Plan de Evaluación de Sistemas ALBAYZIN-08 Verificación de la Lengua (ALBAYZIN-08 VL)

1 Objetivo

El objetivo de esta evaluación es promover el intercambio de ideas, estimular la creatividad y favorecer la colaboración entre los grupos de investigación españoles que trabajan en identificación y verificación de la lengua. Con este fin, se propone una competición entre sistemas en una tarea de verificación de la lengua similar a la que cada dos años organiza la organización estadounidense NIST (*National Institute of Standards and Technology*) a nivel internacional[1], aunque con un menor grado de dificultad, ya que se consideran tan sólo cuatro lenguas (castellano, catalán, euskera y gallego) y se utilizan señales de mejor calidad (ancho de banda y SNR mayores).

2 Tarea

Dado un segmento de señal S y una cierta lengua L_i, la tarea consiste en decidir si el habla contenida en S corresponde a la lengua L_i (lo que se conoce como *prueba de verificación*). Esta decisión deberá basarse únicamente en el análisis automático de la señal de voz, mediante los modelos acústicos, fonéticos, fonológicos, léxicos, sintácticos, etc. que se consideren pertinentes. Por otra parte, para cada segmento S se tomarán N decisiones de verificación, una por lengua, suponiendo que hay N lenguas de interés. Se trata, por tanto, de N tareas de verificación. El rendimiento global del sistema se obtendrá, como se verá más adelante, agregando los resultados obtenidos para esas N tareas en una única función de coste.

Las cuatro lenguas consideradas en esta evaluación se utilizan alternativamente como lenguas objetivo y como lenguas de contraste. Llamamos lengua objetivo a la lengua que se desea detectar en una serie de pruebas de verificación. Lenguas de contraste son todas las que sabemos pueden aparecer en el conjunto de evaluación, es decir, todas las que pueden proponerse como hipótesis alternativa a la lengua objetivo. Así, en una serie de pruebas de verificación el catalán podría actuar como lengua objetivo, tomando castellano, euskera y gallego como lenguas de contraste; en otra serie, el gallego podría ser la lengua objetivo, mientras que castellano, catalán y euskera actuarían como lenguas de contraste; etc.

1 Véase http://www.nist.gov/speech/tests/lre/

3 Evaluación

Para evaluar el rendimiento de un sistema se le someterá a pruebas de verificación sobre un conjunto de evaluación. Cada segmento del conjunto de evaluación se utilizará en una prueba de verificación por cada lengua objetivo. Por tanto, en un cierto segmento podrían detectarse una, dos o más (o ninguna) de las lenguas objetivo. Sin embargo, cada segmento del conjunto de evaluación contendrá habla en una sola lengua. Un sistema de verificación *perfecto* debería detectar *sólo* la lengua objetivo que aparezca en cada segmento.

3.1 Modos de evaluación

Los sistemas se evaluarán en modo cerrado y en modo abierto. En modo cerrado, sólo se contabilizarán los resultados obtenidos sobre aquellos segmentos del conjunto de evaluación que contengan habla en alguna de las lenguas objetivo. En modo abierto, se contabilizarán todos los resultados, también aquéllos obtenidos sobre segmentos con habla en lenguas *desconocidas*. Un sistema de verificación perfecto no debería detectar ninguna de las lenguas objetivo en un segmento que contenga habla en una lengua *desconocida*.

Evidentemente, las premisas de diseño y los elementos que conforman un sistema de verificación serán distintos en uno y otro caso. Al distinguir dos modos de evaluación, se da a los grupos participantes la oportunidad de diseñar sistemas específicos para cada caso. A lo largo de toda la evaluación, incluso cuando se suministren las claves del conjunto de evaluación, la identidad de las lenguas desconocidas permanecerá oculta. Tampoco se suministrarán datos de entrenamiento para ellas. No obstante, junto a los materiales de entrenamiento, se suministrará un conjunto de desarrollo similar al de evaluación, que incluirá también segmentos en lenguas desconocidas, cuya identidad y distribución podrían no coincidir con las del conjunto de evaluación.

