, y una mínima identificación de éste. Un ejemplo, para un archivo *fichero21.pcm*, es:

Archivo: *fichero21.hyp*

```
SPEAKER FICH21 1 0.00 257.93 <NA> <NA> OTROS <NA>
SPEAKER FICH21 1 257.93 12.15 <NA> <NA> LOC2 <NA>
SPEAKER FICH21 1 270.40 25.60 <NA> <NA> OTROS <NA>
...
```

Cada intervención de un locutor se muestra en una línea diferente que contiene los siguientes campos:

- Modo de obtención de la señal: en todos los casos consideraremos este campo igual a SPEAKER.
- Número del fichero a considerar: FICH1, FICH2, ... hasta FICH20
- Canal: en todos los casos consideraremos el canal 1.
- Inicio: representa el segundo y centésima de segundo de comienzo del locutor.
- Duración: representa el segundo y centésima de segundo de la duración de la intervención del locutor.
- Campo 6: <NA>
- Campo 7: <NA>
- Código del locutor. Los códigos posibles para identificar al locutor con los siguientes: LOC1, LOC2, LOC3, LOC4, LOC5 y OTROS. La etiqueta OTROS corresponde con cualquier otro locutor/música/... que no corresponda a alguno de los 5 locutores conocidos.
- Campo 9: <NA>

La representación numérica del tiempo será en segundos y centésimas, pudiendo omitirse los ceros iniciales y finales, así como el punto decimal según los convenios habituales. El carácter decimal deberá ser el '.'. Las cantidades deberán estar separadas por espacios.

Estas hipótesis se compararán mediante la herramienta mencionada con una segmentación de referencia, contenida en archivos con el mismo formato y nombrados con la extensión **.ref**.

Un aspecto que conviene remarcar es no se han etiquetado las pausas breves (menores de 500 ms) entre dos intervenciones de un mismo locutor, considerándose como un mismo fragmento asociado a dicho locutor.

5.- Datos de evaluación

El material de prueba consistirá en 20 ficheros de entre 3 y 5 minutos con intervenciones de estos 5 locutores y/o de otros locutores.

El formato de estos archivos de audio será el mismo que el de los archivos de entrenamiento y se nombrarán con la misma extensión (**.pcm**).

6.- Procedimiento para la evaluación

El procedimiento con las fechas para la evaluación es el siguiente:

- El 21 de Julio de 2006 se dispondrá de los planes de evaluación y se abre el periodo de inscripción.
- La fecha límite de inscripción será el 15 de Septiembre de 2006.
- A partir del 16 de Agosto de 2006 se podrá disponer del material de entrenamiento y desarrollo para las distintas evaluaciones. Es necesario estar inscrito en la evaluación para recibir el material.
- El 18 de Octubre de 2006 se liberarán las bases de datos para la evaluación.
- El 27 de Octubre de 2006 a las 24:00 es la fecha límite para recibir los resultados en el formato y método indicados.
- El 3 de Noviembre de 2006 se enviarán los resultados de la evaluación.

7.- Envío de segmentaciones

Las segmentaciones se enviarán por correo electrónico a la organización. Deberán ser completas, conteniendo por tanto todo el conjunto de datos de evaluación.

Los ficheros de texto con las segmentaciones, con la extensión **“.hyp”**, deberán enviarse comprimidas en un archivo ***.zip** a: **Rubén San-Segundo Hernández** (lapiz@die.upm.es)

Los resultados estarán disponibles una vez se hayan enviado dichos resultados a los participantes. Esto permitirá realizar análisis previos a la celebración de las IV Jornadas de Tecnologías del Habla.

Cada participante deberá remitir una descripción del sistema enviado a la evaluación, que deberá incluir:

- Nombre del sistema
- Condiciones de evaluación (base de datos de entrenamiento)
- Descripción de la aproximación algorítmica

Esta descripción se enviará en formato de texto ASCII o PDF. Las descripciones recibidas se distribuirán como parte del material de análisis de la evaluación.

8.- Evaluación de sistemas

Para la evaluación de los diferentes sistemas se computaron las siguientes medidas:

- **Diarization Error (DER):** porcentaje de habla mal asignada a un locutor respecto el tiempo total de voz.
- **Diarization Error Modified (DERM):** porcentaje de habla mal asignada a un locutor respecto el tiempo total de voz considerando la **identificación** correcta del locutor etiquetado (la diferencia entre estas dos medidas aparece cuando en algún fichero la asignación de los locutores es errónea respecto las etiquetas). Esta medida primero realiza una asignación de etiquetas a los locutores para garantizar el mejor alineamiento (igual al DER) y después comprueba que las etiquetas asignadas y las de referencia sean las mismas.

8.1 Sistema SAUTRELA por EHU

Contacto: Mikel Peñagarikano, email: mpenagar@we.lc.ehu.es,
Tel: 946015310

El sistema presentado por EHU consiste en dos fases:

- **Segmentación:** Se toma una ventana temporal que se divide en dos partes iguales, siendo cada mitad utilizada para generar un modelo de una única gaussiana (N (izq) y N (der)). A partir de ambos modelos se calcula lo que denominaremos disimilitud entre ambos segmentos de señal. Para localizar dichos máximos se utiliza un simple criterio: es máximo todo aquel punto que supere un umbral y no tenga un valor superior en su entorno. El tamaño de ventana, desplazamiento, umbral y entorno han sido escogidos tratando de optimizar la segmentación de la señal de referencia.
- **Identificación:** Tomando las señales de cada locutor, se ha generado para cada uno de ellos un modelo de una sola gaussiana, N (LOCi). Igualmente, se genera un modelo de una sola gaussiana con el propio segmento a identificar. La identificación se realiza eligiendo el locutor que mayor parecido ofrezca. En el caso de que este parecido no supere un umbral (ajustado con el fichero de ejemplo) se etiqueta como OTROS.

