INFORME FINAL DE LA EVALUACIÓN EN SISTESIS DE VOZ

1.- Introducción

El propósito de esta evaluación es el fomento de la investigación en sistemas de síntesis de voz. En esta primera evaluación se quiere evaluar la capacidad de desarrollar sistemas de síntesis que sean capaces de mimetizar la voz de hablantes. Para ello se plantea un plan de evaluación que consiste en el ataque a un sistema de control de accesos con un sistema biométrico de verificación del hablante. Para ello se proporcionará para cada hablante registrado en el sistema de control de accesos un fichero de audio con una frase. La evaluación consistirá en generar de forma sintética una serie de ficheros de audio por hablante con el texto indicado para cada uno de ellos.

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 donde representante deberá ser estudiante.

2.- Medición de prestaciones

Para medir la capacidad de mimetismo de la voz, las frases sintetizadas se aplicarán a un sistema estándar de verificación de locutor basado en Modelos de Mezclas de Gaussianas independiente del texto. Las prestaciones se medirán por el número de aceptaciones que realice el sistema.

3.- Condiciones de evaluación

Para la evaluación se definirán un conjunto de entre 20 y 40 hablantes de los que se proporcionará entre 5 y 10 segundos de voz para la fase de aprendizaje.

Hasta que se abra el periodo de evaluación, los participantes podrán enviar hasta 5 frases sintetizadas de cada uno de los hablantes de aprendizaje para conocer los resultados del ataque al sistema de control de accesos.

Para la fase de evaluación se enviarán de 5 a 10 segundos de segmentos de voz de alrededor de 30 hablantes nuevos. Para cada uno de ellos se definirán las frases que se deben sintetizar, que serán como mínimo del orden de 10 frases y se podrán sintetizar hasta 5 repeticiones de la misma frase por locutor.

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**. y el nombre hará referencia al hablante, a la frase y a la repetición, con el siguiente formato:
loc_#_frase_#_rep_#.pcm

Por ejemplo, para el locator 4, frase 3 y repetición 5, el nombre del fichero será:
Loc_4_frase_3_rep_5.pcm

4.- 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

Los resultados se enviarán por correo electrónico y ftp a la organización. Deberán ser completas, conteniendo por tanto todo el conjunto de datos de evaluación.

Se definirá un procedimiento a través de un servidor ftp para el envío de los ficheros de audio. Para cada equipo se definirá un directorio que se identificara por el acrónimo del equipo y si se evalúan varios sistemas, se definirán subdirectorios para cada sistema.
Para cualquier duda consultar con **Eduardo Lleida Solano** (lleida@unizar.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 con la misma plantilla utilizada para las comunicaciones regulares de las IV Jornadas en Tecnología del Habla. Las descripciones recibidas se distribuirán como parte del material de análisis de la evaluación.

8.- Evaluación de sistemas

Finalmente únicamente se ha presentado un sistema de síntesis a la evaluación por el grupo: **UPC_Grupo_de_Síntesis**
Persona de contacto: Daniel Erro
Email: derro@gps.tsc.upc.es
Telf: 93 401 09 65

Al solo participar un sistema, no se ha podido realizar una comparativa, por ello en este informe simplemente indicamos los resultados obtenidos.

Para la evaluación se han utilizado 20 locutores de los cuales se han proporcionado 5 frases originales y se han sintetizado 5 frases nuevas para cada locutor. Las frases sintetizadas han sido las mismas para los 20 locutores y son las siguientes:

1. “el usuario tendrá a su disposición todas las propuestas de esta institución"
2. "cero tres ocho nueve cuatro uno siete"
3. "la participación en la propuesta de evaluación de sistemas albayzin cero seis"
4. "mi voz es mi contraseña seis cinco dos"
5. "acceso a la base de datos de revistas de información general"

Las señales originales se corresponden con 20 locutores de la base de datos Albayzin. Se ha utilizado la base de datos Albayzin para simular el sistema de control de accesos. La evaluación se ha realizado directamente utilizando el ratio entre la puntuación del modelo del locutor y el modelo universal.

Para registrar a los usuarios del sistema, se usaron modelos universales de hombre y mujer de 512 gaussianas entrenados por ML utilizando 1640 ficheros de todos los locutores de Albayzin. Los modelos de locutor se obtuvieron por adaptacion Bayesiana de medias, varianzas y pesos.

Se utilizaron 20 ficheros para entrenar el modelo y 5 para testear. La simulación con toda la base de datos testeando todos los archivos con todos los locutores da un eer=0.12 % con un umbral thr=-0.71.

Utilizando este umbral para todos los usuarios, en la figura 1 se representa la puntuación para cada uno de las 5 frases de cada locutor (original e impostor).

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

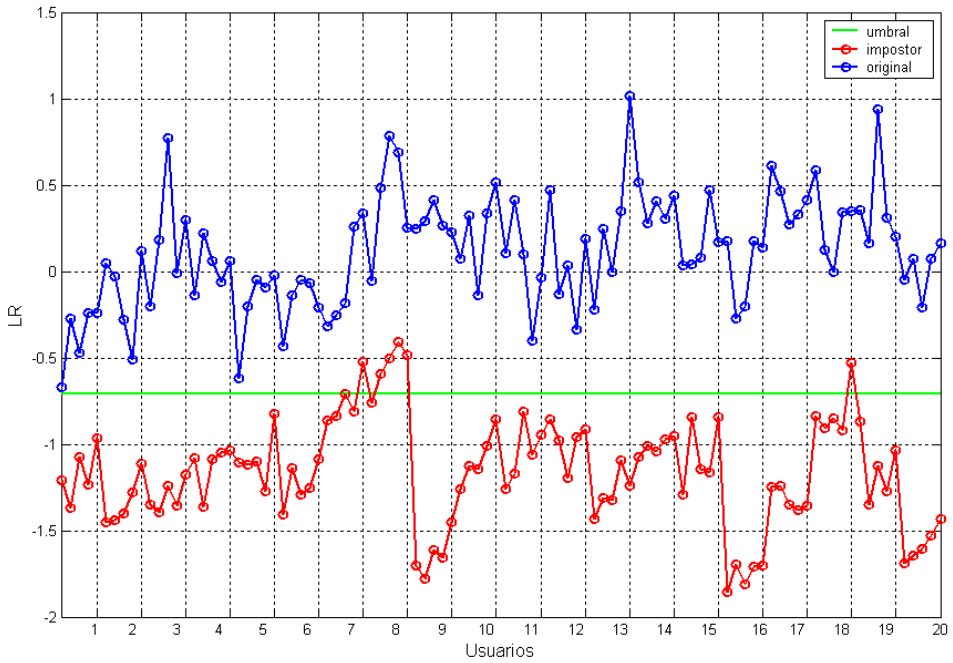

Figura 1. representación de las puntuaciones de las señales originales y de las sintetizadas (impostor). Hay 5 frases por locutor. Los 10 primeros son mujeres y los 10 últimos hombres.

Indicar que el valor medio de la puntuación de los locutores impostores utilizando la base de datos Albayzin es de –1.73 y para las señales sintetizadas es de –1.16, lo que al menos indica una tendencia de mejora de cara a impostar a un locutor. La puntuación media para la voz de los usuarios registrados es de 0.12.