Tecnologías del habla: nuevas oportunidades para los archivos de televisión.

Virginia Bazán-Gil[1], Eduardo Lleida[2], Carmen Pérez[3], Manuel Gómez[4], Alberto de Prada[5],

[1] https://orcid.org/0000-0003-4920-2212 + Fondo Documental RTVE, Madrid, España. virginia.bazan@rtve.es.
[2] https://orcid.org/0000-0001-9137-4013 + Instituto de Investigación en Ingeniería de Aragón. Universidad de Zaragoza, Zaragoza, España. lleida@unizar.es.
[3] https://orcid.org/0000-0002-3325-0546 + Innovación y Estrategia Tecnológica RTVE, Madrid, España. carmen.perez@rtve.es.
[4] https://orcid.org/0000-0001-5128-5905 + Área de Desarrollo Digital RTVE, Madrid, España. manuel.gomez@rtve.es
[5] https://orcid.org/0000-0002-9140-5745 + Fondo Documental RTVE, Madrid, España. alberto.deprada@rtve.es.

Tipo de contribución: Comunicación

Palabras clave: Archivos audiovisuales, tecnologías del habla, metadatado automático

1. Inteligencia Artificial para el análisis de contenidos audiovisuales

La sobreabundancia de contenidos audiovisuales y la dificultad creciente para identificar y describir estos contenidos de forma eficiente ha convertido a la Inteligencia Artificial en un objeto de deseo para los archivos de televisión (Bazán Gil & Guerrero Gómez-Olmedo, 2018). Los futuros procesos de generación automática de metadatos en los archivos se fundamentarán en tres tecnologías complementarias: visión artificial, tecnologías del habla y procesamiento del lenguaje natural. La visión artificial se ocupa del reconocimiento de imágenes, la agrupación y segmentación de escenas y el seguimiento de objetos y personas. Las tecnologías del habla, por su parte, permiten el reconocimiento tanto de voces como de hablantes y sus emociones, así como la segmentación del audio en función de estos hablantes. De forma complementaria, el procesamiento del lenguaje natural permite el reconocimiento de entidades y temas tratados en un discurso y facilita la creación de resúmenes y descripciones, aportando así nuevos metadatos para el archivo (Lleida, 2018). En otras palabras, la visión artificial proporciona la descripción de la capa de imagen, especialmente relevante en los materiales originales en los que el sonido es únicamente ambiente; las tecnologías del habla permiten convertir la capa de audio en un texto que adquiere relevancia tanto en la fase de edición y montaje de un programa, como en las fases posteriores de emisión y difusión a través de la Web (subtitulado); y finalmente, el procesamiento del lenguaje natural contribuye a enriquecer con metadatos el material en la fase de archivo, generando puntos de acceso en forma de entidades para la recuperación.

La aplicación de estas tecnologías no solo facilitará el acceso a un volumen creciente de contenidos audiovisuales, sino que además permitirá alcanzar un nivel de detalle en el análisis hasta ahora impensable en los archivos de televisión (Bazán Gil, 2018). En este nuevo horizonte, las funciones

esenciales de los documentalistas se verán una vez más alteradas, ya que los procesos automatizados requerirán de su implicación en las fases de entrenamiento de los algoritmos en los que se basan estas tecnologías, así como en el control de la calidad de los datos generados de forma automática.

2. Las tecnologías del habla y su aplicación en televisión

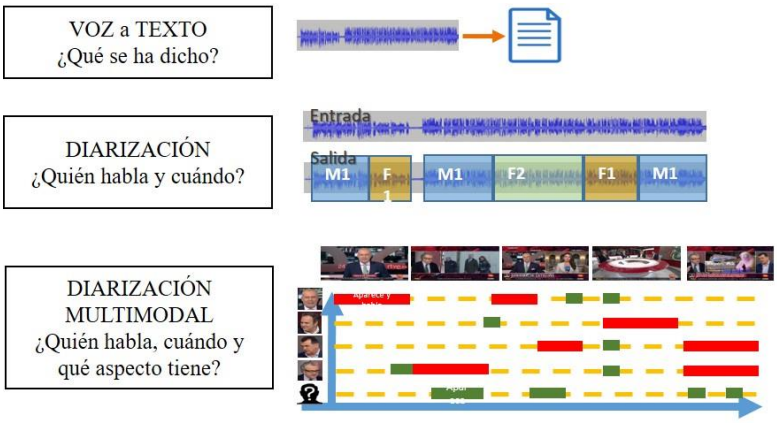
De las tecnologías mencionadas en el epígrafe anterior, las del habla, ya han sido adoptadas por las cadenas de televisión. Gracias a ellas es posible la generación de subtítulos para garantizar la accesibilidad de las personas con discapacidad auditiva a los contenidos audiovisuales. Más del 94% de los contenidos emitidos por RTVE están subtitulados. Los subtítulos, generados por el Área de Accesibilidad están disponibles tanto en la emisión en TDT como en la redifusión en la Web y, en un futuro muy próximo, se incorporarán al Archivo como un punto de acceso más a los contenidos, facilitando con ello las búsquedas por palabras y la recuperación de secuencias concretas dentro de un programa determinado.

