OBAM-PV: una aplicación para el subtitulado de vídeos de Sesiones Plenarias del Parlamento Vasco

OBAM-PV: an application for subtitling Plenary Sessions videos at the Basque Parliament

Germán Bordel García, Mikel Penagarikano Badiola,
Luis Javier Rodríguez-Fuentes, Amparo Varona Fernández
Universidad del País Vasco (UPV/EHU)
Barrio Sarriena s/n, 48940 Leioa,
{german.bordel,mikel.penagarikano,luisjavier.rodríguez,amparo.varona}@ehu.es

Resumen: En este artículo se describe la estructura y los distintos problemas que se han abordado en el desarrollo de una herramienta software para el Parlamento Vasco, que permite la generación de subtítulos para vídeos con las transcripciones textuales disponibles. La mayor parte de las dificultades se encontraron en el pre-procesamiento del texto y en la sincronización de texto y audio. La herramienta es capaz de tratar con recursos multilingües, lo que también ha supuesto una fuente importante de dificultad.
Palabras clave: Subtitulado, Alineamiento, Reconocimiento Automático del Habla, Transcripción Fonética.

Abstract: In this paper, we describe the structure and the various issues that have been addressed in the development of a software tool for the Basque Parliament, that allows the generation of subtitles for videos with the verbatim transcripts available. Most of the difficulties were found performing text preprocessing and synchronizing text to audio. The tool has the ability to deal with multilingual resources which has also been a major source of difficulty.
Keywords: Subtitling, Alignment, Automatic Speech Recognition, Phonetic Transcription.

1 Introducción[1]

Desde septiembre de 2010, dentro de un acuerdo de colaboración que incluye un proyecto de investigación y un contrato de servicio, el grupo GTTS viene proporcionando al Parlamento Vasco (PV) los vídeos con subtitulado de Sesiones Plenarias que ofrece en su web oficial. Estos vídeos combinan las actas de las sesiones con los vídeos originales, permitiendo el cumplimiento de los requerimientos legales sobre accesibilidad que afectan a las administraciones públicas desde el 31 de diciembre de 2008 (RD 1494/2007) .

Los detalles de este servicio y el modo en que se abordaba inicialmente su ejecución se presentaron en Bordel et al. (2011). Posteriormente, en Bordel et al. (2012), se presentó una evaluación del alineador, elemento principal de la solución aplicada.

En este trabajo se presentan los distintos elementos que se han construido y depurado en torno a esta actividad, y que con el tiempo han ido conformando una batería de utilidades y herramientas que pueden combinarse para ofrecer una solución de fácil uso al PV. El objetivo principal a la hora de diseñar esta herramienta fue la sencillez de uso, de forma que pudiera ser utilizada por personal ajeno a las tecnologías implicadas. No se trata de una herramienta de ayuda que permite la interacción con el usuario, al estilo de la presentada en (Álvarez, Pozo y Arruti 2010), o la referida en (Cerisara, Mella y Fohr, 2009), sino una solución más cerrada y específica. En la Figura 1 se muestra la interfaz que permite realizar el subtitulado con tan sólo pulsar dos botones e intercalando entre ellos unas acciones manuales. La información básica

[1] Este trabajo está siendo cofinanciado por la Universidad del País Vasco y el Parlamento Vasco a través del proyecto US11/06, y recibe apoyo del proyecto UPV GIU10/18, y del S-PE12UN55 del Gobierno Vasco dentro del programa SAIOTEK.

