SUBTITULADO EN TIEMPO REAL DE INFORMATIVOS EN DIRECTO PARA LA TELEVISIÓN MEDIANTE RECONOCIMIENTO AUTOMÁTICO DEL HABLA.

Autores:

Alfonso Ortega[1], José Enrique García[1], Eduardo Lleida[1], Emiliano Bernués[2], Daniel Sánchez[3] y Miguel Ferrer[2].

[1] Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza.

[2] Corporación Aragonesa de Radio y Televisión (CARTV)

[3] Aranova.

Resumen:

El subtitulado de informativos, es una importante aplicación que puede facilitar el acceso a la información y la integración de personas sordas o con dificultades en la audición. Sin embargo, el subtitulado en tiempo real de programas emitidos en directo, resulta una aplicación muy costosa que el desarrollo de sistemas completamente automáticos puede ayudar a abaratar.

Existen varias aproximaciones en cuanto al desarrollo de sistemas de subtitulado en tiempo real se refiere que van desde la estenotipia, hasta los

sistemas de subtitulado asistidos por motores de reconocimiento automático del Habla (RAH) [1].

Dentro de este segundo tipo de sistemas, los que hacen uso de sistemas de RAH, podemos encontrar diversos métodos que van desde el uso de sistemas parcialmente asistidos por operadores humanos o sistemas que hacen uso de locutores en la sombra, hasta sistemas que tratan de obtener una transcripción total del audio sin ningún tipo de asistencia humana [2, 3, 4 y 5].

En cuanto a los métodos basados en RAH que requieren de la asistencia de operadores humanos, sus costes pueden ser demasiado elevados para determinadas televisiones de pequeño o mediano tamaño por lo que los sistemas totalmente automáticos suelen ser preferidos. Sin embargo, hoy por hoy esta solución adolece de un conjunto de problemas como son la alta tasa de error del módulo de reconocimiento automático del habla y el elevado retardo en la generación de los subtítulos. [6]

En este trabajo se presenta un método para la generación de subtítulos en tiempo real de forma completamente automática y sin la necesidad de supervisión humana, mediante el uso de técnicas de reconocimiento automático del habla. Dicho sistema utiliza los textos de las noticias obtenidos del sistema informático de redacción realizando el alineamiento temporal entre dichos textos y el audio del informativo a través de un motor de reconocimiento automático del habla. Los subtítulos generados por el módulo de RAH son enviados al gestor del teletexto que los mostrará a través de la página 888.

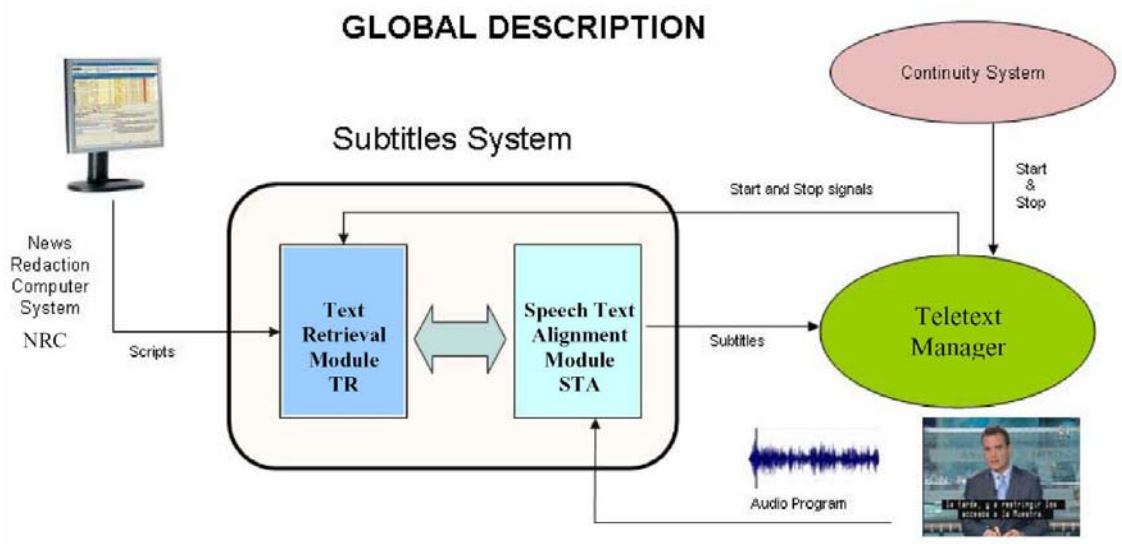

Figura 1. Esquema general de la aplicación de subtitulado en tiempo real.