Aún más, la transcripción de voz a texto puede realizarse sobre los rodajes originales que se utilizan en la edición de un programa de televisión, evitando a los redactores la tediosa labor de transcribir manualmente las entrevistas. En esta línea, el Área de Innovación y Estrategia Tecnológica y el Fondo Documental RTVE trabajan, junto al equipo del programa Crónicas, en un proyecto piloto para la transcripción automática de entrevistas con sistemas de transcripción de voz a texto. Este proyecto no sólo supondrá una optimización del tiempo dedicado a la producción del programa, sino que aumentará sustancialmente la accesibilidad de un material, los rodajes, que habitualmente cuentan con escasos metadatos asociados cuando llegan al archivo.

El uso de esta tecnología, tanto en la producción como en la emisión, garantiza la creación de una capa de información relevante para la preservación y reutilización de los contenidos. Es fundamental, no solo generar los datos, sino establecer además las sinergias entre áreas y los flujos de trabajo necesarios para que estos datos acompañen a los contenidos en su viaje hacia el archivo. En este sentido son relevantes ya que constituyen puntos de acceso a los contenidos preservados.

Como veremos con más detalle, las principales tecnologías implicadas en estos procesos son la transcripción de voz a texto (*Speech to text*) y la diarización o segmentación y agrupación por hablantes. Los sistemas de voz a texto nos permiten conocer qué se ha dicho mientras que la diarización nos ayudará a identificar quién habla, cuándo habla y qué aspecto tiene si aplicamos, además, técnicas de reconocimiento facial tal y como muestra la figura 1.

Figura 1: Tecnologías del habla aplicadas a archivos de televisión



3. RTVE Ibersepeech 2018: el reto tecnológico de la Cátedra RTVE - Universidad de Zaragoza

En este contexto, en el que la Inteligencia Artificial brinda a los archivos de televisión nuevas oportunidades, la Corporación RTVE y la Universidad de Zaragoza formalizaron en julio de 2017 la creación de una Cátedra cuyo principal objetivo es la realización de actividades de formación, investigación, estudio y divulgación de las Tecnologías de la Información y de las Comunicaciones relacionadas con el Big Data y su aplicación al análisis de contenidos audiovisuales y sonoros (Cátedra RTVE Universidad de Zaragoza, 2017). En ella, junto al equipo de la Universidad de Zaragoza colaboran distintas áreas de RTVE como son: el Fondo Documental, el área de Innovación y Estrategia Tecnológica y el Área de Desarrollo Digital.

Como parte de sus actividades, y con el objetivo de impulsar la investigación en tecnologías del habla en español, en mayo de 2018 la Catedra lanzó el RTVE Iberspeech Challenge 2018 (RTVE, 2018). Un reto tecnológico que puso a disposición de la comunidad científica más de 500 horas de contenidos emitidos junto a sus correspondientes subtítulos, además de los subtítulos de los programas de producción propia emitidos por el Canal 24h a lo largo de 2017 (Lleida, Ortega, Miguel, Bazán, Pérez, Zotano, et al., 2018)

El conjunto de datos liberado, conocido como base de datos RTVE2018, incluye cerca de una veintena de programas de distintos géneros y temática, producidos y emitidos por RTVE entre 2015 y 2018. Estos programas presentan distintas dificultades desde el punto de vista de las tecnologías del habla como son: la diversidad de acentos del español, la superposición de diálogos, el habla espontánea, la variabilidad acústica, el ruido de fondo o el vocabulario específico, factores todos ellos que inciden en el rendimiento de los sistemas de transcripción.

Sobre este conjunto de programas, los 22 grupos participantes debían detectar automáticamente, etiquetar y transcribir segmentos de habla. Para ello se definieron tres tareas distintas: Voz a texto, diarización y diarización multimodal. En las tareas de voz a texto y diarización los participantes podían optar por dos modalidades de participación: entrenamiento cerrado y abierto. El entrenamiento cerrado suponía entrenar el sistema únicamente con los datos proporcionados en la base RTVE2018, mientras que el entrenamiento abierto permitía utilizar cualquier otra base de datos documentada. En la tarea de diarización multimodal, los participantes contaban con imágenes y vídeos adicionales de cada una de las personas que debían reconocer.