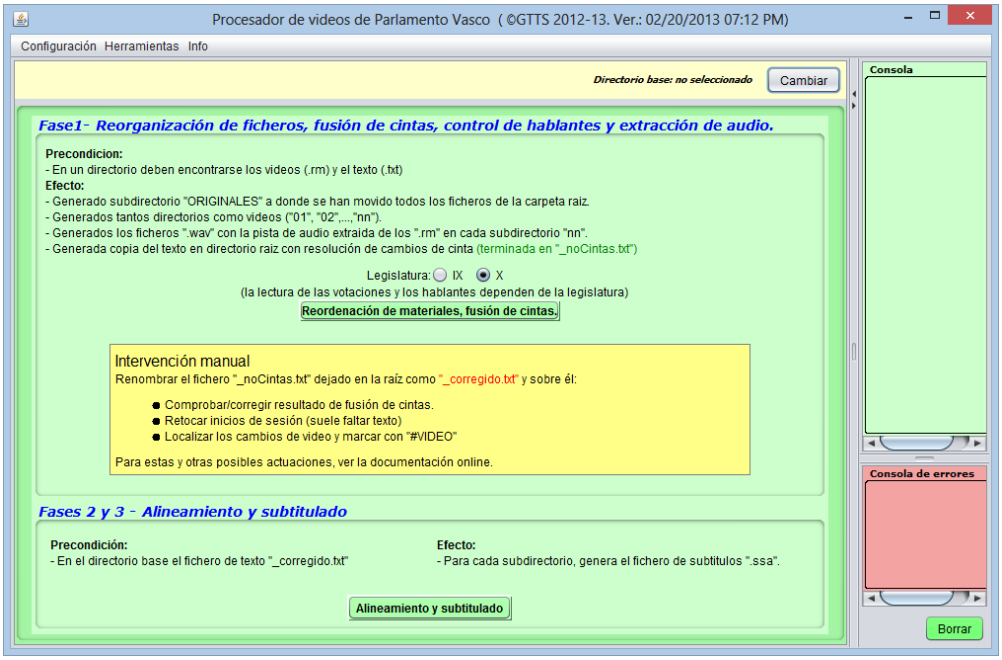

Figura 1: Interfaz de la aplicación. La fase 1 es específica para el caso del Parlamento Vasco, mientras que las fases 2 y 3 son procesos generales que se ejecutan conjuntamente. En el menú se dispone de todas las herramientas que dan soporte a las dos acciones disponibles en la interfaz.

para que la tarea se ejecute correctamente se presenta en la misma interfaz, de modo que el usuario no necesite acudir a la documentación si no surge ninguna situación especial.

El artículo está organizado como sigue. En la sección 2 se describe de un modo somero la estructura general de la aplicación, que no difiere sustancialmente de la presentada con anterioridad en Bordel et al. (2011). En la sección 3 se analizan las dificultades encontradas y soluciones aportadas en cada módulo. Por último, en la sección 4 se detallan las mejoras actuales y previstas de la herramienta.

2 *Estructura del sistema*

El procesamiento está estructurado en tres etapas, donde la central es la que resuelve el problema de la sincronización entre dos "streams" de entrada (audio y texto), mientras que la previa y la posterior preparan los datos y postprocesan los resultados respectivamente (véase la Figura 2). Esta pieza central que alinea texto y audio es la clave del sistema. Pese a su complejidad, se ejecuta como un transformador entrada/salida de un solo paso (el acrónimo que se le dio originalmente obedece a este objetivo: One Button Alignment Machine - OBAM)

La primera etapa de adecuación de las entradas es específica para los requerimientos del PV, aunque también usa recursos generales (particularmente el núcleo del alineador). La última etapa se limita a procesar la salida para generar subtítulos.

La OBAM está formada por un alineador de secuencias de símbolos que toma por entradas las transformaciones de audio y texto a fonemas generadas por dos módulos diferenciados. Para el audio se utiliza un módulo de decodificación acústico-fonética que puede ser tanto HTK (Young et al., 2006) como el sistema de GTTS Sautrela (Penagarikano y Bordel, 2005). La transcripción de texto a fonemas se lleva a cabo mediante un transcriptor multilingüe desarrollado también por GTTS.

3 *Dificultades y soluciones reseñables*

Cada uno de los elementos que conforman el sistema ha planteado una serie de retos. Pero antes de entrar en materia, abordaremos un aspecto de interés para su funcionamiento conjunto: el mecanismo de anotación del texto de entrada.