La implementación está basada en una arquitectura distribuida en la que cada tarea está realizada por un módulo distinto pudiendo estos residir en ordenadores diferentes, realizándose la comunicación entre ellos a través del envío de paquetes XML mediante el protocolo TCP/IP, con la excepción del envío de los subtítulos desde el motor de RAH al gestor del teletexto que hace uso del protocolo UDP por exigencias del fabricante del equipamiento del teletexto. En la figura 1, puede verse un diagrama de bloques de la aplicación de subtitulado en tiempo real para informativos en directo.

A rasgos generales, podríamos decir que el sistema está compuesto por dos módulos principales que son el módulo encargado de recoger los textos de las noticias del sistema informático de la redacción (*Text Retrieval Module TR,* en la figura 1) y el módulo encargado del alineamiento temporal entre el audio y los textos (*Speech-Text Alignment Module*, *STA*, en la figura 1).

El sistema completo está asistido por el sistema de continuidad que envía indicaciones del comienzo y el final del informativo al módulo encargado de la recogida de los textos de las noticias. Por otro lado, la altura de los subtítulos está controlada por la rotuladora de manera que cada vez que se inserta un texto en pantalla, el módulo de generación de los subtítulos recibe dicha información elevando la altura a la que va a presentarse el subtítulo con el objetivo de que la lectura del texto presentado no se vea obstaculizada por los subtítulos de la página 888.

Un aspecto importante a tener en cuenta en la subtitulación de informativos en tiempo real, es la alta tasa de modificaciones que sufren las noticias a lo largo del informativo. Por este motivo el sistema encargado de recuperar el contenido de cada una de ellas, el módulo de recuperación de los textos de las noticias, desarrollado por la empresa Aranova, debe estar constantemente monitorizando el sistema informático de redacción y enviar las modificaciones que encuentre al motor de reconocimiento automático del habla. Este módulo de recuperación de los textos de las noticias envía al módulo de alineamiento temporal la lista completa de las noticias instantes antes del comienzo del informativo. Sin embargo, puesto que constantemente se producen modificaciones en el contenido del informativo (surgen nuevas noticias, se caen algunas noticias existentes o se modifica el texto de las mismas) el módulo TR envía constantemente paquetes de cambios al módulo STA.

El módulo de alineamiento temporal entre el audio y los textos está basado en un motor de Reconocimiento Automático del Habla desarrollado por el grupo de

investigación GTC (Grupo de Tecnologías de las Comunicaciones) del Instituto de Investigación en Ingeniería de Aragón de la Universidad de Zaragoza. Se trata de un sistema de RAH que hace uso de modelos ocultos de Markov continuos (continuos HMM) con unidades acústicas contextuales donde cada unidad está modelada con una mezcla de gaussianas (GMM) de 16 componentes. Los parámetros acústicos empleados son 12 Mel-Frequency Cepstral Coefficients (MFCC) más el coeficiente de energía normalizado junto con la primera y la segunda derivada de éstos. El audio empleado es digitalizado a 16 Khz. con 16 bits por muestra. En cuanto al modelo de lenguaje, tras la recepción del texto de cada noticia, se genera una gramática de estados finitos a la que se le añaden redes de fonemas tras cada pausa y al comienzo de la noticia. El motivo de utilización de la red de fonemas es que en ocasiones las noticias contienen videos, declaraciones de personajes o locuciones del reportero de las que no se tiene el texto. En estos casos, los reporteros que preparan la noticia, marcan el lugar de inserción de los fragmentos sin texto asociado de manera que durante esos pasajes no se enviarán subtítulos al no contar con el texto asociado. De este modo, la red de fonemas será la encargada de modelar acústicamente estos trozos del informativo evitando que la gramática generada con el texto de la noticia progrese erróneamente.