Transcripción de voz a texto

En esta tarea los sistemas de reconocimiento debían transcribir cerca de 39 horas de emisión correspondientes a los programas: "Al filo de lo imposible" (AFI), "Arranca en Verde" (AV), "Dicho y hecho" (DH), "España en Comunidad" (EC), "La mañana" (LM), "La tarde en 24h. Tertulia" (LT24HTer), "Latinoamérica en 24H" (LA24H) y "Saber y ganar" (SG) (Lleida, Ortega, Miguel, Bazán, Pérez, Gómez, et al., 2018b). El rendimiento de los sistemas se calculó en función de la tasa de error por palabras (WER) comparando la transcripción automática con su transcripción real:

$$WER = \frac{N_{INS} + N_{BOR} + N_{SUS}}{N}$$

donde N_{INS} es el número de inserciones (palabras reconocidas que no se han pronunciado), N_{BOR} es el número de borrados (palabras no reconocidas pero que se han pronunciado), N_{SUS} es el número de sustituciones (palabras reconocidas equivocadamente) y N el número total de palabras a reconocer.

Diarización de hablantes

Esta tarea supone segmentar los documentos de audio en función de los hablantes y agrupar todos los segmentos correspondientes a un mismo hablante, sin conocer a priori ni el número de hablantes ni su identidad. Los participantes pusieron a prueba sus sistemas con aproximadamente 22 horas de emisión correspondientes a los programas "España en Comunidad", "La mañana", "La tarde en 24h. Tertulia" y "Latinoamerica en 24H"(Ortega et al., 2018). En este caso, el rendimiento de los sistemas se midió en función del Error de Diarización (DER) comparando la segmentación obtenida por los sistemas con una segmentación real supervisada. El DER contabiliza la fracción de tiempo que no se atribuye correctamente a un hablante específico:

$$DER = \frac{T_{PER} + T_{FA} + T_{HAB}}{T_{VOZ}}$$

donde T_{PER} es la cantidad de voz considerada como no voz (silencio, música, etc.), T_{FA} es la cantidad de no voz considerada como voz, T_{HAB} es la cantidad de voz asignada a un hablante equivocado y T_{VOZ} la cantidad total de voz a evaluar.

Diarización multimodal

La diarización multimodal permite segmentar contenidos audiovisuales en función de los hablantes y enlazar los segmentos a una misma voz y a una misma cara. En este caso los personajes (39) eran previamente conocidos. (Lleida, Ortega, Miguel, Bazán, Pérez, Gómez, et al., 2018a) Los sistemas se pusieron a prueba sobre 4 horas de emisión de los programas "La tarde en 24h. Tertulia" y "La mañana", y su rendimiento se midió en función de la tasa de Error de Diarización (DER) en la segmentación de hablantes y de caras.

4. Resultados de Iberspeech 2018

Transcripción de voz a texto

En esta tarea, y para la condición de entrenamiento abierto, participaron 7 equipos internacionales con 14 sistemas de reconocimiento, basados tanto en sistemas comerciales, como tecnologías propietarias (Jorge et al., 2018) o en tecnologías *open source* como Kaldi (Povey et al., 2011) o

DeepSpeech 2 (Amodei et al., 2016). El conjunto de horas empleado para el entrenamiento de los sistemas osciló entre las 109 y las 3800 horas.

En la figura 2 se muestra la distribución de tasas de error por programas para los 14 sistemas evaluados (Iberspeech, 2018). Como era de esperar existe una gran dependencia del tipo de programa. Programas con entrevistas o conversaciones espontaneas y con audio en exteriores como “Dicho y Hecho”, “La Mañana” o “Arranca en Verde” son los programas que suponen el mayor reto con tasas de error muy por encima del 20% para la mayoría de los sistemas. El sistema mejor evaluado obtiene una tasa de error del 16,45% y el peor evaluado un 35,80%

La figura 3 muestra la distribución de las tasas de error por programas para el sistema ganador. Cabe destacar que para una gran mayoría de programas mantiene unas tasas de error por debajo del 20% e incluso para un programa como “Latinoamérica en 24H”, con distintos acentos del español, la tasa de error se sitúa por debajo del 10%. Este sistema ha sido entrenado con 3.800 horas de voz transcrita de vídeos de distintas fuentes de español tanto peninsular como de América Latina.

Figura 2. Distribución de la tasa de error WER de los 14 sistemas evaluados por programas.

