UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

PROYECTO FIN DE CARRERA

ALINEAMIENTO DE AUDIO Y TEXTO PARA EL APRENDIZAJE DEL IDIOMA INGLÉS

Ingeniería de Telecomunicación

Darwin Patricio Córdova Lucero

Septiembre 2012

Alineamiento de texto y audio para el aprendizaje del idioma inglés.

AUTOR: Darwin Patricio Córdova Lucero

TUTOR: Doroteo Torre Toledano

Área de Tratamiento de Voz y Señales (ATVS)

Dpto. de Tecnología Electrónica y de las Comunicaciones

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Septiembre 2012

RESUMEN

El propósito de este proyecto es obtener un sistema cuyo resultado se aplique directamente sobre un método de aprendizaje del idioma inglés implementado en el sitio web www.inglesdivino.com. Dicho método consiste básicamente en aprender o mejorar el idioma inglés de forma entretenida mediante canciones. Y para tal propósito dos son las características fundamentales en que se apoya dicho método: La primera es que cada palabra de la letra de una determinada canción es resaltada en el momento justo en que es pronunciada, y la segunda es que se ofrece la posibilidad de ir a cualquier parte del audio con tan sólo un clic sobre cualquier palabra de la letra de una determinada canción.

Para que sea posible lo dicho anteriormente es necesario disponer de los tiempos de inicio y fin de cada palabra presentes en el audio. Hasta ahora lo que se ha venido haciendo en inglesdivino es tomar dichos tiempos de forma manual (algo muy costoso). Con este proyecto lo que se pretende es acelerar dicho proceso mediante el uso de una tecnología presente en de las tecnologías del reconocimiento de voz denominada *alineamiento forzado*.

Los alineamientos de audio y texto se harán usando dos métodos: El primero está basado en modelos HMM de fonética inglesa preexistentes, y el otro (método sin modelos preexistentes) está basado en el entrenamiento de modelos HMM desde cero usando el audio a alinear (y posiblemente audios similares complementarios).

Las pruebas se realizarán con canciones y noticias en inglés. Para las pruebas con canciones, se elegirán 3 canciones de cantantes y estilos diferentes. Para las noticias, se elegirán cuatro fragmentos de YouTube que contienen cuatro locutores: dos mujeres y dos hombres. A partir de estos datos se estudiará el nivel de influencia en el alineamiento de factores como la velocidad al hablar presente en los audios, la cantidad de información de partida, el género y el estilo musical de las canciones, etc.

Ya por último y no menos importante, se hará también un análisis subjetivo de la degradación de la calidad (en la satisfacción del usuario) del alineamiento automático respecto al alineamiento manual. De esta manera seremos capaces de determinar de forma aproximada el error permisible en el alineamiento automático respecto al alineamiento manual de forma que no sea perceptible por el usuario.

PALABRAS CLAVE

Alineamiento, Canciones y letras, Noticias, Aprendizaje de idiomas, HMM, YouTube, Inglesdivino, Vídeos musicales, Reconocimiento, Entrenamiento, Audio y transcripción

ABSTRACT

The purpose of this project is to obtain a result which is applied directly on a learning English language method implemented on the website www.inglesdivino.com. This method consists basically of learning or improving English language in an entertaining way through songs. And for that purpose there are two key features that support this method: The first one is that every word of the lyrics of a particular song is highlighted at the right time as it is pronounced, and the second one is that it offers the possibility of going to any part of the audio recording with just one click on any word in the lyrics of a particular song.

In order to carry out the above it is necessary to have the beginning and ending time for each word in the audio recording. So far in inglesdivino those times have been taken manually (very high cost). The aim of this project is to accelerate this process by using a technology of speech recognition known as *forced alignment*.

The alignments of text and audio will be performed using two methods: the first one is based on pre-existing English phonetic HMM models, and the other one (model-free method) is based on training HMM models from scratch using the audio to align (and possibly some similar complementary audios).

Testing will be done with songs and news in English. For testing songs, 3 songs from different singers and different styles will be chosen. For broadcast news, four segments from YouTube containing four speakers will be chosen: two females and two males. From these data we study the level of influence in the alignment of factors such as the speed in speaking of speech sounds, the starting amount of information, the gender and the musical styles of the songs, etc.

Finally, we do a subjective analysis of the degradation of the quality of the automatic alignment with respect to the manual alignment in user satisfaction. From this we will determine approximately the allowable error in automatic alignment with respect to manual alignment so that it is not perceptible by the user.

KEY WORDS

Alignment, Songs and Lyrics, Broadcast News, Language Learning, HMM, YouTube, Inglesdivino, Recognition, Training, Audio and transcription.

A mi madre, Isabel,

A quien le estoy inmensamente

Agradecido por sus años de sacrificio.

AGRADECIMIENTOS

En primer lugar me gustaría agradecer a mi tutor Doroteo Torre Toledano por haberme brindado la oportunidad de realizar este proyecto, por su apoyo mostrado desde el primer momento en que me propuse realizar este proyecto. Sin duda, sin sus ánimos iniciales no habría tenido la suficiente motivación para enfrentarme a este reto, sus palabras fueron claves en la toma de mi decisión. Me gustaría también agradecer a mis compañeros de laboratorio que siempre han estado allí para prestarme su ayuda cuando lo he necesitado. No sólo ha sido un apoyo a nivel académico, sino también a nivel personal. Muchas horas compartiendo el mismo ambiente crea cierto vínculo emocional, y es fundamental una buena relación para poder llevar a cabo un trabajo de forma satisfactoria, en mi caso así ha sido y lo agradezco.