Otro aspecto importante a considerar dentro del sistema presentado es la determinación de la siguiente noticia a subtitular. A pesar de que al comienzo del informativo se cuenta con los textos de las noticias que lo van a componer, debido a los múltiples cambios sobre el informativo o a fallos en la finalización

de la subtitulación de una noticia puede suceder que el sistema de alineamiento pierda momentáneamente el sincronismo y tenga preparada una noticia para subtitular que no se corresponde con la que realmente va a ser locutada. Los motivos de la pérdida de sincronismo pueden ser múltiples, que el presentador no ha dicho el texto de la noticia correctamente, que haya habido una alteración en el orden de las noticias no comunicada, un fallo del motor de reconocimiento, audio de baja calidad,… Para permitir que el subtitulado del resto del informativo se desarrolle correctamente es preciso que ante esas eventualidades el sistema salte a la siguiente noticia a ser subtitulada. Puesto que no se cuenta con información externa fiable y en todos los casos que indique cuál será la siguiente noticia a subtitular, el sistema aquí presentado cuenta con un detector de noticias basado también en un motor de RAH. La decisión se realiza mediante la decodificación acústica del audio del programa de manera que si el reconocedor devuelve un camino en la decodificación suficientemente largo (alrededor de 10 palabras) sobre el texto de alguna de las noticias que siguen a la noticia que presumiblemente iba a ser la próxima en locutarse, se decide que debe producirse un salto en el orden del informativo, yendo a la noticia indicada por el mencionado detector de noticias. Existe asimismo una segunda medida de seguridad para garantizar el correcto seguimiento del orden de las noticias basado en un atributo de las noticias dentro del sistema informático de la redacción. Este atributo cambia de estado (de CUED a PLAY) cada vez que una noticia que contenga una pieza de video pasa a ser la siguiente en entrar, en ese caso, el sistema pasa a considerarla como la siguiente noticia a subtitular. Lamentablemente no todas las noticias cuentan con dicho atributo y el cambio en dicho atributo puede producirse

incluso antes de que la noticia anterior haya finalizado por lo que el empleo de este mecanismo de control debe restringirse a una medida de segundo orden para el caso en el que el detector de noticias no haya funcionado adecuadamente.

El sistema de subtitulado aquí presentado ha sido adoptado para sus informativos por Aragón Televisión, la televisión pública de Aragón y lleva funcionando desde Junio de 2008 con resultados satisfactorios. En concreto, se ha llevado a cabo un estudio de prestaciones recabando información a lo largo de un mes de informativos. En él se ha medido el porcentaje de noticias que han sido completadas con éxito, es decir, sin ningún error; la proporción de noticias que han finalizado parcialmente subtituladas, es decir, alguno de los subtítulos no ha sido enviado y el número de noticias que no ha sido subtitulado en ninguna medida.

Los informativos de Aragón televisión están organizados en tres ediciones diarias, mediodía, tarde y noche. La primera edición, a las 14:00 horas, es la más extensa, con más de 90 noticias por informativo y en la que más cambios se producen por lo que va a ser la edición más difícil de abordar por el sistema presentado. Por otro lado, la segunda edición, a las 20:30 posee una duración de 45 minutos con alrededor de 75 noticias y en ella los cambios, aunque frecuentes, no llegan a un número tan elevado como en la primera edición Excepcionalmente, los domingos, esta edición se ve reducida en su duración, limitándose a unos 15 minutos con unas 20 noticias. Por último, la tercera edición, al filo de las 0:00 horas, es el informativo de menor duración con unas

50 noticias por programa. En ella el número de cambios es relativamente bajo. En la figura 2 se muestra el histograma de noticias por informativo en Aragón Televisión, donde se puede ver gráficamente el reparto de noticias por informativo descrito anteriormente.