3.2 Duración de los segmentos del conjunto de evaluación

El conjunto de evaluación (también el de desarrollo, que, como se ha dicho, tendrá características similares) presentará segmentos de tres duraciones distintas, de aproximadamente 30, 10 y 3 segundos, que contendrán mayoritariamente voz y sólo algunos fragmentos de silencio o ruido de fondo. Ello permitirá medir el rendimiento de los sistemas de verificación frente a distintas cantidades de habla. No se identificará la duración de cada segmento. Internamente, los segmentos contendrán íntegra y literalmente un fragmento de la grabación original, sin cortes ni manipulaciones de ninguna clase. Por último, aunque cada segmento contendrá habla en una sola lengua, el habla podría provenir de varios locutores distintos.

3.3 Sistemas libres vs sistemas restringidos

Los grupos participantes deberán preparar sus sistemas de verificación preferiblemente a partir de los materiales específicos de entrenamiento y desarrollo que les serán suministrados, aunque también podrán utilizar cualesquiera otros materiales, bien de forma directa, bien de forma indirecta, incorporándolos en subsistemas auxiliares (reconocedores de habla, decodificadores acústico-fonéticos, etc.). Teniendo en cuenta los materiales utilizados para desarrollar los sistemas, se distinguirán: (1) *sistemas restringidos*, que sólo utilizarán los datos de entrenamiento y desarrollo suministrados específicamente para esta evaluación; y (2) *sistemas libres*, que podrán incorporar cualesquiera datos y subsistemas.

3.4 Sistemas primarios vs sistemas alternativos

Atendiendo a los modos de evaluación y a los materiales utilizados para desarrollar los sistemas, se distinguirán 4 competiciones distintas: CL (modo cerrado, sistema libre), CR (modo cerrado, sistema restringido), AL (modo abierto, sistema libre) y AR (modo abierto, sistema restringido). Cada grupo podrá enviar un único sistema primario a cada competición, y tantos sistemas alternativos como desee. Al efecto de obtener el ranking de grupos, en cada competición sólo se tendrán en cuenta los resultados de los sistemas primarios, aunque se mostrarán los resultados de todos los sistemas presentados. En todos los casos, los resultados se desglosarán por duraciones.

3.5 Entradas y salidas del sistema de verificación

Por cada sistema presentado deberán enviarse los resultados de las pruebas de verificación para las 4 lenguas objetivo. Es decir, cada segmento del conjunto de evaluación participará en 4 pruebas de verificación, una por cada lengua objetivo. El sistema no se considerará válido si faltan los resultados de verificación correspondientes a alguna de las 4 lenguas objetivo.

Para cada prueba de verificación, el sistema utilizará exclusivamente la siguiente información:

- Un segmento de señal con habla
- La especificación de la lengua objetivo
- La especificación de las lenguas de contraste

y entregará como salida:

- La decisión (SI o NO) sobre si el segmento contiene la lengua objetivo
- Una puntuación, tanto más grande (más positiva) cuanto mayor sea la probabilidad de que el segmento contenga la lengua objetivo.

Los grupos participantes podrán indicar si sus puntuaciones representan *Log Likelihood Ratios (LLR)*, al efecto de utilizarlas para calcular una medida alternativa de rendimiento (C_{LLR}, véase el apartado 3.7).

3.6 Medida básica de rendimiento

Para medir el rendimiento de un sistema, se compararán las decisiones tomadas por el sistema en las pruebas de verificación con las decisiones correctas. En primera instancia, se estimarán las probabilidades de error de rechazo y error de aceptación, y a partir de éstas se calculará una medida básica de rendimiento.