Otra de las personas responsables de que este proyecto haya salido a la luz ha sido la que en todo momento me ha prestado su apoyo incondicional, incluso antes de nacer ya velaba por mí, me refiero a mi madre, que a lo largo de prácticamente toda mi vida ha ejercido de padre y madre en todas las acciones de mi día a día, y, por supuesto, a lo largo de toda mi trayectoria académica. Es imposible no mencionar a mi hermano Fabricio, que siempre ha estado a mi lado apoyándome en todo lo que ha podio, mi fiel compañero de prácticas que siempre parecía tener una solución donde yo sólo encontraba problemas.

Agradezco también al todo el personal docente que durante todos estos últimos años han depositado parte de sus conocimientos en cada uno de nosotros contribuyendo a nuestra riqueza tanto a nivel profesional como personal.

Atrás quedan unos años de duro trabajo y sacrificio, pero también de experiencias agradables que nunca olvidaré. Aquellos días de prácticas interminables, de días sin almorzar, de pérdidas de la noción del tiempo. Aquellas noches o más bien madrugadas estudiando sin parar, también aquellas pesadillas de media noche de suspensos de última convocatoria. No ha sido nada fácil, pero el esfuerzo ha merecido la pena. Indudablemente, lo aprovechado durante estos años me servirá para el resto de mi vida.

Y ya por último me gustaría pedir disculpas si me olvido de alguien sin agradecer, pero en todo caso que quede claro que agradezco a “todo” aquel que haya contribuido de alguna u otra manera al desarrollo de este proyecto de fin de carrera. Dentro de ese “todo” está Fernando, un buen amigo y una de las personas más estupendas que he conocido, que sabe tan bien como yo las cosas dulces y amargas de esta aventura. Sin más que decir me despido diciéndoles que pasen y vean lo que las siguientes hojas de este proyecto les tienen preparado para ustedes. Muchas gracias.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1

INTRODUCCIÓN

1.1. MOTIVACIÓN

Hoy en día el idioma inglés se ha convertido en algo fundamental para el desarrollo de cualquier actividad profesional, sobre todo en una sociedad tan competitiva como en la que vivimos, donde un buen manejo de este idioma juega mucho a nuestro favor. En España es especialmente crítica la dificultad para el dominio de esta lengua, así lo revelan los barómetros de febrero del 2010 realizados por el CIS: Un 63.1 % de la población española no habla, no escribe, ni lee en inglés. Sólo un 22.9% habla y escribe en inglés. Estas cifras desoladoras nos sitúan a la cola de Europa en cuanto al conocimiento de este idioma.

Una de las causas de que esto suceda es sin duda que estamos poco habituados a escuchar y pronunciar en esta lengua, y si lo hacemos lo hacemos obligados o por el simple hecho de aprobar una asignatura, pero no realmente con el entusiasmo de aprender. Quizá sea necesario encontrar métodos más amenos que nos ayuden en esta labor.

Es con ese objetivo que surgió la página web www.inglesdivino.com. La diferencia fundamental del método empleado en dicha página con respecto a otros radica en que además de didáctico puede ser entretenido, y así aprender inconscientemente a base de escuchar y repetir. Estamos hablando de aprender o mejorar nuestro inglés con canciones en inglés.

En la actualidad existen infinidad de páginas en internet que ofrecen esta posibilidad, pero de manera muy pobre, se limitan a poner una canción y su letra correspondiente, ignorando el hecho de que la mayoría de las veces un estudiante de este idioma se pierde al seguir la letra y no es capaz de saber en cada instante que palabra está siendo pronunciada. En inglesdivino se solventa este problema haciendo que las letras

de las canciones “cobren vida”, gracias a que cada palabra es remarcada en el momento de ser pronunciada.

Según las últimas estadísticas de inglesdivino, la página está teniendo una muy buena aceptación entre los usuarios. La única publicidad de la que se hace uso es el boca a boca y las redes sociales. Crecemos a un ritmo de 10.000 usuarios nuevos por mes. Actualmente 2000 usuarios se conectan cada día a nuestro sitio web para aprender y entretenerse con este método. Conforme aumentan los usuarios, aumenta la demanda de canciones, y sin un sistema automático que nos ayude en la labor de adaptación de las canciones a este método nos es muy difícil cumplir con sus necesidades. Es por eso que se ha decidido entrar al grupo ATVS para desarrollar un sistema basado en las tecnologías del habla que nos ayude en esta labor y así poder satisfacer la demanda cada vez mayor por parte de nuestros usuarios.

1.2. OBJETIVOS

El objetivo de este proyecto es diseñar un sistema automático cuyos resultados, tras una corrección manual, sean aplicables directamente a la página de inglesdivino [1]. El sistema tendrá que ser capaz de reconocer de la forma más precisa posible el momento en el que empieza y termina cada palabra de un determinado audio. Esta información de tiempos de inicio y final son los que se necesitan en inglesdivino para dotar de dinamismo a las letras de las canciones, y es también la información que más cuesta conseguir debido a su alto coste en tiempo.

Lo que se pretende con este proyecto es reducir en más del 50% el tiempo empleado en adaptar cada canción al método de aprendizaje de inglés empleado en inglesdivino. Lo ideal sería que el sistema que se pretende diseñar no tuviese ningún error localizando los tiempos, pero lamentablemente, como todos sabemos, eso es prácticamente imposible. Es por eso que no aspiramos a un 100% de cierto, pero si a un porcentaje que se aproxime lo máximo posible a dicho valor.