Dado que la OBAM no debía requerir de ninguna interacción ni ajuste durante su ejecución, y debía ser capaz de transmitir entre su entrada y su salida cierta meta-información no predeterminada —p.ej. cambios de oradores, u otras anotaciones para los módulos de postpro-

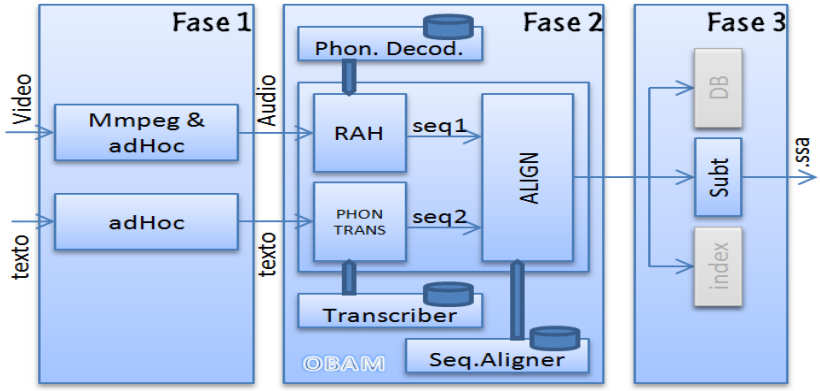

Figura 2: estructura de la aplicación.

ceso—se ha establecido un mecanismo genérico de anotación que permite que cada módulo sea transparente o atienda a determinadas marcas. De este modo el texto de entrada puede contener líneas de comentario que son transmitidas o interpretadas por los diferentes módulos.

En concreto, para el postproceso que nos ocupa —el que da lugar al subtitulado— son especialmente importantes las marcas que permiten mantener diferencias entre el texto que se pretende mostrar en los subtítulos y lo realmente hablado, como puede observarse en la Figura 3.

```
#<SPK txt="LEHENDAKARIAK (Tejeria Otermin)" id="33.A" />
#<ALTSUBT txt="">
Egunon guztioi osoko bilkurari hasiera emango diogu
mesedez bakoitza bere eserlekuan eser dadila
#</ALTSUBT>
Gai-zerrendari ekin baino lehen, gogoratu egingo dizuet
Arantza Quiroga eta Natalia Rojo legebiltzarkideak botoa
delegatu egin dutela.
```

Figura 3: El uso de marcas introducidas como comentarios en el texto original permite transmitir información a los módulos de postproceso (tales como cambios de hablante o subtítulos alternativos).

A continuación se presentan las funciones de cada módulo y las dificultades encontradas en cada caso.

3.1 Fase 1. Adecuación de entradas

La mecánica de alineamiento texto-voz es más efectiva cuanto más próximo es el texto transcrito a la literalidad de lo hablado. Por ello, en la primera fase se pretende adecuar ambas entradas lo más posible de modo que el resultado del alineamiento no se vea afectado por disparidades evitables. El resultado de la sincronización entre las dos entradas será tanto mejor cuanto más correlación exista entre ellas inicialmente.

Como se aprecia en la Figura 1, el botón de la parte superior permite aplicar todas las adecuaciones automáticas, y a continuación se muestran las instrucciones para las posibles actuaciones manuales que permiten cumplir este objetivo.

Evidentemente, estas acciones se refieren básicamente a la transcripción textual, puesto que en principio la única actuación necesaria respecto al audio es su extracción del contenedor multimedia. No obstante, en ocasiones surge la necesidad de eliminar zonas amplias de habla no transcrita[2], por lo que esta primera fase incluye la extracción del audio en lugar de admitirse directamente el vídeo como entrada en la segunda fase.

En todo caso, siempre existirá una disparidad entre ambas secuencias, puesto que sólo a partir del texto no puede inferirse con total seguridad cómo se ha producido el discurso.

De la experiencia concreta con las sesiones plenarias del Parlamento Vasco, observamos cuatro causas de disparidad:

- la mecánica específica con que se realizan las transcripciones,
- la codificación de determinados elementos estructurales del discurso,
- las disfluencias presentes en el habla,
- la dificultad propia del proceso de transcripción.