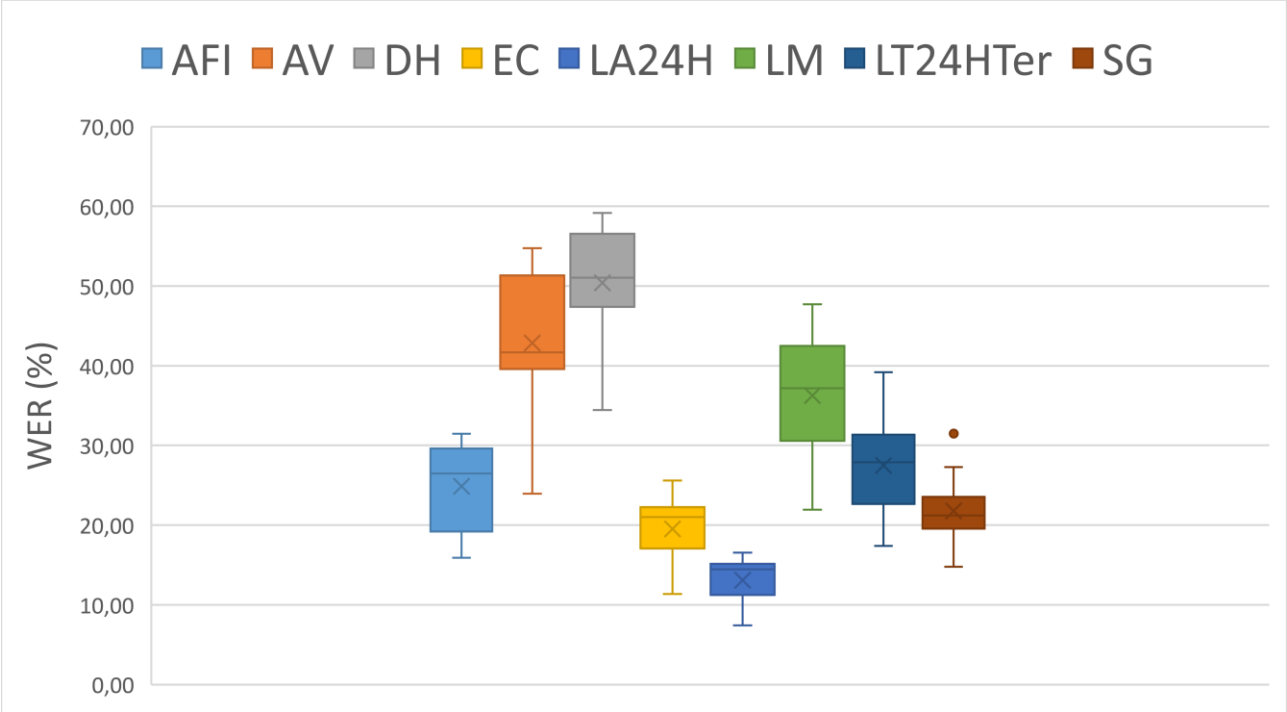
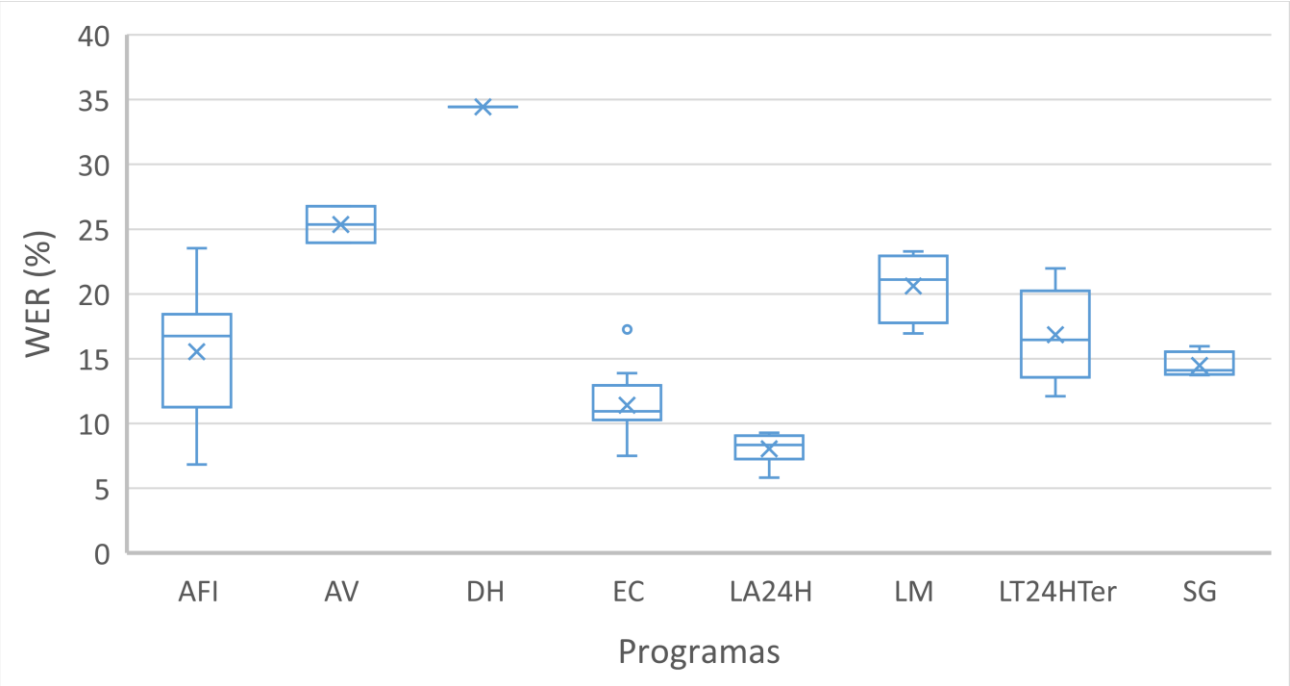