Además debemos tener en cuenta que en este caso al tratarse de canciones tenemos un factor más en contra nuestra, y es la música de fondo que tiene casi el mismo nivel de potencia que la propia voz o incluso a veces mayor.

El sistema a desarrollar tendrá como datos de entrada únicamente un audio y su transcripción correspondiente. Esta última afirmación no es del todo cierta, ya que como veremos a lo largo del desarrollo de este proyecto, en uno de los métodos que emplearemos para realizar el alineamiento necesitaremos más de un audio si queremos obtener mejores resultados.

Los audios con los que se tratará son canciones y noticias en inglés. Actualmente en inglesdivino nos interesa obtener unos buenos resultados en alineamiento con

canciones, ya que es en lo que se basa fundamentalmente esta página. Pero en un futuro pensamos ampliarlo a noticias, por lo que también ser hará un análisis en este tipo de audios. Aunque ya suponemos que si el sistema funciona bien con canciones, los resultados con noticias serán mejores o como mínimo serán igual de buenos, ya que en este caso no se tiene la música como ruido de fondo.

Todos los datos multimedia con los que funciona inglesdivino son proporcionados por el portal YouTube [14]. Por este motivo, los experimentos de este proyecto que requieran de datos de entrada multimedia también procederán del mismo portal. En cuanto a las transcripciones de los audios y las referencias manuales, éstas serán extraídas de las bases de datos de inglesdivino.

1.3. ORGANIZACIÓN DE LA MEMORIA

En este proyecto se empieza presentando un problema de la vida real, que se pretende solventar parcialmente con el desarrollo de este proyecto. También se describen (en el apartado del estado del arte) aplicaciones similares con el mismo problema. Posteriormente se pasa a hablar de los posibles métodos a emplear en este PFC para solventar dicho problema. Finalmente se hace un análisis de los métodos que mejor se adaptan a los audios dependiendo de su naturaleza (si son canciones o noticias). Toda esta información se recoge en 5 capítulos a lo largo de toda la memoria. A continuación se pasa a describir de forma general cada una de ellas:

CAPÍTULO 1. INTRODUCCIÓN

En este capítulo se hace una pequeña introducción a lo que se desarrollará a lo largo de toda esta memoria. Se intenta dejar claro las necesidades por las que surge este proyecto y los objetivos factibles a los que se aspira. Por último, se describe la estructura que se ha decidido adoptar para el desarrollo de este trabajo.

CAPÍTULO 2. ESTADO DEL ARTE

En este apartado se hará alusión a algunos estudios científicos similares a los desarrollados en este trabajo, relacionados con el alineamiento de audio y texto sobre canciones y voz hablada. También se verán aplicaciones de la vida real que comparten el mismo problema que se intenta solventar en este PFC. En particular, se hará una descripción en profundidad de inglesdivino [1], sitio web al que irán aplicados los resultados obtenidos en este proyecto.

CAPÍTULO 3. DESARROLLO DEL SISTEMA

En esta sección se describen paso a paso todos los elementos necesarios para implementar el sistema completo. Se empieza describiendo la fuente de los datos con los que se trabajará, se prosigue con la descripción del desarrollo del sistema haciendo

una breve alusión teórica y explicando en detalle los dos métodos de alineamiento empleado en este PFC. Por último, se describe de forma general todo el proceso en conjunto mediante un esquema global.

CAPÍTULO 4. EXPERIMENTOS Y RESULTADOS

En este capítulo se desarrolla una explicación de todos los experimentos realizados con el sistema completo descrito en el capítulo anterior. Una vez vistos dichos experimentos se pasa a mostrar los resultados obtenidos a partir de ellos.

CAPÍTULO 5. CONCLUSIONES Y TRABAJO FUTURO

En este capítulo se extraen las principales conclusiones a partir de los resultados observados en el capítulo anterior. De dichas conclusiones se elegirá qué método de alineamiento de los estudiados es más adecuado para cada aplicación dependiendo del tipo de fuente de información.

También se sugieren posibles líneas de trabajo futuro que se pueden seguir para continuar con el desarrollo de este tipo de sistemas para lograr optimizar los resultados obtenidos en este proyecto.

2

ESTADO DEL ARTE EN ALINEAMIENTO DE AUDIO Y TEXTO

2.1. INTRODUCCIÓN

En el contexto de este proyecto entenderemos por alineamiento al hecho de, dados un audio y su transcripción, localizar los instantes de tiempo en los que empiezan y terminan los fonemas presentes en un determinado audio. La transcripción estará compuesta de palabras, y, a su vez, éstas podrán descomponerse en una secuencia de fonemas que representarán la pronunciación de dichas palabras.

La aplicación final a la que irán destinados los resultados de este proyecto no necesita de una precisión a nivel de fonema, pero se hará de esta forma ya que es más preciso que hacerlo a nivel de palabra.

Una vez se dispone del alineamiento a nivel de fonema, es fácilmente extraíble el alineamiento a nivel de palabra (lo que en realidad se necesita en inglesdivino [1]). Para ello, y teniendo en cuenta que una palabra está compuesta por una secuencia de fonemas, simplemente extraemos el tiempo de inicio del primer fonema y el tiempo final del último fonema de la palabra en cuestión.