[2] En las sesiones del Parlamento Vasco esto sucede en determinadas votaciones en las que se llama a todos los parlamentarios desde la mesa para depositar su voto, mientras que el acta sólo refleja el resultado final.

```
== Reordenación y fusión (Fri Apr 26 14:50:04 CEST 2013)   ==
Procesando C:\PV\10-024-2013abr18

CONTROL DE HABLANTES:
     Correcto. Todos los hablantes y cargos identificados
VOTACIONES: 17
CONTROL DE SECUENCIA DE CINTAS:
    Secuencia de cintas correcta.
UNIÓN DE CINTAS: Longitud coincidente (en palabras)
(01-02: 25)(02-03: 20)
(03-04:  1 ***ERROR*** No ejecutada ninguna acción)
(04-05: 28)(05-06: 11)(06-07:  5)(07-08: 13)(08-09:  6)
(09-10: 10)(10-11: 10)(11-12: 15)(12-13: 16)(13-14: 20)
(14-15: 12)(15-16: 16)(16-17: 10)(17-18:  9)(18-19: 31)
(19-20:  0 ***ERROR*** No ejecutada ninguna acción)
(20-21: 22)(21-22:  9)(22-23:  9)(23-24:  8)(24-25: 17)
(25-26:  8)(26-27:  5)
(27-28:  3 ***ERROR*** No ejecutada ninguna acción)
(28-29: 12)(29-30:  6)(30-31: 10)(31-32:  5)(32-33: 19)
(33-34: 18)
CAMBIOS DE CINTA: 33, con problemas: 3
== Fin de Reordenación y fusión (Fri Apr 26 14:50:06 CEST 2013) ==
```

Figura 4: Información obtenida sobre el proceso de adecuación del acta, donde puede verse el resultado de los intentos de fusión de cintas de transcripción.

A continuación se analiza brevemente cada uno de los casos.

3.1.1 Mecánica específica de la transcripción

La tarea de transcribir manualmente un discurso en tiempo real no es sencilla, ya que requiere la habilidad de mecanografiar a la velocidad suficiente, tomar decisiones en determinadas situaciones de ambigüedad, conocer términos en ocasiones poco comunes, etc. Por ello, este proceso conlleva un cierto nivel de imperfección en los resultados, muy dependiente de la capacitación de las personas que lo llevan a cabo. Lo que nos interesa particularmente en este punto, es que se trata de una actividad que no se puede mantener sostenidamente durante un largo periodo de tiempo. Por este motivo, es práctica habitual que las transcripciones de discursos largos se lleven a cabo por al menos dos personas, que se van turnando para permitir periodos de descanso. El resultado es una transcripción formada por fragmentos con unos puntos de conexión que no son exactamente coincidentes.

En el caso concreto de las transcripciones bilingües del Parlamento Vasco, la transcripción manual está contratada a una empresa que tiene establecidos turnos de unos 15 minutos, de modo que una sesión de duración media (5 horas) puede presentar unos 20 puntos de corte en las transcripciones. Estos puntos de corte muestran un solapamiento entre el final de la transcripción realizada por la persona que acaba su turno y el principio de quien la empieza. Normalmente estos segmentos no coinciden, principalmente porque la interpretación de la prosodia es ambigua y da lugar a puntuaciones diferentes, pero también es frecuente encontrar diferencias en la secuencia de palabras (véase la Figura 5).

Gobierno de España y de todos los gobiernos.
No nos parece ningún drama que este tipo de obras puedan sufrir determinados retrasos. **Yo creo que lo más grave es que puedan**

#PASO A CINTA 30 NO RESUELTO AUTOMATICAMENTE --- PENDIENTE

Yo creo que efectivamente lo más grave es que puedan sufrir retrasos las políticas sociales, el impulso a la actividad económica o a la promoción del empleo y todas aquellas políticas que sirvan para salir de la crisis. En fin, esto

Figura 5: En ocasiones los solapamientos en los turnos de transcripción no presentan un encaje suficiente como para ser resueltos automáticamente.

Este problema es parcialmente resuelto por la aplicación alineando ambos extremos del texto mediante la misma herramienta de alineamiento que se utilizará en la fase 2 para alinear texto y audio. Si se obtiene una secuencia de palabras coincidente de una longitud que sobrepase un determinado umbral, se unen adecuadamente. En caso contrario, se introduce un comentario que permite localizar el problema en el fichero para corregir manualmente la situación.

3.1.2 Codificación de elementos estructurales

Las sesiones parlamentarias tienen una mínima estructura que se ve reflejada en las actas. Es decir, no todo es transcripción literal de la co-

rrespondiente pista de audio, sino que hay anotaciones. Estas no son muy numerosas pero tienen la dificultad de no obedecer a unos criterios predefinidos, por lo que estos se han inferido a partir de ejemplos. Se trata por tanto de criterios que pueden cambiar y requieren de cierta capacidad de configuración.