Figura 3. Distribución de la tasa de error WER por programas para el mejor sistema



En la condición de entrenamiento cerrado han participado 3 grupos con 6 sistemas. Las tasas de error conseguidas en esta modalidad se sitúan entre el 19,57% y el 26,66%. Se trataba de una condición más exigente ya que los sistemas contaban únicamente con los subtítulos y no con las transcripciones reales, hecho que ha incidido en el rendimiento al igual que el número limitado de horas de entrenamiento. Finalmente, dos equipos enviaron resultados incluyendo signos de puntuación (puntos y comas) con tasas de error incrementadas un 25% relativo.

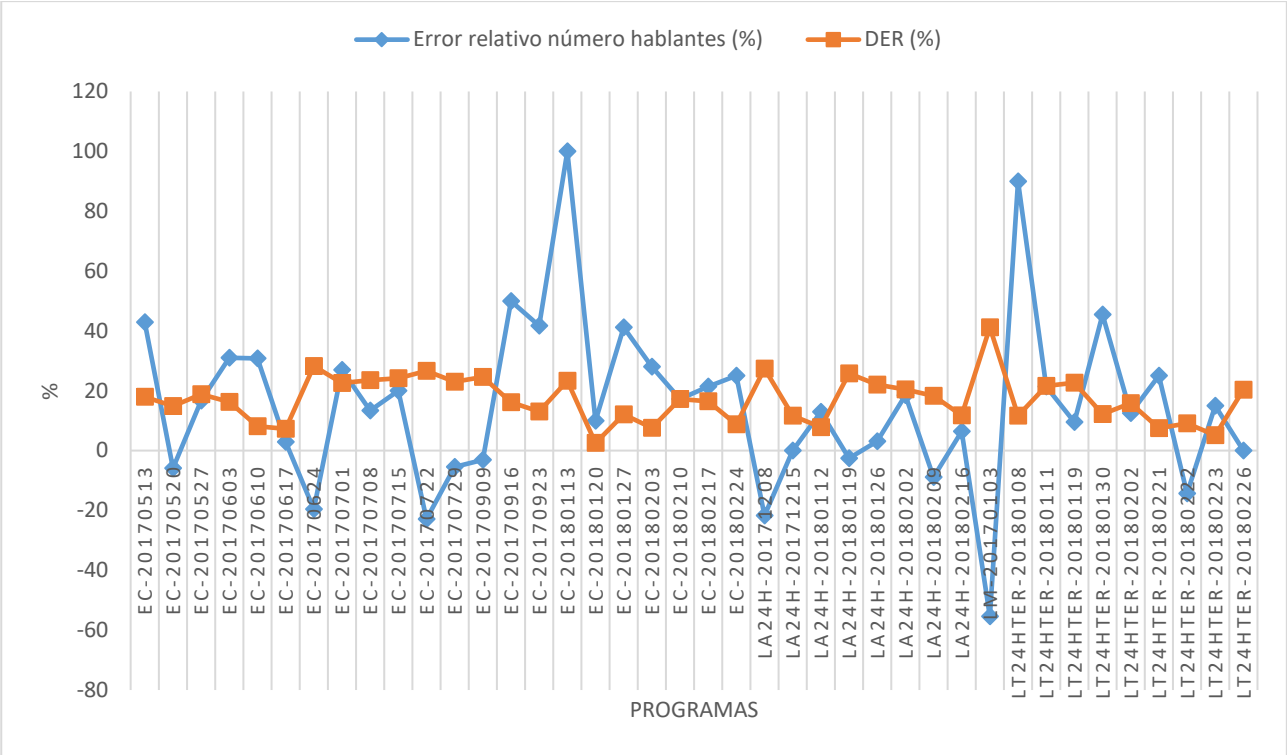
Diarización de hablantes

En el reto de diarizacion de hablantes han participado 8 equipos internacionales, 6 bajo la condición cerrada y 4 bajo la abierta. Todos los sistemas son propietarios de cada equipo participante (Iberspeech, 2018). En la condición cerrada, donde únicamente se podían utilizar los datos proporcionados en la base de datos RTVE2018, las tasas DER se sitúan entre el 17.27% del mejor sistema y el 39,09% del sistema con peores prestaciones. En la condición abierta las tasas de DER se sitúan entre el 26% y el 31%.