Para llevar a cabo lo dicho anteriormente es necesario el uso de un Alineador, un sistema que automáticamente alinea en tiempo señales de audio y su texto correspondiente. La tarea de tal Alineador está muy relacionada con la del reconocimiento automático de voz. En reconocimiento, estamos interesados en las palabras que han sido pronunciadas y no nos importan los puntos temporales de comienzo y final de cada segmento (generalmente una palabra o un símbolo de un extracto de pausa, música, etc.) Sin embargo, para el alineamiento el comienzo y final de un segmento son de nuestro pleno interés, mientras que se da por conocida cada

palabra pronunciada (se dispone de su transcripción). Por tanto, parece bastante natural modificar un reconocedor existente que nos haga el trabajo.

El término *alineamiento* empleado en este proyecto está muy relacionado con el de *segmentación* empleado en [2]. Al igual que en el alineamiento, en segmentación la tarea principal es la de localizar tiempos de inicio y final en unidades lingüísticas, ya pueden ser éstas fonemas, palabras o frases enteras. El objetivo final es el mismo: localizar de la forma más precisa posible dichas fronteras temporales en un determinado audio.

El desarrollo de este tipo de tecnologías no sólo contribuiría a resolver problemas a los que nos enfrentamos en inglesdivino [1], sino también sería de gran utilidad para aplicaciones estrechamente relacionadas con el alineamiento de audio y texto: Subtitulado de televisión, entretenimiento (ejemplo: Karaoke), juegos basados en audio sincronizado, etc.

El aporte es también importante dentro del ámbito del reconocimiento de voz. Concretamente en las metodologías basadas en Datos (Data-Driven) y Bases de Datos (como se menciona en [2]), donde es necesario el entrenamiento de unos modelos a partir de unos datos que requieren de una segmentación previa a nivel fonético. Dichos modelos serán los que representarán los sonidos correspondientes a un cierto conjunto de unidades, y a partir de las cuales se podrán construir todos los vocabularios que serán capaces de manejar los reconocedores de voz.

2.2. ALINEAMIENTO: TRABAJOS CIENTÍFICIOS SIMILARES

Ya se ha visto que el alineamiento en el contexto de nuestro proyecto no es más que identificar los tiempos de inicio y fin de los fonemas presentes en un determinado audio. Concretamente, este Proyecto Fin de Carrera se centra fundamentalmente en el alineamiento de canciones en inglés (dados el audio de la canción y su letra correspondiente). También se extiende el estudio para audios de noticias en inglés como ampliación, ya que es de suponer que una vez superadas las barreras del alineamiento sobre canciones se den por superadas también la de los audios de noticias, ya que estos últimos son audios más fácilmente tratables que las canciones al incluir menos ruido y menos alteraciones en la pronunciación de los fonemas. Esta suposición se confirmará posteriormente en el apartado de experimentos.

Cabe aclarar que en este proyecto se emplean algunas veces los términos *alineamiento* y *segmentación* de forma casi indistinta, teniendo claro que la segmentación es la localización de fronteras temporales de los segmentos presentes en un audio. En este proceso, una vez hecha la segmentación se desconoce la etiqueta de cada segmento, para saberlo, es necesario un proceso de etiquetado, que consiste básicamente en asignarle a cada segmento troceado una etiqueta. Una buena explicación de las

diferencias entre segmentación y etiquetado puede encontrase en [2]. En concordancia, el alineamiento está muy ligado a la segmentación, ya que se trata de segmentar a partir de un audio y una secuencia de etiquetas, y obtener como resultado fragmentos segmentados cada uno con su etiqueta correspondiente.

Una vez dicho esto, a continuación se muestran algunos ejemplos muy significativos del interés científico que tiene el alineamiento de audio y texto sobre voz y canciones:

- [2], en donde se hace un profundo estudio sobre la segmentación (paso fundamental en el alineamiento) como paso previo al desarrollo de reconocedores de voz mucho más sofisticados que los convencionales.

- [3], en donde, al igual que en [2], se utilizan técnicas de refinamiento local para mejorar la segmentación realizada mediante el uso de modelos HMM.

- [4], en donde se hace un alineamiento entre un set de fonemas extraídos de un audio y un set de fonemas extraídos de una transcripción para posteriormente, tras un análisis de los mismos, reportar imperfecciones en dicha transcripción.

- [5], que se ocupa de la creación automática de enlaces entre texto y audio basado en alineamiento para aplicación en el idioma alemán.

- [6], en donde se plantea el problema de realizar una segmentación en sílabas detectando para ello el núcleo silábico.

- [7], que en lugar de segmentar los fonemas en las fronteras fonéticas, trata de encontrar sus centros.

Las referencias citadas anteriormente están más relacionadas con alineamientos sobre audios de carácter general. Como se expuso anteriormente, este proyecto se centra sobre todo en audios de canciones y noticias. A continuación se hace mención de la acogida que tiene el alineamiento de canciones en el panorama científico, pero antes de pasar a tratar dicho tema, primero se mostrará algunos trabajos que tienen que ver sobre todo con el alineamiento en noticias: El tema de los subtítulos en noticias ha sido más estudiado debido a su clara aplicación, en particular para permitir a las personas sordas acceder al contenido de las noticias. El subtitulado de noticias se puede afrontar de dos maneras: usando reconocimiento de voz para obtener la transcripción y luego realizar el alineamiento sobre el audio, como se hace en [11], o haciendo uso del conocimiento previo de la transcripción usada por los locutores de noticias para alinear texto y audio como se hace en [12]. En este proyecto se usará siempre la transcripción ya sea de las canciones o de las noticias.