Las marcas más habituales son las que determinan los comienzos de los discursos de cada hablante, que se conforman por sus dos apellidos y el término "andrea" o "jauna" ("señora" o "señor") terminando en dos puntos. También se indican las horas de comienzo y finalización de las sesiones y recesos (con frases indistinguibles del discurso si no es por su semántica), así como algunas situaciones como: (Geldiunea), (Barreak), (Txaloak), (Berbotsak), etc. (pausa, risas, aplausos, habla de fondo,...). Todas estas anotaciones pasan a través de los diversos mecanismos aplicados por el sistema y llegan a la salida sincronizados de modo que puedan ser utilizados si fuera preciso.

La situación más compleja es la que se da con las votaciones, puesto que el acta no recoge la lectura que se hace de los resultados por parte de la presidencia, sino que muestra una redacción estandarizada. Para estas situaciones, el sistema sustituye el texto por el que se considera "lectura más probable" (en la Figura 1 puede verse un selector de legislatura, que es tenido en cuenta en este caso, puesto que las "lecturas más probables" dependen de las personas que las realizan). Aunque la secuencia introducida no siempre sea la correcta, los resultados son satisfactorios gracias a la robustez del mecanismo de alineamiento frente a errores puntuales, así como al hecho de que los números son parte significativa de estas frases y suponen generalmente secuencias textuales correctas.

3.1.3 Disfluencias

Normalmente los transcriptores manuales filtran lo hablado para recoger un discurso literal pero libre de disfluencias. Esto no es algo que esté perfectamente definido, puesto que en ocasiones cabe la interpretación alternativa de que algo que podría considerarse una disfluencia sea un recurso retórico. Un caso muy claro es el de las repeticiones: en la medida en que un transcriptor considere que una repetición lleva una carga retórica y la consigne en el texto, la transcripción será más fiel al audio que si es filtrada.

En todo caso, el nivel de disfluencias puede llegar a ser no despreciable, aunque el efecto negativo de su filtrado se mantiene bastante acotado, al ser un fenómeno distribuido a lo largo del discurso que raramente afecta intensamente a un segmento de una longitud importante.

3.1.4 Dificultad propia del proceso de transcripción

Como ya se ha explicado al analizar los solapamientos entre segmentos de transcripción, dos transcriptores manuales diferentes generan resultados distintos, lo que necesariamente implica que la relación del texto con el audio presentará diferencias notables.

Aparte de la corrección de disfluencias, los transcriptores toman decisiones bajo la presión de mantener el ritmo de tiempo real, lo que puede llevarles a suprimir palabras valoradas como no estrictamente necesarias desde el punto de vista del resultado final (véase un ejemplo en la Figura 5). Pueden asimismo transcribir fonéticamente términos desconocidos para ellos, cometiendo errores (muy frecuentemente es el caso de siglas y nombres de empresas o de organismos). Y por último, deben generar transcripciones incluso cuando la calidad del audio se deteriora por muy diversos motivos (por causa del hablante, del entorno, o incluso por causas técnicas).

En Hazen (2006) se presentan algunos datos sobre la calidad de las transcripciones manuales. El material acumulado a lo largo del periodo en que GTTS viene trabajando con el PV permitiría un estudio en profundidad de estos fenómenos.

3.2 Fase 2. Alineamiento

Una vez que disponemos de dos secuencias de entrada —audio y texto— lo más próximas posible, la fase 2 se encarga de relacionarlas temporalmente. Como puede apreciarse en la Figura 2, esta fase es totalmente automática. Para ello utilizamos un alineador de símbolos que exige que convirtamos ambas secuencias a un alfabeto simbólico común.

Esto se lleva a cabo convirtiendo ambas entradas en secuencias fonéticas mediante un decodificador acústico-fonético para el audio, y mediante un transcriptor ortográfico-fonético multilingüe para el texto.