Supongamos que hay N lenguas objetivo y que el conjunto de evaluación E es la unión de $N+1$ subconjuntos disjuntos: E_j con segmentos que contienen habla en la lengua L_j, $j \in [1,N]$, y E_0 con segmentos que contienen habla en una lengua L_0 desconocida. Consideremos las decisiones del sistema al verificar la lengua objetivo L_i sobre el conjunto de evaluación. En primer lugar, se contarán los segmentos de E_i que han sido rechazados, lo que permitirá estimar la probabilidad de error de rechazo $P_{miss}(i)$. A continuación, para cada L_j, *con* $j \neq i$, se contarán los segmentos de E_j que han sido aceptados, lo que permitirá estimar la probabilidad de error de aceptación $P_{false\_alarm}(i,j)$. Estas dos probabilidades se combinan en una medida $C(i,j)$ que llamaremos *coste bilateral*, que refleja el coste de las decisiones del sistema con respecto al par L_i como lengua objetivo y L_j como lengua de contraste, de acuerdo a un cierto modelo de coste que depende de la aplicación:

$$C(i,j) = C_{miss} \cdot P_{target} \cdot P_{miss}(i) + C_{false\_alarm} \cdot (1 - P_{target}) \cdot P_{false\_alarm}(i,j)$$

El modelo de coste viene dado por los valores de los parámetros C_{miss}, C_{false_alarm} y P_{target}. En esta evaluación se aplicarán los mismos valores que en la campaña 2007 de evaluación de sistemas de verificación de la lengua del NIST:

$$\begin{aligned} C_{miss} &= C_{false\_alarm} = 1 \\ P_{target} &= 0.5 \end{aligned}$$

Por último, se obtendrá un coste promedio, sumando las contribuciones para todas las lenguas objetivo y para todas las lenguas de contraste, como sigue:

$$C_{avg} = \frac{1}{N} \sum_{i=1}^{N} \left\{ \begin{array}{l} C_{miss} \cdot P_{target} \cdot P_{miss}(i) \\ + \quad \sum_{\substack{j=1 \\ j \neq i}}^{N} C_{false\_alarm} \cdot P_{non\_target} \cdot P_{false\_alarm}(i,j) \\ + \quad C_{false\_alarm} \cdot P_{OOS} \cdot P_{false\_alarm}(i,0) \end{array} \right\}$$

donde *N* es el número de lenguas objetivo, P_{non_target} es la probabilidad a priori de las lenguas de contraste (suponiéndolas equiprobables) y P_{OOS} es la probabilidad a priori de que un segmento contenga una lengua desconocida (*OOS: Out-Of-Set*). En esta evaluación, se utilizarán los siguientes valores:

$$P_{OOS} = \begin{cases} 0.0 & \text{en modo cerrado} \\ 0.2 & \text{en modo abierto} \end{cases}$$

$$P_{non\_target} = \frac{1 - P_{target} - P_{OOS}}{N - 1}$$

El coste promedio C_{avg} se calculará de manera separada para cada uno de los tres subconjuntos de segmentos de duraciones 3, 10 y 30 segundos.

3.7 Medida alternativa de rendimiento

En los casos en que las puntuaciones suministradas por un sistema de verificación de la lengua representen *Log Likelihood Ratios* (LLR), dichas puntuaciones se utilizarán para calcular una medida alternativa de rendimiento, conocida como C_{LLR}, en la que no aparecen parámetros de coste dependientes de la aplicación, por lo que tendrá un carácter más general.