Volviendo al tema del alineamiento de canciones, encontramos que también ha habido un amplio interés científico en estos últimos años. Así lo demuestran trabajos como los siguientes:

- [8], en donde, como paso previo al alineamiento sobre canciones, se utiliza un algoritmo de extracción de voz. Tras esta etapa, se realiza el alineamiento usando modelos HMM preexistentes.

- [9], en donde se usa programación dinámica y un método sin modelos.

- [10], en el que se intenta primero aislar la voz y la música, y luego se adaptan modelos HMM preexistentes a la canción con música incluida. También, como etapa del preprocesado se eliminan las zonas en las que no hay voz mediante un método de detección de actividad vocálica.

- [13], en donde se presenta el sistema *LyricAlly.* Este sistema realiza el alineamiento en dos fases: primeramente alinea la estructura a nivel superior de una canción. Y Luego, dentro de las fronteras de las secciones detectadas, refina el alineamiento usando una estimación uniforme de la duración de un fonema en lugar de utilizar un método basado en reconocimiento de fonema.

Como vemos, el alineamiento en canciones ha acaparado numerosas publicaciones científicas. De todas ellas podemos decir que la mayoría se basan en la utilización de modelos HMMs preexistentes. Y muchas de ellas tienen en común un preprocesado que consiste en la extracción de ruido o música de fondo de las canciones. En este proyecto se hará un preprocesado ligeramente distinto en las canciones, que consiste en la extracción de las zonas donde no hay voz, pero en nuestro caso se hará de forma manual. Dichas zonas sin voz serán eliminadas si son aproximadamente mayores a 2 segundos. Al igual que en la mayoría de los trabajos realizados en este campo, también se seguirá un enfoque basado en modelos HMM. A parte de utilizar modelos HMMs preexistentes, en este proyecto también se utilizará un método libre de modelos pre-existentes que más adelante se verá en qué consiste.

2.3. INGLESDIVINO: ALINEAMIENTO A NIVEL DE PALABRA

En este apartado se explica en detalle la aplicación web que motivó principalmente la realización de este proyecto. Se explicaran sus principales características, su situación actual y su relación con este proyecto.

2.3.1. ¿Qué es inglesdivino?

Inglesdivino [1] es un sitio web creado con el objetivo de ayudar a los estudiantes de inglés a aprender o mejorar dicho idioma de una manera divertida. Para lograr tal objetivo lo que se hace es incorporar al proceso de aprendizaje un factor que motive a estudiar este idioma, no sólo por obligación o necesidad, sino también por entretenimiento. Y esto se pretende lograr con la ayuda del uso de canciones en inglés, aunque también con otras aplicaciones como juegos o karaoke en inglés.

En inglesdivino [1] somos conscientes de la infinidad de páginas web existentes que emplean este mismo método en enseñanza. Es por eso que decidimos añadir algo más a nuestro sitio web, algo que ninguna otra página ofreciera. Es entonces cuando surge una idea sencilla que llevada a la práctica se convierte en potencialmente atractiva. Seguramente todo aquel que ha estudiado inglés, alguna vez intentó aprenderse una canción en inglés, y seguramente al principio cuando quiso seguir el audio y la letra a la vez se perdió porque no sabía que palabra estaba siendo pronunciada aunque tuviese delante la letra de la canción. Esto se soluciona en inglesdivino haciendo que la letra de la canción y el audio estén perfectamente sincronizados en tiempo, de manera que cuando se pronuncie una determinada palabra ésta sea remarcada. En inglesdivino [1] también se permite que un usuario pueda ir a cualquier parte de la canción (hacia delante o hacia atrás), sin tener que esperar a que pase el audio, con sólo clicar sobre la palabra a la que se quiera ir. Una vez clicada la palabra, automáticamente se buscará la correspondencia en tiempo y se reproducirá la canción justo en el momento en el que se pronuncia la palabra clicada. En la Figura 2.1 se muestra a modo de ejemplo cómo se muestra una palabra cuando está siendo pronunciada.

Figura 2.1: Ejemplo de la palabra “moment” en el momento de ser pronunciada.

2.3.2. Relación de inglesdivino con este PFC

Para conseguir sincronizar audio y texto con la precisión que se hace en inglesdivino es necesario disponer de los tiempos de inicio y fin de cada palabra presentes en el audio. Actualmente esto se hace de forma manual, labor que conlleva muchas horas de trabajo y paciencia.

La labor de tomar los tiempos de inicio y final de una palabra se simplifican bastante teniendo en cuenta que lo normal es que una palabra comience en el instante en que ha acabado la anterior. Así, al tomar el tiempo final de una determinada palabra, estamos también tomando el principio de la siguiente. Obviamente esto no siempre es así, ya que en los finales de frases hay pausas largas hasta que empieza la siguiente palabra. Pero dado que hasta que no se empieza a pronunciar la siguiente palabra, estará marcada la palabra actual aunque no esté siendo pronunciada (porque se ha pronunciado ya y ahora hay una pausa musical), nos interesará entonces sólo el tiempo de inicio de cada palabra (que la mayoría de las veces será el final de la anterior). Es por eso que no es del todo correcto decir que sólo tomamos los tiempos de inicio de las palabras, ya que intrínsecamente tomamos el final de la mayoría de ellos. Una vez hecha esta aclaración, a continuación se pasa a ver la relación de esa página web con este PFC.