La elección del fonema como unidad básica es un elemento crucial para el funcionamiento de esta fase. Por un lado, permite la utilización de un decodificador fonético con independencia de los idiomas a reconocer, con la sola condición de que disponga de un conjunto de mode-

los fonéticos que les dé cobertura (como se muestra en Bordel et al. (2011) para el caso de Euskera y Español), y por otro, al tratarse de un vocabulario reducido (lo que no sucede si se trabaja a nivel de palabra u otra unidad similar), permite que el alineamiento de símbolos encaje ambas secuencias con una granularidad muy fina, donde los desajustes entre secuencias no tienen efectos a larga distancia y los puntos sincronizados correctamente no distan mucho entre sí. Una evaluación del alineamiento fonético se presentó en Bordel et al. (2011), donde se observa un rendimiento sólo ligeramente inferior al método clásico de Moreno et al. (1998), sin su elevado coste ni sus dificultades para el manejo de múltiples idiomas.

A continuación se detallan las dificultades encontradas y las soluciones adoptadas para cada uno de los elementos que componen esta etapa.

3.2.1 El decodificador fonético

Para la decodificación del audio se utiliza un decodificador sencillo, puesto que no es preciso obtener unas tasas de reconocimiento muy elevadas cuando la tarea a resolver es la identificación de los instantes de tiempo en que se pronuncian los fonemas. Puede llegar a ser admisible una tasa de reconocimiento relativamente baja siempre y cuando los errores se distribuyan homogéneamente y no haya zonas especialmente mal reconocidas. En este sentido, es más importante la robustez del sistema frente a ruidos, tipos de canal, etc. En particular, la tasa de reconocimiento en el caso de la aplicación que nos ocupa está en torno al 60% que, atendiendo a los resultados obtenidos, resulta más que suficiente.

Pueden encontrarse más detalles sobre el decodificador utilizado en Bordel et al. (2012).

3.2.2 El transcriptor grafema-fonema

Esta es una pieza clave del sistema en lo que se refiere a las dificultades que presenta tratar con recursos multilingües.

El proceso se resuelve mediante la utilización de un transcriptor multilingüe (TMU) que se apoya en transcriptores monolingües (TMO) y toma decisiones en aquellos casos en que la respuesta de éstos a una entrada presenta ambigüedad (es decir, cuando más de uno proporciona una transcripción o cuando no lo hace ninguno). En estos casos, el TMU estima un idioma para la unidad a transcribir en función de su entorno más próximo, o en función de uno más amplio si éste también presenta ambigüedades[3].

Los TMOs son consultados para ver si disponen de una transcripción tabulada en un diccionario, de modo que reconozcan la unidad como propia. Si ningún TMO reconoce la unidad, el TMU decide cuál de ellos debe proporcionar una transcripción, y éste la genera a partir de reglas. El sistema proporciona al usuario un listado con las palabras desconocidas transcritas mediante reglas para permitirle su validación y que de este modo pasen a formar parte del diccionario.

Los TMOs resuelven siempre los casos de transcripción de números y diversos símbolos, por lo que en estas situaciones siempre se está a expensas de la estrategia aplicada en el nivel multilingüe.

También en el nivel multilingüe, y previamente a las consultas a los TMOs, se detectan (mediante expresiones regulares) una serie de elementos específicos, como fechas, numeraciones de artículos, porcentajes, etc., que se resuelven específicamente. En ocasiones, estos elementos son dependientes del idioma (como fechas y porcentajes en Español y Euskera, que se ordenan de modo inverso), y esto es tenido en cuenta en la elección del TMO.

3.2.3 El alineador

El alineador de secuencias de símbolos es la pieza central del proceso. La principal dificultad a superar es la computacional: se están afrontando alineamientos de secuencias que duran por término medio 2,5 horas[4], lo que supone secuencias de unos 100.000 símbolos.

El método habitual para el alineamiento global de secuencias de símbolos es el de Needleman y Wunsch (1970), pero este es un proceso cuadrático tanto en tiempo como en espacio, por lo que la dimensión media de nuestro problema es del orden de 10^{10}. Esto no es excesivo en términos de coste temporal para un procesador medio actual que puede ejecutar del orden de 10^8 instrucciones por segundo, pero sí lo es en términos de coste espacial ya que trabajando

[3] Conviene resaltar que esta estrategia no siempre supondrá un "acierto" puesto que en ocasiones se pronuncian palabras aisladas en un idioma diferente al de su entorno.