Para el sistema con menor tasa DER, el 13,7% del DER se corresponde con la asignación incorrecta del hablante, es decir, un segmento de voz de un hablante se ha asignado a otro, el 1,1% se corresponde con hablantes no detectados y un 2,5% a segmentos sin hablantes que se han asignado a alguno de ellos. Examinado los errores en la detección del número de hablantes, encontramos que el mejor sistema tiene de media un error relativo del 23,46%, siendo en un 70% sobreestimaciones del número de los que hablan .

La figura 4 muestra el DER y el error relativo en la detección del número de personas que hablan por programa. Se puede ver que un error grande en la estimación del número estas no implica un DER elevado ya que en muchas ocasiones los hablantes no detectados, o detectados de más, se corresponden con segmentos de duración muy corta.

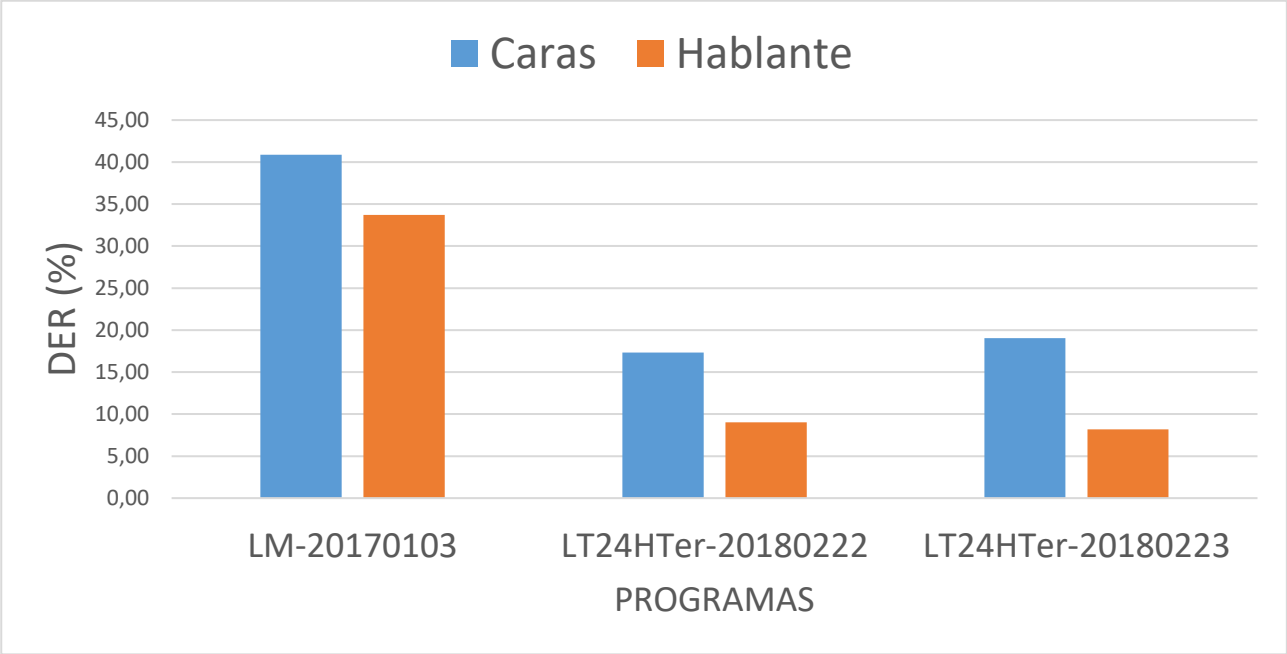
Figura 4. Distribución de la tasa de error DER y error relativo en el número de hablantes