Después de todo lo mencionado anteriormente parece quedar claro que es necesario un mecanismo que nos ayude en la labor de toma de tiempos. Con ese objetivo surge este proyecto. No se pretende que el sistema a diseñar en este PFC tome los tiempos de las palabras con un acierto del 100%, pero si al menos con más del 50% en la mayoría de los casos. Los tiempos que no fuesen correctos se corregirían de forma manual. Con lo que finalmente se obtendría un sistema combinado basado en un alineamiento automático con posterior corrección manual. Lograr esto supondría un notable ahorro en tiempo.

Para ello el sistema será básicamente un Alineador, que se encargará de recibir como entrada un audio y su transcripción y dará como resultado los tiempos de inicio y final de cada palabra presente en el audio.

2.3.3. Estadísticas de inglesdivino

Ya hemos hablado de las bondades de inglesdivino, pero ¿está recibiendo una buena aceptación por parte de los usuarios? Las estadísticas nos dicen que sí, y así los demuestran la gráficas del número de usuario por mes. Teniendo en cuenta que no se utiliza publicidad pagada, sino sólo la publicidad boca a boca y el poder de las redes sociales como Facebook o Twitter, es un buen indicador el ritmo de crecimiento que estamos teniendo: 10000 usuarios más cada mes, en la actualidad alrededor de 2000

personas se conectan a diario para aprender inglés con este método. En la Figura 2.2, se puede ver el ritmo de crecimiento de los últimos meses.

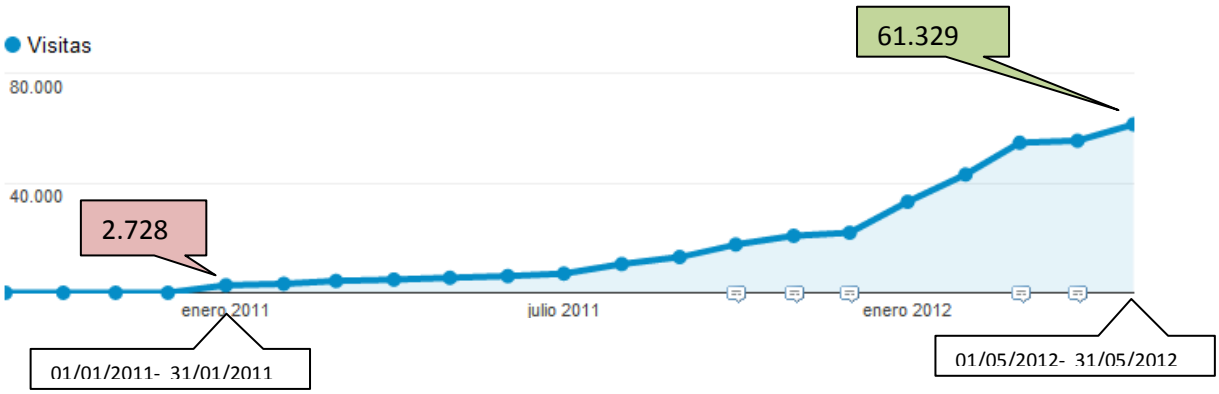


Figura 2.2: Número de vistas mensuales de inglesdivino

En cuanto a la procedencia de los visitantes de inglesdivino, la mayoría proceden sobre todo de países hispanohablantes. En la Figura 2.3 se pueden ver los países más activos en todo lo que lleva de vida inglesdivino (aproximadamente un año).

	País/territorio	Visitas	% Visitas
1.	Mexico	110.088	22,22%
2.	Spain	85.366	17,23%
3.	Colombia	62.193	12,55%
4.	Peru	46.014	9,29%
5.	Argentina	45.571	9,20%
6.	Chile	26.944	5,44%
7.	United States	23.736	4,79%
8.	Venezuela	13.999	2,83%
9.	Ecuador	13.906	2,81%
10.	(not set)	7.805	1,58%

Figura 2.3: Principales países visitantes de inglesdivino

Un indicador también muy significativo de cualquier página web es el porcentaje de usuarios que vuelven frente a los que son nuevos. Si se representan los dos en una misma gráfica, no está claro cuál sería la mejor distribución de porcentajes entre los mismos. Ya que si el 100% son visitantes que vuelven significaría, que la página no está creciendo ya que no hay ningún visitante nuevo. Si, por el contrario, hubiese un 100% de visitantes nuevos, tampoco sería nada bueno, ya que eso significaría que no hay ningún usuario que esté volviendo. Luego ahí hay un compromiso entre los porcentajes que dependerá mucho del tipo de página. En la Figura 2.4 se muestra esta gráfica para el caso de inglesdivino.

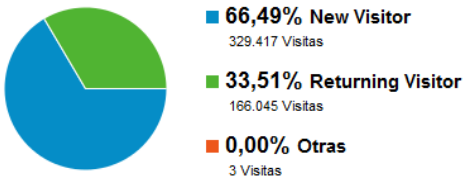


Figura 2.4: Visitantes nuevos vs Visitantes que vuelven de inglesdivino

Observando la Figura 2.4, vemos que, del total de visitantes, más del 50% son nuevos, lo que es un buen indicador de la velocidad de crecimiento de la página. Que sólo el 33.51% sean usuarios que vuelven nos es un mal indicador, ya que este porcentaje varía sobre todo dependiendo de los usuarios nuevos que vengan a inglesdivino.

2.3.4. Problemas generales por resolver en inglesdivino

El principal problema en inglesdivino [1] es el alineamiento manual que se tiene que realizar. Este elevado coste en tiempo se deriva en que no se dispongan del suficiente número de canciones como para poder satisfacer a todos los usuarios. Además se necesita un personal cualificado a la hora de tomar tiempos de palabras de las canciones. Las consecuencias de este gran problema se intenta amortiguar con la realización de este PFC, que esperamos que nos reduzca el tiempo en al menos un 50%, cosa que ya sería bastante aceptable.