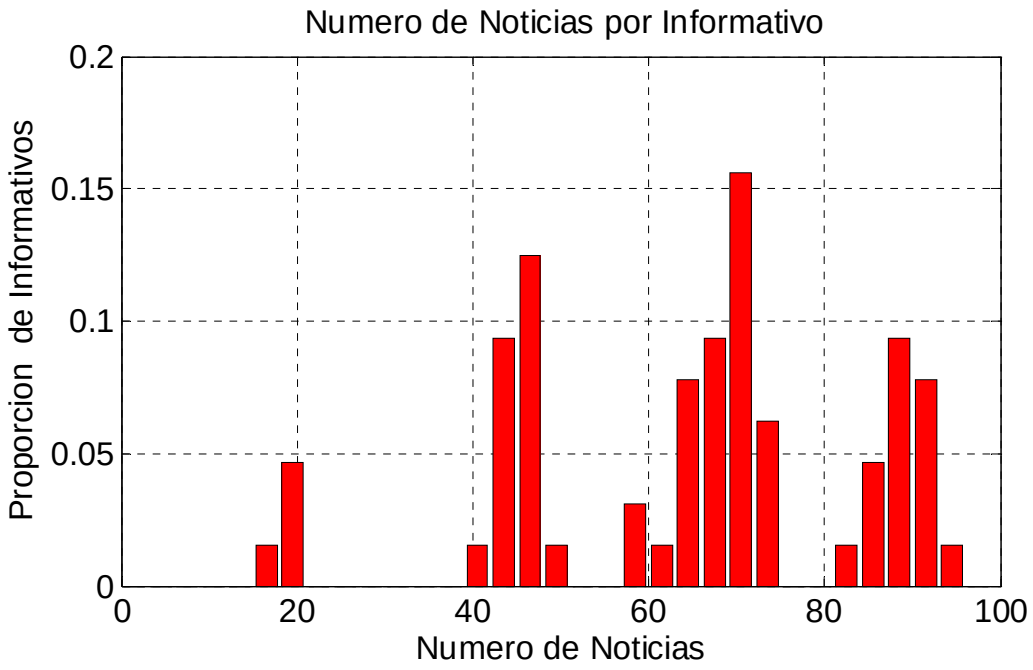

Figura 2. Histograma del número de noticias por informativo.

Gracias a este estudio se ha podido comprobar cómo ningún informativo ha terminado con menos del 75% de noticias totalmente subtituladas y más del 55% ha concluido con más del 90% de las noticias correctamente subtituladas, como puede verse en la figura 3. Ningún noticiario ha concluido con más del 20% de las noticias parcialmente subtituladas y cerca del 70% contienen menos del 10% de las noticias parcialmente subtituladas, como se muestra en la figura 4. Por último, más del 90% de los informativos concluyen con menos del 5% de las noticias sin subtitular, como puede observarse en la figura 5.

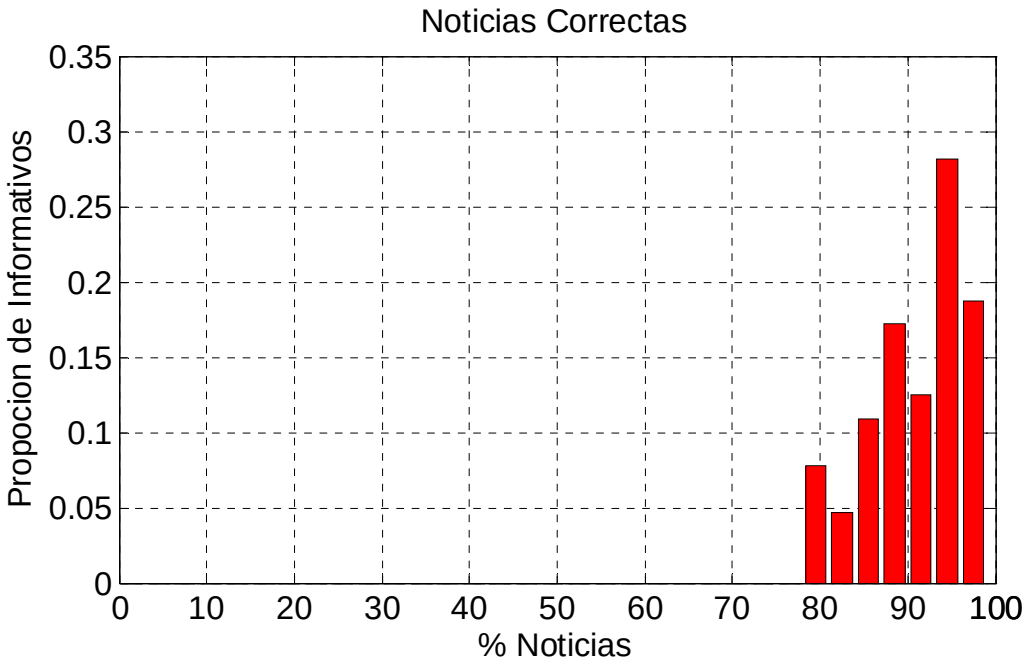

Figura 3. Histograma del porcentaje de noticias subtituladas correctamente.