Sea $LR(S,L_i)$ el *Likelihood Ratio* correspondiente al segmento *S* y la lengua objetivo L_i. En función de las probabilidades condicionales de *S* con respecto a las dos hipótesis alternativas, a saber, que *S* contenga la lengua objetivo L_i o que *S* contenga otra lengua, el *Likelihood Ratio* viene dado por:

$$LR(S, L_i) = \frac{Prob(S|L_i)}{Prob(S|\bar{L}_i)}$$

Considerese de nuevo el conjunto de evaluación *E*, resultado de la unión de *N+1* subconjuntos disjuntos: E_j, $j \in [1,N]$, con segmentos que contienen habla en la lengua L_j, y E_0, con segmentos que contienen habla en una lengua L_0 desconocida. Se definen los costes bilaterales $C_{LLR}(L_i,L_j)$, con $i \in [1,N]$ y $j \in [0,N]$, como sigue:

$$C_{LLR}(L_i, L_j) = \begin{cases} \frac{1}{|E_i|} \cdot \sum_{S \in E_i} \log_2 \left(1 + LR(S, L_i)^{-1}\right) & \text{si j = i} \\ \frac{1}{|E_j|} \cdot \sum_{S \in E_j} \log_2 \left(1 + LR(S, L_i)\right) & \text{si j} \neq \text{i} \end{cases}$$

El coste promedio $C_{LLR\text{-}avg}$ se define[2]:

$$C_{LLR-avg} = \frac{1}{N}\sum_{i=1}^{N}\left\{\begin{array}{l} P_{target}\cdot C_{LLR}(L_i,L_i) \;\; + \\ \sum_{\substack{j=1\\ j\neq i}}^{N} \; P_{non\_target}\cdot C_{LLR}(L_i,L_j) \;\; + \\ P_{OOS}\cdot C_{LLR}(L_i,L_0) \end{array}\right\}$$

3.8 Representación gráfica del rendimiento (curvas DET)

Al igual que en las evaluaciones del NIST, a partir de las puntuaciones se generarán curvas DET (*Detection Error Tradeoff*)[3] que mostrarán el rendimiento de los sistemas en distintos puntos de operación (correspondientes a distintos valores del umbral de verificación) y que permitirán discernir visualmente cuál o cuáles de ellos tienen un mejor comportamiento. Estas curvas, generadas mediante software del NIST[4], incluirán sendas marcas para el punto de operación especificado por las decisiones del sistema y el punto de operación de coste mínimo.

4 Materiales de entrenamiento, desarrollo y evaluación

Los materiales de entrenamiento, desarrollo y evaluación provienen todos ellos de programas de televisión (informativos, documentales, debates, entrevistas, reportajes, magazines, etc.), de tres conjuntos disjuntos e independientes. Esto significa, por ejemplo, que un programa utilizado para entrenamiento no aparecerá ni en desarrollo ni en evaluación.

Para acceder a estos datos es necesario registrarse en la Evaluación ALBAYZIN-08 VL, y aceptar las condiciones de participación que se detallan más abajo. Entre ellas, cabe destacar el compromiso de utilizar estos datos exclusivamente con fines de investigación, de no distribuirlos ni comerciar con ellos y de referenciarlos adecuadamente en cualesquiera publicaciones a que dieran lugar.

Las señales se han adquirido a través de un mismo dispositivo (una grabadora digital Roland Edirol R-09) y se han depositado en ficheros WAV (monocanal, 16 Khz, 16 bits/muestra). Los materiales incluyen diversos tipos de habla: leída, planificada,

2 Las razones para usar esta función de coste, así como sus posibles interpretaciones, se tratan con detalle en un artículo de Niko Brummer y Johan du Preez sobre reconocimiento del locutor: "Application-independent evaluation of speaker detection", Computer Speech and Language, Vol. 20, N. 2-3, April-July 2006, pp. 230-275. Esta misma función de coste se relaciona con la verificación de la lengua en otro artículo de Niko Brummer y David A. van Leeuwen: "On Calibration of Language Recognition Scores", Proceedings of 2006 IEEE Odyssey -The Speaker and Language Recognition Workshop.

3 Véase "The DET Curve in Assesment of Detection Task Performance", Proceedings of Eurospeech 1997, Vol.4, pp. 1895—1898, accesible vía web en: http://www.nist.gov/speech/publications/index.htm.