Otros de los problemas que tiene que ver menos con el alineamiento manual de canciones es la disponibilidad. Actualmente se está produciendo un incremento masivo en el uso de dispositivos móviles a nivel mundial. Y por desgracia, la inmensa mayoría de ellos no soportan la tecnología flash, ingrediente fundamental para que nuestra página web funcione, ya que, entre otras cosas, los videos insertados de YouTube, están basados en esa tecnología.

Afortunadamente, recientemente se ha anunciado por parte de los ingenieros de Google que se lanzarán APIs para los desarrolladores cuyas páginas webs estén basadas en videos de YouTube. Estamos a la espera, y de ser cierto, creemos que atraeríamos a gran parte de usuarios que apenas usan el ordenador, dando prioridad a este tipo de dispositivos.

Otros problemas de menor relevancia tienen que ver con la falta de algunos contenidos, pero que se solventarían fácilmente dedicando el tiempo necesario a cada tarea. Cosas como ampliar el contenido teórico, extender el contenido a otros idiomas, incluir más juegos, incluir chat, etc. son los que se podrían categorizar dentro de este tipo de problemas.

2.4. OTRAS APLICACIONES: ALINEAMIENTO A NIVEL DE FRASE

En la sección anterior se ha descrito una aplicación (inglesdivino) que tiene como ingrediente fundamental el alineamiento a nivel de palabra. A continuación se describirán dos aplicaciones en las que el alineamiento se hace a nivel de frase.

2.4.1 Transcriptor de YouTube

YouTube [14] es un portal de vídeos, propiedad de Google [15], muy conocido a nivel mundial. En los últimos años, en Google [15], se han estado desarrollando tecnologías relacionadas con el reconocimiento de voz para aplicación directa sobre los vídeos de dicho portal. Hace ya algún tiempo que en YouTube [14] se permite en mucho de sus vídeos a los usuarios puedan ver la transcripción automática de los audios presentes en los vídeos. Esta transcripción se realizada de forma completamente automática basándose en las tecnologías de reconocimiento de voz. Es por eso que, en mucho de los casos, se obtiene una transcripción de una calidad muy mala, sobre todo si hay mucho ruido de fondo. Si los vídeos tienen una calidad aceptable y los locutores pronuncian lo más correctamente posible, las transcripciones que se obtienen pueden llegar a ser muy buenas. Pero lo normal es que los usuarios suban vídeos que en los hablan sin pensar que un reconocedor de voz va tener la labor de intentar averiguar las palabras que dicen. En la Figura 2.5 se muestra cómo acceder a la transcripción (o subtítulos) automática del portal YouTube [14].

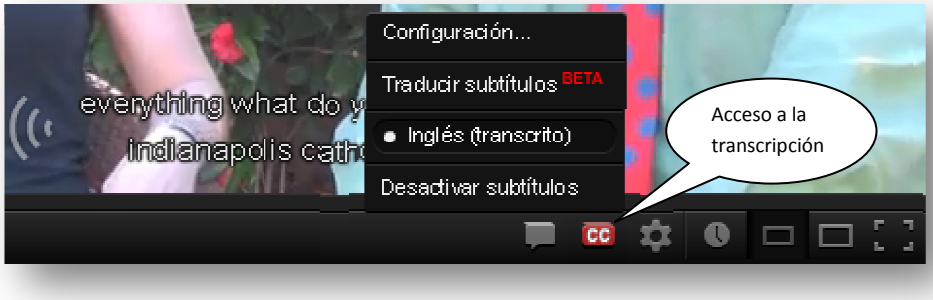


Figura 2.5: Transcripción automática del portal YouTube

Recientemente, en todos los vídeos que incluyen la funcionalidad anterior, se permite también ver un alineamiento a nivel de frase. Las frases de de dicho alineamiento son también, por supuesto, obtenidas automáticamente, de hecho son las mismas que aparecerían en el vídeo si elegimos la funcionalidad de subtitulado con en la Figura 2.5. En la Figura 2.6 se muestra una ilustración de dicho alineamiento y cómo acceder a ella.