8.2 Sistema TALP

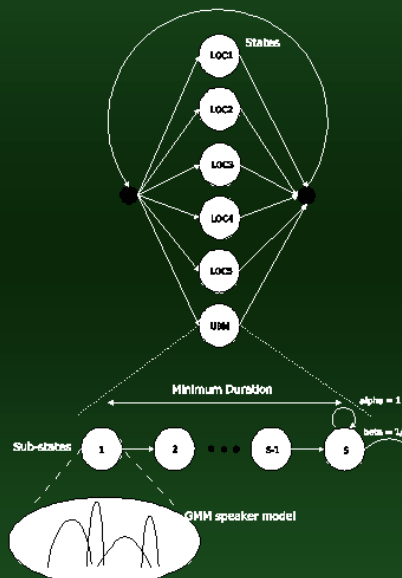
Contacto: Jorge Luque, email: aluque@gps.tsc.upc.edu,
Tel: 934010994.



SEGMENTACIÓN (III)

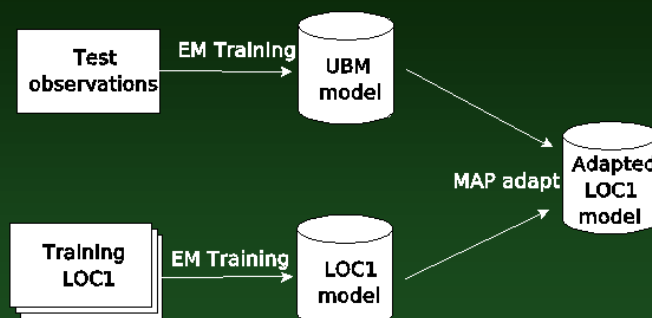
- SISTEMA TALP (UPC):

- 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 (IV)

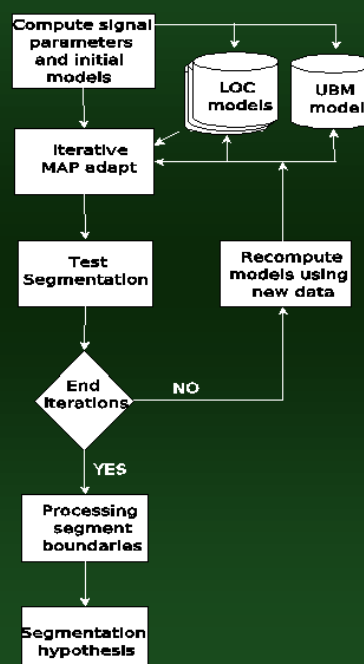
- The signal spectrum is estimated with 12 MFCC obtained every 10 ms
- Each speaker model is MAP adapted (mean + weight) from a UBM model
- UBM model is EM trained with all the show data





SEGMENTACIÓN (V)

- 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



8.4 Resultados finales

Los principales resultados de los sistemas comentados son los siguientes

	DER	DERM
EHU (SAUTRELA)	15.20	19.50
TALP (UPC1)	17.44	17.44
TALP (UPC2)	18.03	18.03

Como se puede observar el sistema de la EHU (SAUTRELA) presenta un muy buen error de Diarization (DER) pero empeora cuando se obliga a que la identificación del locutor sea la correcta (DERM). Aunque los sistemas de la UPC tienen un mayor DER, el hecho de obligar a que la identificación del locutor sea correcta hace que el sistema UPC1 se comporte mejor.

Un aspecto que conviene remarcar en esta evaluación son las diferencias principales con la evaluación propuesta desde NIST en la modalidad de Speaker Diarization. Estas diferencias son las siguientes:

- En primer lugar, es obligatorio la identificación correcta del locutor para que se considere como un acierto.
- En segundo lugar, en esta evaluación el número de locutores es fijo (LOC1-LOC5+OTROS, 6 en total) y conocido, mientras que en la

evaluación de NIST el número de locutores es indefinido, incrementando sensiblemente la dificultad de la tarea y la complejidad de los sistemas desarrollados.

Una vez presentados los resultados, desde el equipo SAUTRELA (Mikel y LuisJa) propusieron una modificación de la métrica DERM a la que denominaremos DERM2 y que se define a continuación:

- **Diarization Error Modified 2 (DERM2):** la diferencia principal con la medida DERM es que en lugar de comparar las etiquetas obtenidas del mejor alineamiento con las etiquetas referencia, se comparan las etiquetas reales de los ficheros de hipótesis con las etiquetas de referencia.

La medida DERM penaliza en exceso el error en la identificación de un locutor puesto que las etiquetas de referencia se comparan con las etiquetas asignadas en el mejor alineamiento obtenido, y no con las etiquetas de los ficheros de hipótesis. En este caso pueden ocurrir situaciones en las que tengamos una asignación de etiquetas que cometa algunos errores que no aparecen en los ficheros de hipótesis. Observada esta diferencia se propone esta segunda medida para futuras evaluaciones en segmentación/identificación de locutores.

Los resultados considerando todas las medidas son:

	DER	DERM	DERM2
EHU (SAUTRELA)	15.20	19.50	17.25
TALP (UPC1)	17.44	17.44	17.44
TALP (UPC2)	18.03	18.03	18.03

En este los resultados del sistema SATÚRELA mejoran sensiblemente igualándose al mejor sistema presentado por UPC.