Diarización multimodal

En la tarea de diarización multimodal han participado 4 equipos internacionales y se han evaluado 10 sistemas todos ellos basados en tecnología propietaria (Iberspeech, 2018). El reto consistía no solo en agrupar segmentos correspondientes al mismo hablante/cara sino que además debían asociarlo a alguno de los 39 personajes definidos. La figura 5 presenta los mejores resultados DER de caras y hablantes por programas, donde destacan las mejores prestaciones de la segmentación de hablantes sobre la de caras en los tres programas y la clara diferencia entre un programa de variedades como es el de “La mañana”, sobre uno de tertulias en estudio como es “La Tarde en 24H, Tertulia”.

Figura 5. Mejores DER caras y hablante por programa



Conclusiones

Uno de los objetivos de la Cátedra RTVE de la Universidad de Zaragoza es la evaluación de tecnologías aplicables al análisis de contenidos audiovisuales. Con este propósito se ha organizado el primer reto a nivel internacional para evaluar sistemas de reconocimiento automático del habla, diarización de hablantes y multimodal para la lengua española en el contexto de programas de televisión de contenido diverso. Se ha conseguido una amplia participación, con un total de 22 de grupos de origen tanto nacional como internacional que han valorado muy positivamente la organización de un reto tecnológico como el propuesto. Los resultados muestran las dificultades tecnológicas que el reto suponía. Las tasas de error en transcripción de voz a texto varían, para el mejor sistema, entre un 6% y un 35% dependiendo del contenido del programa.

En relación a la diarización tanto de hablante como multimodal, los resultados obtenidos reflejan la dificultad en estimar el número de hablantes, la variabilidad en las prestaciones con el tipo de contenido y el mejor comportamiento de la diarización de hablante frente a la de cara.

Como resultado del esfuerzo de organización del reto, se pone a disposición de la comunidad científica una base de datos de contenido audiovisual con más de 550 horas de programación y con los protocolos de entrenamiento, desarrollo y evaluación para la comparativa de sistemas de reconocimiento automático del habla, diarización de hablantes y multimodal.

Estos resultados deben abordarse además desde la perspectiva del usuario, es decir desde la utilidad de las tecnologías para su integración en los flujos de trabajo de una organización como RTVE.

En los últimos 5 años los reconocedores de voz han mejorado de forma espectacular. Aun así, todavía existe una gran variabilidad en la tasa de error, entre un 5 % y un 50%, que limita su utilidad en todos los escenarios posibles: desde el material original hasta los programas emitidos. Los factores que inciden en el buen funcionamiento de los sistemas son múltiples, destacando el entorno acústico, la

expresión oral del hablante y el contexto léxico-sintáctico. Tanto los resultados de Iberspeech, como las pruebas realizadas con el equipo del programa "Crónicas", ponen de manifiesto que donde el oído humano no llega los reconocedores tampoco aciertan. Por ejemplo, las entrevistas en grupo con un único micrófono en las que predominan expresiones locales, muletillas, palabras incorrectas, errores fonéticos, sintácticos y léxicos constituyen un verdadero reto tanto para los reconocedores como para el oído humano.

En estos momentos en RTVE se están poniendo a prueba las tecnologías del habla en tres escenarios distintos: la producción, la emisión (tanto en Televisión como en la Web) y el Archivo. De estas pruebas se obtienen varias conclusiones:

- La importancia no solo de generar metadatos en función de las necesidades de las distintas áreas sino de establecer los mecanismos adecuados para garantizar la reutilización de estos datos en la cadena de producción – emisión – archivado.

- La correlación entre tipo de programa y resultado de la transcripción. En este sentido, y desde el punto de vista del archivo, es necesario comprender el funcionamiento de la tecnología y bajo qué premisas tiene un mayor rendimiento. Este conocimiento debe ponerse en relación con las necesidades reales de trabajo y con las expectativas de las distintas áreas en cuanto la calidad de los resultados.

- El grado de tolerancia respecto a los errores varía en función del escenario para el que se generan los metadatos. Tasas de error por encima del 15% pueden ser tolerables para el archivo cuando se carece de otros datos para la recuperación, pero no son admisibles desde el punto de vista de la emisión de los contenidos.