4 http://www.nist.gov/speech/tools/DETware_v2.1.targz.htm

conversacional formal, espontánea, etc. Asimismo, aunque la SNR es bastante buena en casi todos los casos, las condiciones ambientales y de canal son también muy diversas: entrevistas en estudio sin ruido de fondo, reportajes desde la calle, desde una fiesta, desde una manifestación, llamadas telefónicas en directo, reportajes con una ligera música de fondo, programas concurso o de humor con risas y aplausos, etc.

El conjunto de evaluación, que se suministrará en la última fase del proceso a través de un DVD en el formato que se detalla en la sección 5.2.1, tendrá no más de 8 horas de duración, constará de un total de no más de 2000 ficheros, con habla en las 4 lenguas objetivo y en otras lenguas desconocidas, de 3 duraciones distintas (3, 10 y 30 segundos), con al menos 100 ficheros por cada duración y lengua, mezclados aleatoriamente y con nombres también aleatorios. En todo caso, cada señal contendrá mayoritariamente voz y sólo algunos fragmentos de silencio o ruido de fondo.

El conjunto de entrenamiento constará de aproximadamente 8 horas por lengua (unas 32 horas en total), en ficheros de duración variable, que, al igual que los de evaluación, contendrán mayoritariamente voz (en condiciones ambientales y de canal diversas) y sólo pequeños fragmentos de silencio o ruido de fondo. Como se ha dicho, está permitido utilizar, directa o indirectamente, otros datos de entrenamiento, con la condición de incluir en la descripción del sistema suficiente detalle sobre el volumen y composición de los datos y subsistemas adicionales, así como del uso que se ha hecho de ellos.

Junto al conjunto de entrenamiento se suministrará un conjunto de desarrollo de características similares al de evaluación. También se suministrará el *script* de evaluación y el fichero de claves del conjunto de desarrollo, para que cada grupo pueda evaluar los sistemas que vaya desarrollando. Este *script* será prácticamente idéntico al utilizado por el NIST en la *2007 Language Recognition Evaluation*[5], salvo por la adición de una nueva tarea de verificación y los identificadores correspondientes a las 4 lenguas objetivo (castellano, catalán, euskera y gallego).

5 Condiciones generales de participación

5.1 Procedimiento general

Tras el envío de los materiales de entrenamiento y desarrollo, los grupos participantes dispondrán de tres meses para desarrollar sus sistemas. A continuación, tras el envío del conjunto de evaluación, dispondrán de tres semanas para procesarlo y enviar los resultados. Puesto que la evaluación se llevará a cabo mediante el *script* del NIST, los resultados de verificación deberán enviarse en el formato requerido por dicho software: un

5 http://www.nist.gov/speech/tests/lang/2007/score_lre07.v01b.tgz

fichero de texto con una línea por cada prueba de verificación y 6 campos por línea, que hacen referencia a la competición, la lengua objetivo, el modo de evaluación, el nombre del fichero de test, la decisión tomada con respecto a la detección de la lengua y la puntuación asignada por el sistema. Esta puntuación se interpretará de modo que cuanto mayor sea su valor, más alta será la probabilidad de detectar la lengua objetivo. El ranking de sistemas en todas las competiciones y subcategorías se efectuará tomando como referencia el coste promedio, según se ha definido en el apartado 3.6. También se calcularán y mostrarán las gráficas DET, y en aquellos casos en que la puntuación suministrada se identifique como *Log Likelihood Ratio* (LLR), se indicará el valor promedio de la medida alternativa C_{LLR}, tal como se ha definido en el apartado 3.7.

Como se ha dicho en el apartado 3.4, cada grupo podrá enviar hasta 4 sistemas primarios, uno por competición: CR, CL, AR y AL, y tantos sistemas alternativos como desee. Al efecto de establecer el ranking de grupos en cada competición, sólo se tendrán en cuenta los resultados de los sistemas primarios. Por último, el premio ALBAYZIN-08 de Evaluación de Sistemas de Verificación de la Lengua se entregará al ganador (esto es, al que obtenga un menor coste promedio C_{avg}) en la competición CR dentro de la subcategoría de segmentos de 30 segundos. El grupo organizador de esta evaluación (GTTS, de la UPV/EHU) no podrá optar al premio ALBAYZIN-08.