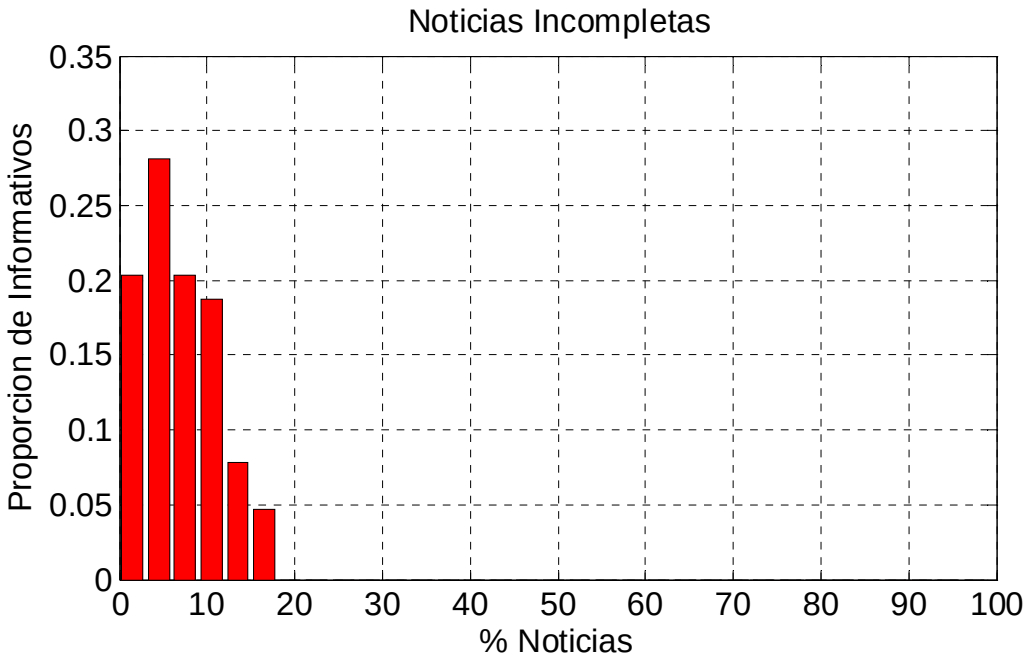

Figura 4. Histograma del porcentaje de noticias subtituladas parcialmente.