Para los archivos audiovisuales la incorporación de las tecnologías del habla a sus procesos diarios es más que una prioridad, pero la pregunta es ¿por dónde empezar? En un escenario ideal, en el que estas herramientas estén ya integradas en los flujos de trabajo, el Archivo recibirá el subtitulado generado por el Área de Accesibilidad lo que, junto con los metadatos generados tanto por la producción del programa como por el área de emisiones, proporcionará suficiente información para la recuperación y reutilización de los contenidos emitidos. En lo que al material original se refiere, el archivo recibirá del equipo de producción del programa, las transcripciones con las correcciones realizadas por los redactores.

Pero ¿qué hacer con el material original o emitido preservado hasta ahora con unos metadatos mínimos en el archivo? ¿Qué criterios debemos establecer para su tratamiento con sistemas de reconocimiento automático? Si usamos la tecnología indiscriminada corremos el riesgo de incurrir en un gasto en recursos económicos, tecnológicos y humanos sin la garantía de obtener unos resultados óptimos. No basta con disponer de la tecnología y usarla, es necesario definir primero cómo se va a aplicar. Pero ¿Qué criterios debemos emplear? ¿Los relacionados con su uso futuro como producción, reemisión o comercialización de los contenidos? ¿Aquellos relacionados con la posible fidelidad de la transcripción? En otras palabras ¿Es mejor tener información, aunque no sea exacta, que no tener ninguna?

Estas son las preguntas a las que debemos dar respuesta si queremos integrar de manera eficaz estas tecnologías en el Archivo.

Referencias

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., … Zhu, Z. (2016). *Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin*. Retrieved from http://proceedings.mlr.press/v48/amodei16.pdf

Bazán Gil, V. (2018). El renacimiento de los archivos: inteligencia artificial y semántica aplicada a la descripción de contenidos audiovisuales. Retrieved January 9, 2019, from https://es.slideshare.net/Artium_Vitoria/ix-encuentros-de-centros-de-documentacin-de-arte-contemporneo-en-artium-virginia-bazn-gil

Bazán Gil, V., & Guerrero Gómez-Olmedo, R. (2018). Descripción automática de archivos audiovisuales: NeuralTalk, un modelo de video2text aplicado al archivo de RTVE Cita recomendada. *BiD: Textos Universitaris de Biblioteconomia i Documentació*, (41). https://doi.org/10.1344/BiD2018.41.7

Cátedra RTVE Universidad de Zaragoza. (2017). Cátedra RTVE de la Universidad de Zaragoza. Retrieved January 9, 2019, from http://catedrartve.unizar.es/

Iberspeech. (2018). Iberspeech 2018. In *Iberspeech 2018*. Barcelona. Retrieved from https://www.isca-speech.org/archive/IberSPEECH_2018/

Jorge, J., Martínez-Villaronga, A., Golik, P., Giménez, A., Albert Silvestre-Cerdà, J., Doetsch, P., … Sanchis, A. (2018). MLLP-UPV and RWTH Aachen Spanish ASR Systems for the IberSpeech-RTVE 2018 Speech-to-Text Transcription Challenge. In *Iberspeech 2018* (pp. 257–261). Barcelona. https://doi.org/10.21437/IberSPEECH.2018-54

Lleida, E. (2018). Tecnologías para al análisis y metadatado de contenidos audiovisuales. Retrieved January 9, 2019, from http://www.rtve.es/contenidos/documentos/instituto/4_Jornada_Archivos_tv.pdf

Lleida, E., Ortega, A., Miguel, A., Bazán, V., Pérez, C., Gómez, M., & De Prada, A. (2018a). *Albayzin Evaluation: IberSPEECH-RTVE 2018 Multimodal Diarization Challenge*.

Lleida, E., Ortega, A., Miguel, A., Bazán, V., Pérez, C., Gómez, M., & De Prada, A. (2018b). *Albayzin Evaluation: IberSPEECH-RTVE 2018 Speech to Text Transcription Challenge*.

Lleida, E., Ortega, A., Miguel, A., Bazán, V., Pérez, C., Zotano, M., & De Prada, A. (2018). *RTVE2018 Database Description*.

Ortega, A., Viñals, I., Miguel, A., Lleida, E., Bazán, V., Pérez, C., … De Prada, A. (2018). *Albayzin Evaluation: IberSPEECH-RTVE 2018 Speaker Diarization Challenge*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., … Vesely, K. (2011). The kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 1–4. https://doi.org/10.1017/CBO9781107415324.004

RTVE. (2018). La Cátedra RTVE en la Universidad de Zaragoza presenta su primer reto tecnológico a la comunidad científica - RTVE.es. Retrieved January 9, 2019, from http://www.rtve.es/rtve/20180521/catedra-rtve-universidad-zaragoza-presenta-su-primer-reto-tecnologico-comunidad-cientifica/1737360.shtml