5.2 Formato de los datos

5.2.1 Datos de evaluación

Los datos de evaluación se distribuirán en un único DVD. Este contendrá una única carpeta denominada VL08-Eval, que a su vez contendrá dos elementos:

- *seg.ndx*: fichero de texto con la lista de segmentos del conjunto de evaluación
- *data*: carpeta con los segmentos del conjunto de evaluación, ficheros de señal en formato WAV (monocanal, 16 kHz, 16 bits) cuyos nombres serán cadenas alfanuméricas pseudo-aleatorias seguidas de la extensión *.wav*

5.2.2 Resultados de verificación

Los resultados de verificación de un sistema irán todos ellos en un único fichero de texto, con una línea por cada prueba de verificación y 6 campos por línea, separados por blancos, en el orden siguiente (tal como exige el *script* de evaluación del NIST):

1. El tipo de sistema utilizado en la evaluación: "**VL08-Eval-R**" (sistema restringido) o "**VL08-Eval-L**" (sistema libre).
2. La lengua objetivo ("**castellano**", "**catala**", "**euskera**" o "**galego**").
3. El modo de evaluación: "**closed-set**" (cerrado) u "**open_set**" (abierto).

4. El nombre del fichero de test (sin la extensión .wav).
5. La decisión tomada con respecto a la detección de la lengua objetivo: "**T**" (lengua detectada) o "**F**" (lengua no detectada).
6. La puntuación asignada por el sistema (un número real, interpretado de modo que cuanto mayor sea su valor, más alta será la probabilidad de detectar la lengua objetivo).

5.3 Procedimiento de envío

Los ficheros de resultados de cada sistema, así como la descripción del sistema o sistemas, se enviarán como anexos por correo electrónico a luisjavier.rodriguez@ehu.es. En el correo electrónico deberá indicarse también, por cada sistema enviado, si las puntuaciones suministradas por dicho sistema se pueden interpretar como *log likelihood ratios*. Los nombres de los ficheros de resultados se compondrán de acuerdo al siguiente esquema:

<id_grupo>_<id_competición>_<id_sistema>.out ,

donde *id_grupo* es el identificador del grupo participante, *id_competición* el identificador de la competición (CR, CL, AR o AL) e *id_sistema* el identificador del sistema (primario, alt1, alt2, etc.). Así, por ejemplo, si el grupo GTTS envía dos sistemas, uno primario y otro alternativo, a la competición CR, los nombres de estos ficheros serían:

GTTS_CR_primario.out
GTTS_CR_alt1.out

5.4 Descripción de los sistemas

Cada grupo participante deberá elaborar una descripción de su sistema o sistemas. Si se envían varios sistemas a la misma competición, uno de ellos deberá identificarse como sistema primario, y el resto como sistemas alternativos. El propósito de esta descripción es dar una idea lo más clara y precisa posible del funcionamiento de un sistema, teniendo en cuenta los siguientes aspectos:

- La audiencia estará compuesta por desarrolladores de sistemas, científicos o tecnólogos familiarizados con algunas de las tecnologías implicadas, pero no forzosamente con las técnicas y algoritmos aplicados en un sistema determinado. Por ello, se trata de dar explicaciones claras, evitando en lo posible la jerga especializada y abreviaturas.
- La descripción será precisa y completa. Desde el punto de vista de otros desarrolladores, una visión general (pero superficial) del sistema no tiene ninguna utilidad. Sin llegar a dar detalles a pequeña escala, se trata de incluir todas las

características relevantes del sistema, de modo que otro desarrollador pueda construirlo por su cuenta. Para no ser excesivamente prolijos en la descripción de todos los elementos del sistema, se recomienda utilizar referencias bibliográficas en la medida de lo posible.