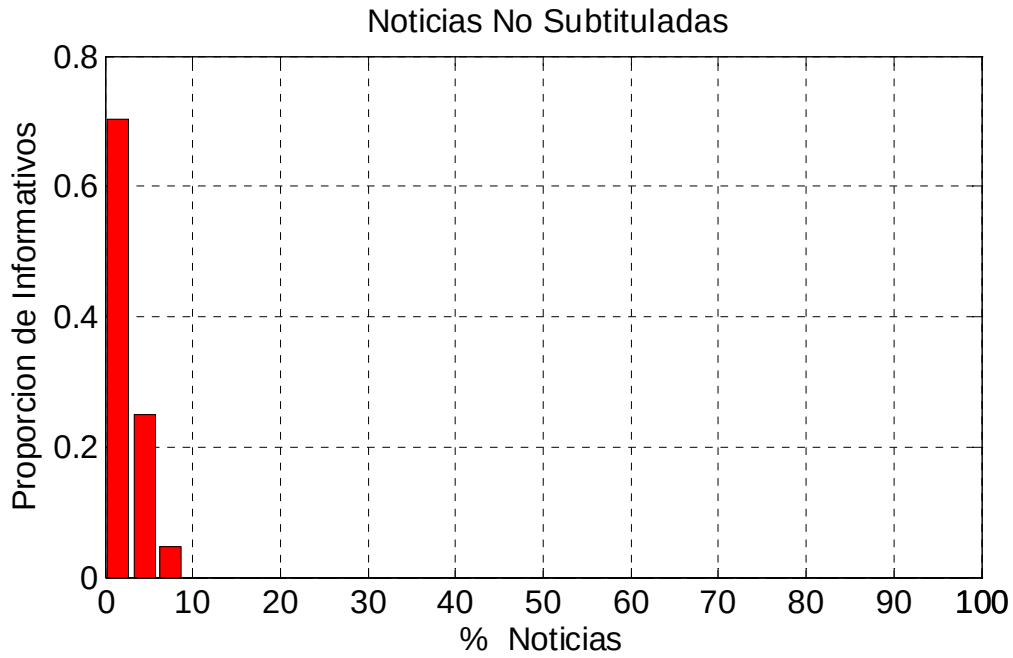

Figura 5. Porcentaje de noticias no subtituladas.

A modo de resumen, en la Tabla 1 se muestra el porcentaje de noticias completadas con éxito, parcialmente subtituladas y no subtituladas para cada una de las ediciones del informativo. En ella podemos ver cómo en media, el 91,52% de las noticias se subtitulan correctamente, quedando un 6,7% de las mismas parcialmente subtituladas y tan solo un 1,78 % sin subtitular. Dado que la primera edición del informativo es en la que más cambios se producen, es en la que las prestaciones del sistema más se degradan aunque manteniéndose en niveles muy elevados, ya que alrededor del 90% de las noticias se subtitulan satisfactoriamente. Por otro lado, la tercera edición es en la que menos cambios se producen y por lo tanto, la que mejores resultados ofrece, llegando hasta cerca de un 95% de noticias correctamente subtituladas y menos de un 1% de las noticias no subtituladas.

EDICIÓN	COMPLETADAS CON ÉXITO	PARCIALMENTE SUBTITULADAS	NO SUBTITULADAS
PRIMERA	88,64%	8,55%	2,81%
SEGUNDA	91,50%	6,90%	1,60%
TERCERA	94,42%	4,65%	0,93%
MEDIA	91,52%	6,70%	1,78%

En cuanto al retardo en la generación de los subtítulos del sistema presentado, en ocasiones denominado latencia, podríamos decir que es despreciable e imperceptible debido al hecho de que los subtítulos se envían al gestor del teletexto tan pronto como el motor de reconocimiento automático del habla detecta la primera de las palabras del subtítulo. Únicamente un pequeño retardo intencionado se añade al comienzo de cada noticia y tras la emisión de cada fragmento sin texto asociado para evitar la emisión de subtítulos de forma anticipada debido a las locuciones sin texto asociado.

Como conclusión, podríamos decir que el sistema aquí presentado permite la subtitulación en tiempo real de informativos emitidos en directo a través del uso de un motor de reconocimiento automático del habla. Éste recibe del sistema informático de la redacción los textos de las noticias y realiza el alineamiento temporal entre el audio y los textos. Posteriormente, los subtítulos correctamente sincronizados con el audio se envían al gestor del teletexto que los mostrará a través de la página 888. Un estudio estadístico de las prestaciones nos revela que la precisión del sistema puede llegar a un 91.5 % de las noticias correctamente subtituladas con menos de un 2% de las noticias no subtituladas. Además, podríamos decir que el sistema cuenta con un retardo

imperceptible debido a la rapidez en la generación de los subtítulos del módulo de alineamiento. El sistema está funcionando en Aragón Televisión desde Junio de 2008 satisfactoriamente.

Bibliografía:

[1] Virginia Fuentes Bueno, Israel González Carrasco y Belén Ruiz Mezcua, “Subtitulado en Tiempo Real. Sistemas y Tecnología”, I Congreso de Accesibilidad a los Medios Audiovisuales para personas con Discapacidad AMADIS’06. Madrid, Julio 2006.

[2] J. Brousseau, J.F. Beaumont, G. Boulianne, P. Cardinal, C. Chapdelaine, M. Comeau, F. Osterrath and P. Ouellet, “Automated closed-captioning of live TV broadcast news in French”, in Proceedings of Eurospeech 2003, Geneva, Switzerland, 2003.

[3] G. Boulianne, F.F. Beaumont, M. Boisvert, J. Brousseau, P. Cardinal, C. Chapdelaine, M. Comeau, P. Ouellet, F. Osterrath, “Computer-assisted closed-captioning of live TV broadcasts in French”, in Proceedings Interspeech 2006, Pittsburgh, USA, 2006.