[4] Las sesiones tienen una duración media de unas 5 horas, pero por limitaciones del material audiovisual utilizado se recogen en fragmentos que no pasan de 3 horas.

con enteros de 4 bytes estamos hablando de un requerimiento inicial de 40GB de memoria.

Hirschberg (1975) publicó un algoritmo lineal en espacio para resolver el problema de la determinación de la subsecuencia de máxima longitud común a dos secuencias de símbolos. Este algoritmo resuelve también el problema que se nos presenta. Es más, permite fragmentar el problema de modo recursivo en subproblemas cada vez más pequeños, lo que permite explotar adecuadamente las capacidades multihilo de los procesadores actuales. De este modo, el sobrecoste temporal frente al método directo es compensado por la mayor eficacia de la implementación. La adaptación del algoritmo para que haga un uso eficaz del procesamiento multihilo ha sido la principal dificultad a resolver en esta etapa.

El algoritmo de Hirschberg valora positivamente cada par de símbolos coincidentes, en lo que puede ser visto como el uso de un "kernel" que tiene como función objetivo la maximización del número de coincidencias ("aciertos"). Con un "kernel" que valore, por el contrario, negativamente las diferencias (sustituciones, inserciones y borrados), la función objetivo tratará de minimizar el número de operaciones de edición que relacionan ambas secuencias (la distancia de Levenshtein). En principio no está claro cuál de las dos estrategias puede ser la más beneficiosa para la tarea del alineamiento texto-voz. Tras el estudio del problema, en el sistema se implementa un kernel convenientemente ajustado para resultar en una función objetivo que maximiza el número de aciertos, minimizando a su vez como segundo objetivo la longitud (es decir, de entre todas las posibles secuencias de edición que maximizan los aciertos, se selecciona una con longitud mínima).

Algunos detalles del alineamiento y de su eficacia real pueden encontrarse en Bordel et al. (2012).

3.3 Fase 3. Subtitulado

Una vez que se dispone de la información de sincronización de las dos secuencias fonéticas, con toda la información que permite reconstruir las palabras, la puntuación e incluso otro tipo de informaciones (como turnos de hablantes, etc.), pueden realizarse diferentes postprocesamientos, de entre los cuales el subtitulado es el que está siendo explotado para dar servicio al PV.

Como puede apreciarse en la Figura 2, esta fase es totalmente automática y se ejecuta conjuntamente con la anterior.

Con esta sincronización fonética, puede reconstruirse el texto con las marcas de tiempo asociadas a cada palabra, y el problema del subtitulado se reduce a la segmentación de éste en porciones que sean adecuadas para su presentación al usuario.

El algoritmo implementado procesa repetidas veces el texto, partiendo originalmente de su segmentación en frases, para ir buscando lugares de corte de los fragmentos demasiado largos (conforme a un umbral configurable) y para unir fragmentos excesivamente cortos (igualmente conforme a otro umbral configurable). Los puntos de corte se estiman de acuerdo a los caracteres de puntuación, atendiendo a un conjunto de prioridades y buscando siempre las particiones más equilibradas. En cuanto a las uniones, hay que mencionar que se tienen en cuenta los cambios de hablantes para que los textos presentados sean siempre de un mismo hablante.

4 Mejoras previstas y en curso

En las secciones anteriores se han analizado las dificultades que ha sido preciso superar para disponer de una aplicación operativa. El uso y la experimentación con la herramienta viene mostrando las fortalezas y debilidades de las soluciones adoptadas, y como consecuencia de ello está en constante fase de mejora, afectando actualmente a varios de los algoritmos implementados:

- En primer lugar, la fusión de las cintas de transcripciones manuales se lleva a cabo al alcanzar un umbral en la longitud de la máxima secuencia coincidente al hacer un alineamiento. Está planteado depurar este criterio mediante la consideración de un umbral para una "tasa de acierto" entre todas las posibles subsecuencias alineadas (lo que permitiría considerar correcto el ejemplo de la Figura 5).
- El mecanismo de alineamiento es bastante robusto frente a los errores de transcripción distribuidos uniformemente por todo el texto, pero en caso de darse disparidades importantes entre texto y audio en zonas amplias, se produce un cierto "efecto de borde" que afecta a zonas que de otro modo se ajustarían correctamente. Si este fenómeno llega a producirse de un modo importante, será preciso insertar texto manualmente o borrar zonas de audio. Actualmente se está experimentando con mecanismos automáti-

cos basados en el uso de símbolos capaces de "absorber" las zonas no coincidentes.