Para mantener una cierta homogeneidad formal, las descripciones de los sistemas se editarán mediante las plantillas WORD o LaTeX de las VJTH, accesibles desde la siguiente página web: http://jth2008.ehu.es/e_articulos.html. El documento solicitado debería incluir al menos las siguientes secciones:

1. Introducción
2. Sistema A (nombre o identificador del sistema)
   2.1. Descripción del sistema
   Explicar claramente los métodos y modelos utilizados en el sistema A
   2.2. Datos de entrenamiento
   Dar cuenta de todos los datos que, directa o indirectamente, se hayan utilizado para desarrollar el sistema A. Debe incluirse la fuente de los datos, las condiciones de adquisición, volumen y tipo de datos, año de publicación y cualquier otra información que se considere relevante.
   2.3. Tiempo de procesamiento
   Se deberá calcular la razón de tiempo real, esto es, el tiempo de CPU empleado por el sistema para procesar el conjunto de evaluación para todas las lenguas objetivo dividido por la duración del conjunto de evaluación. El tiempo de CPU incluye las operaciones de E/S y el procesamiento necesario para tomar las decisiones de detección, y se mide en términos del tiempo total empleado por una sola CPU, de principio a fin. Los sistemas que no procesen los datos de evaluación de una sola vez no se verán penalizados, ya que sólo se deberá contabilizar el tiempo de procesamiento, no los lapsos comprendidos entre las diferentes etapas de procesamiento. No se considerará tampoco el tiempo empleado en inicializar el sistema (esto es, cargar los modelos en memoria) ni en preprocesar las señales para eliminar ruido o eco. En todo caso, deberán indicarse las especificaciones de CPU y memoria de la máquina o máquinas utilizadas.
3. Sistema B (nombre o identificador de otro sistema)
   Esta sección es similar a la sección 2 pero para otro sistema B. Si el sistema B fuera un sistema alternativo al sistema A (primario), deberán destacarse las diferencias entre uno y otro. Deberá añadirse una sección como ésta por cada sistema enviado a la evaluación.

4. Referencias

   Lista de referencias a libros, artículos, ponencias, etc. sobre aspectos relevantes de los sistemas presentados a la evaluación.

5.5 Reglas de participación

Se resumen aquí las reglas básicas que deben observar los participantes en esta evaluación:

- Cada grupo participante constará de uno o más investigadores.
- Los grupos que deseen participar en esta evaluación deberán inscribirse antes del 31 de julio de 2008 a través del formulario accesible desde la página web de la Evaluación ALBAYZIN-08 VL: http://gtts.ehu.es:8080/RTTH-LRE08/Formulario.jsp. Una vez validada su inscripción, y a partir del 30 de junio de 2008, les serán enviados los materiales de entrenamiento y desarrollo. La información sobre los grupos participantes será publicada en la página web a medida que vayan formalizándose las inscripciones. Los interesados pueden dirigirse directamente al comité organizador para consultar aspectos que no estén claros sobre la evaluación, por e-mail a luisjavier.rodriguez@ehu.es y por teléfono al número 946012716.
- Los datos suministrados (entrenamiento, desarrollo y evaluación) se utilizarán exclusivamente con fines de investigación. En el futuro, los grupos participantes podrán utilizar estos datos para desarrollar o evaluar sus propios sistemas, pero no podrán distribuirlos ni comerciar con ellos. Si dieran lugar a publicaciones, deberán referenciar la base de datos como sigue:

  > KALAKA. Speech database created for the 2008 Language Recognition Evaluation on Spanish Languages, organized by the Spanish Network on Speech Technology. Produced by the Software Technologies Working Group (GTTS, http://gtts.ehu.es), University of the Basque Country.