[4] A. Ando, T. Imai, A. Kobayashi, H. Isono, and K. Nakabayashi, “Real-Time Transcription System for Simultaneous Subtitling of Japanese Broadcast News Program”, IEEE Transactions on Broadcasting, Vol. 46, No. 3, September 2000.

[5] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, “Broadcast news subtitling system in Portuguese”, in Proc. ICASSP 2008, Las Vegas, USA, 2008.

[6] H. Meinedo, M. Viveiros, J. Paulo and S. Neto, “Evaluation of a Live Broadcast News Subtitling System for Portuguese”, in Proc of Interspeech 2008, Brisbane, Australia, 2008.

Currículo de los autores:

Alfonso Ortega:

Ingeniero de Telecomunicación y doctor por la Universidad de Zaragoza, premio extraordinario de doctorado y premio Cátedra Telefónica. Es profesor contratado doctor en la Universidad de Zaragoza, centrando su actividad en las Tecnologías del Habla. Ha participado en más de 30 proyectos de investigación, es autor de más de 50 artículos en revistas o congresos científicos internacionales de reconocido prestigio y es inventor de una patente.

Edificio Ada Byron, María de Luna 1. 50018, Zaragoza

Tlfno. +34 976 76 2363

ortega@unizar.es

José Enrique García:

Ingeniero de Telecomunicación por el Centro Politécnico Superior (Universidad de Zaragoza) desde 2006. Actualmente trabaja como ingeniero de proyectos en el ámbito de las tecnologías del habla dentro del Grupo de Tecnologías de las Comunicaciones GTC del Instituto de Investigación en Ingeniería de Aragón de la Universidad de Zaragoza.

Edificio Ada Byron, María de Luna 1. 50018, Zaragoza

Tlfno. +34 976 76 2705

jegarlai@unizar.es

Eduardo Lleida

Doctor Ingeniero de Telecomunicación por la Universidad Politécnica de Cataluña (1990). Catedrático de Universidad en el Centro Politécnico Superior de la Universidad de Zaragoza. Sus líneas de trabajo se centran en las tecnologías del habla. Ha participado en más de 60 proyectos y contratos de I+D. Es autor de más de 100 artículos publicados en revistas, capítulos de libro, congresos internacionales y nacionales. Es inventor en 4 patentes.

Edificio Ada Byron, María de Luna 1. 50018, Zaragoza

Tlfno. +34 976 76 2372

lleida@unizar.es

Emiliano Bernués:

Ingeniero de Telecomunicación y Doctor por la Universidad Politécnica de Madrid. Es profesor de la Universidad de Zaragoza y ha sido director técnico de la Corporación aragonesa de radio y televisión, donde se encargó de la puesta en marcha de ambos medios. Actualmente es Director Gerente de Aragón Telecom, empresa pública encargada de gestionar el despliegue de la TDT en Aragón.

ARAGON TELECOM. World Trade Center, Torre Este, planta15, of. B.

Avd María Zambrano 31. 50018 Zaragoza

Tfno: 620 849773

ebernues@aragontelecom.es

Daniel Sánchez

Es técnico superior en desarrollo de aplicaciones informáticas de gestión. Su actividad profesional se ha desarrollado en la empresa IASoft durante 6 años centrándose en desarrollo de aplicaciones para la administración pública y en la espesa Aranova durante 4 años desarrollando soluciones audiovisuales.

ARANOVA. Valle de Broto 9-11, 1ª planta, oficina 4 – 50015, Zaragoza.

Tfno: 627525660

daniel.sanchez@aranova.es

Miguel Ferrer:

Ingeniero de Telecomunicación por el Centro Politécnico Superior (Universidad de Zaragoza) desde 2009. Inicia su carrera profesional como ingeniero en el Departamento Técnico de Corporación Aragonesa de Radio y Televisión, por medio del que ha sido becado durante el año 2008. Su trabajo se mueve en el ámbito de nuevos proyectos audiovisuales, siendo el responsable del HDLAB dirigido a la implantación de tecnologías en Alta Definición.

Corporación Aragonesa de Radio y Televisión (CARTV)

Av. María Zambrano, 2. 50018, Zaragoza

Tlfno. +34 876 256613

mferrer@cartv.es