- Por otro lado, también en la fase de alineamiento, se está experimentando con el uso de kernels que tienen en cuenta la caracterización de los errores cometidos por el decodificador acústico-fonético, así como con un mecanismo de valoración contextual de inserciones y borrados que podría ser positivo para el alineamiento voz-texto.
- En el apartado del subtitulado también se corrigen puntualmente determinados fenómenos que se observan como mejorables. Actualmente se está incluyendo el tiempo como criterio de corte, en particular para no presentar conjuntamente segmentos cortos excesivamente distantes.

5 Conclusiones

Se ha presentado una aplicación específica desarrollada para facilitar el servicio de subtitulado de vídeos de sesiones plenarias que se está ofreciendo al Parlamento Vasco desde septiembre de 2010. La presentación se ha centrado en la descripción de los problemas encontrados, las soluciones adoptadas y las mejoras que se encuentran en desarrollo y prueba. Respecto a la eficacia de la herramienta, se ha mencionado que los resultados son altamente satisfactorios (está siendo utilizada en producción), remitiendo a Bordel et al. (2012) donde se encuentra una evaluación objetiva del mecanismo de alineamiento que implementa. Del mismo modo, para lo que tiene que ver con detalles concretos de los componentes del sistema, se remite a Bordel et al. (2011).

6 Agradecimientos

Los autores quieren mostrar su agradecimiento a los miembros de la Mesa del Parlamento Vasco que aprobaron el apoyo económico a las iniciativas que dan lugar al trabajo presentado, así como a las personas que apostaron inicialmente por él, y quienes lo hacen posible en la actualidad: Juan Luis Mázmela, Andoni Aia, Juanjo Arruza y Andrés Serrano.

Bibliografía

Alvarez, A., A. del Pozo, y A. Arruti, 2010. Apyca: Towards the automatic subtitling of television content in Spanish. *International Multiconference on Computer Science and Information Technology* IMCSIT, páginas 567–574. Wisla, Polonia, Octubre, 18-20,

Bordel, G., S. Nieto, M. Penagarikano, L. J. Rodríguez Fuentes, y A. Varona, 2011. Automatic subtitling of the Basque parliament plenary sessions videos. *in Twelfth Annual Conference of the International Speech Communication Association, INTERSPEECH*, páginas 1613–1616. Florencia, Italia, agosto 28-31.

Bordel, G., M. Penagarikano, L. J. Rodríguez Fuentes, y A. Varona, 2012. A simple and eficient method to align very long speech signals to acoustically imperfect transcriptions. Páginas *Interspeech*, Portland (OR), EEUU, Septiembre 9-13.

Cerisara, C., O. Mella, y D. Fohr, 2009. JTrans, an open-source software for semi-automatic text-to-speech alignment. *Proceedings of Interspeech*, páginas 1823-1826. Brighton Reino Unido.

Hazen, T. 2006. Automatic alignment and error correction of human generated transcripts for long speech recordings. *Ninth International Conference on Spoken Language Processing, Interspeech-ICSLP*.

Hirschberg, D. 1975. A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, 18(6):341-343.

Moreno, P., C. Joerg, J. Thong, y O. Glickman. 1998. A recursive algorithm for the forced alignment of very long audio segments. *Fifth International Conference on Spoken Language Processing*.

Needleman, S. B., y C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443-453.

Penagarikano M., y G. Bordel, 2005. Sautrela: A highly modular open source speech recognition framework. *Proceedings of the ASRU Workshop*. Páginas pp. 386–391. San Juan, Puerto Rico, December,

Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Lui, G. Moore, J. O. andDave Ollason, D. Povey, V. Valtchev, y P. Woodland, 2006. The HTK Book (for HTK Version 3.4). *Cambridge, UK: CUED, Cambridge University Engineering Department.*