- Una vez recibidos los datos de evaluación (29 de septiembre de 2008), los grupos participantes deberán procesarlos y devolver los resultados de verificación de su sistema o sistemas dentro del plazo límite establecido (19 de octubre de 2008, 24:00 hora de España peninsular), según formato y método especificados en los apartados 5.2.2 y 5.3 de este plan.
- La información disponible por un sistema en cada prueba de verificación se limitará a la especificada en el apartado 3.5. En particular:

- cada segmento de evaluación debe ser procesado de manera independiente y sin información alguna sobre el resto de los segmentos, y
- no está permitido escuchar las señales del conjunto de evaluación.

□ Los resultados de verificación de un sistema deben incluir todas las señales del conjunto de evaluación y todas las lenguas objetivo. En particular, no se considerarán válidos los resultados de verificación si faltan los correspondientes a alguna lengua objetivo.

□ En cada competición, cada grupo podrá enviar tantos sistemas alternativos como desee, pero deberá enviar uno (y sólo uno) identificado como primario.

□ Los envíos deberán incluir la identificación del grupo y un fichero de resultados por cada sistema, indicando a qué competición se presenta (CR, CL, AR, AL) y si las puntuaciones corresponden o no a *log likelihood ratios*.

□ Se entregará el premio ALBAYZIN-08 de Evaluación de Sistemas de Verificación de la Lengua al grupo ganador (esto es, al que obtenga un menor coste promedio C_{avg}) en la competición CR dentro de la subcategoría de segmentos de 30 segundos.

□ Cada grupo deberá enviar, junto a los resultados, un documento en formato PDF donde se describirán con suficiente detalle las características del sistema o sistemas (ya sean primarios o alternativos) que se presenten a la evaluación. Para homogeneizar el aspecto de estas contribuciones, que se publicarán en el CDROM de las Jornadas junto con el informe final del comité organizador de la Evaluación ALBAYZIN-08, se utilizarán como base las plantillas LaTeX o WORD de las Jornadas.

□ Cada grupo participante deberán enviar uno o más representantes al Workshop de Evaluaciones ALBAYZIN-08, dentro las V Jornadas en Tecnología del Habla, para presentar los detalles técnicos de su sistema o sistemas, así como los resultados obtenidos, y discutir la problemática de esta evaluación. Este Workshop se realizará en sesión abierta a todos los participantes de las Jornadas.

□ Este plan podría sufrir modificaciones, debido a restricciones o necesidades no previstas inicialmente, a errores o a imprecisiones, que en todo caso serán comunicadas a todos los grupos participantes.

6 Calendario

- 5 de mayo de 2008
  - Publicación del plan de evaluación en la web de las VJTH
  - Apertura del plazo de inscripción
  - Habilitación de un formulario de inscripción en línea accesible desde la web de las VJTH
- 30 de junio de 2008
  - Comienzo del envío de los materiales de entrenamiento y desarrollo, junto con el script de evaluación, a los grupos participantes.
  - Habilitación de una web colaborativa, para consulta de dudas, aportación de comentarios, etc. con una parte pública y otra de acceso restringido a los grupos participantes
- 31 de julio de 2008
  - Fin del plazo de inscripción
- 29 de septiembre de 2008
  - Envío de los datos de evaluación a los grupos participantes
  - Apertura del plazo de envío de resultados
- 19 de octubre de 2008 (24:00, hora de España peninsular)
  - Fin del plazo de envío de resultados
- 31 de octubre de 2008
  - Notificación a cada grupo participante de los resultados de la evaluación de su sistema o sistemas
  - Liberación del etiquetado de lenguas del conjunto de evaluación
- 12 de noviembre de 2008
  - Workshop de Evaluaciones ALBAYZIN-08 (12:00-13:30)
    - Entrega de documentación con el resumen y el análisis de los resultados de evaluación de todos los sistemas presentados
    - Presentación de los sistemas a cargo de los grupos participantes
    - Discusión
  - Presentación del resumen de resultados de la Evaluación ALBAYZIN-08 VL en sesión plenaria (17:00-17:45) dentro de las VJTH