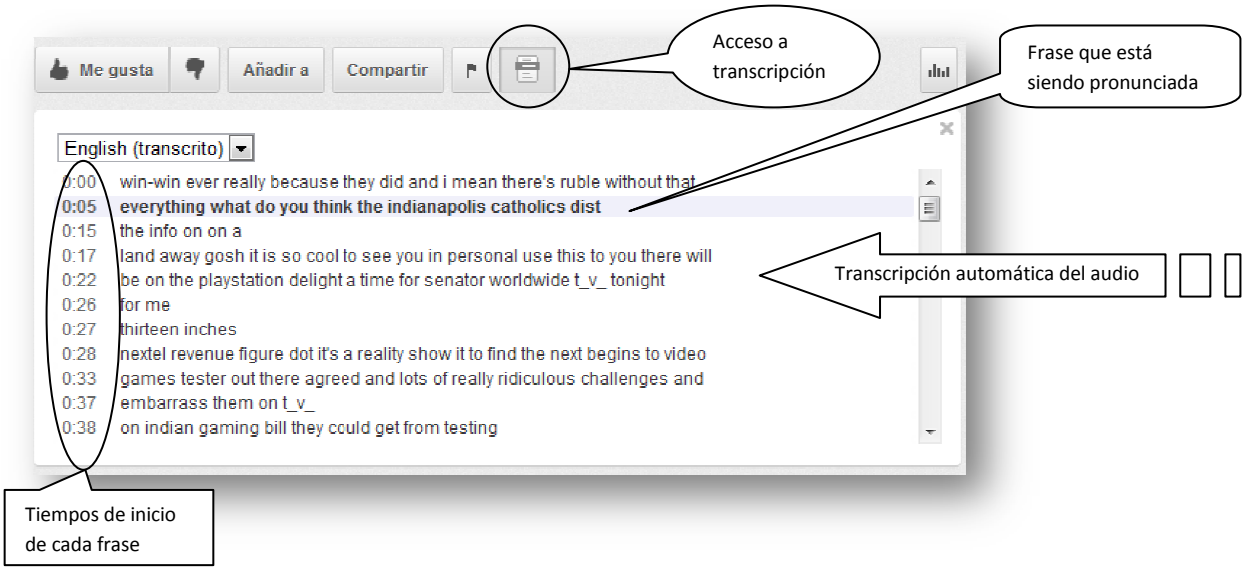


Figura 2.6: Alineamiento a nivel de frase del portal YouTube

2.4.2 Lyricstraining

Lyricstraining [16] es una página web similar a inglesdivino [1]. Su objetivo es enseñar idiomas mediante canciones. A parte del inglés, también se puede practicar idiomas como el francés o el alemán.

En esta página web, las letras de las canciones, al igual que inglesdivino, no son generadas automáticamente, sino que se obtienen previamente de forma manual.

A diferencia de inglesdivino [1], Lyricstraining [16] hace un alineamiento a nivel de frase. Y al obtenerse la transcripción de forma manual, no hay apenas errores en las letras de las canciones (salvo posibles erratas). En la figura 2.7 podemos ver la apariencia que tiene el alineamiento en dicha página.

Figura 2.7: Alineamiento a nivel de frase de Lyricstraining

3

DESARROLLO DEL SISTEMA

3.1. INTRODUCCIÓN

Este capítulo empieza describiendo cosas generales de este proyecto como, por ejemplo, los recursos que han sido necesarios para llevarla a cabo, el enfoque empleado en este PFC, los fundamentos teóricos detrás del alineamiento de audio y texto, y el tratamiento inicial de los datos de entrada del sistema.

Posteriormente se pasa a describir el procedimiento completo seguido para el desarrollo del sistema Alineador. En dicha descripción se incluirán los detalles en profundidad de cada uno de los elementos que componen el sistema final. También se analizarán en detalle los dos métodos de alineamiento empleados en este proyecto.

Finalmente se describirá el proceso de fusión de los resultados, dados por el sistema final, con el sitio web inglesdivino [1]. En él se detallaran tanto la obtención de los datos desde el sitio inglesdivino [1], como la subida de los resultados del sistema al mismo. Y, para terminar, se concluirá dando una visión global del sistema completo.

3.2. GENERALIDADES SOBRE EL DESARROLLO DE ESTE PFC

Este Apartado describe una serie de aspectos genéricos sobre el desarrollo de este PFC, como los medios materiales utilizados, el software empleado, las bases de datos empleadas para las pruebas, etc.

3.2.1. Medios Materiales

Para este Proyecto Fin de Carrera sólo ha sido necesario el uso de un equipo informático, empleado tanto para los desarrollos y las pruebas, como para la preparación de la documentación.

- Características del equipo:

  Los desarrollos y las pruebas, así como la documentación se han realizado en un ordenador portátil con dos sistemas operativos: Linux Ubuntu 11.10 y Windows vista. Bajo el primer sistema operativo se han realizado los desarrollos y las pruebas, y bajo el segundo, con el paquete de herramientas Microsoft Office incluido, la redacción de la memoria.

3.2.2 Software

Se han empleado dos sistemas operativos: Linux Ubuntu 11.10 y Windows vista. Para la realización de la documentación se ha empleado el paquete de herramientas Microsoft Office. En este Apartado no se pretende describir ese software, ya que se trata de un software bastante convencional y ampliamente conocido. En lugar de ello se va a describir el software que se ha empleado en los desarrollos y en las pruebas realizadas en este PFC, que sí resulta un software mucho más específico.

- *Hidden Markov Models Toolkit (HTK) v3.4*

  Se trata de un conjunto de herramientas software que facilitan las tareas de entrenamiento y realización de pruebas con modelos ocultos de Markov (HMMs). Se utilizará este software para afrontar el problema del alineamiento automático basados en HMMs. Su uso se facilita gracias a un estupendo manual de usuario [17], que incluye una buena introducción teórica a los modelos ocultos de Markov y a su utilización en reconocimiento de voz.

- *SoX 14.4.0*

  Es un software diseñado para el tratamiento de audio. Entre algunas de sus funciones se encuentran la de filtrar, cambiar frecuencia, sumar y restar señales de audio, recortar, etc. En nuestro proyecto lo utilizaremos principalmente para pasar audios en modo estéreo a modo mono, para cambiar frecuencia de muestreo, y sobre todo para recortar fragmentos de audio.

- *Gnuplot 4.6.0*

  Se trata de un software utilizado para representación gráfica de datos. En nuestro proyecto nos será muy útil para dibujar sobre todo histogramas de los resultados obtenidos.

3.2.3 Bases de datos

En el desarrollo de este PFC se ha empleado parte de la base de datos de la que se hace uso en inglesdivino [1]. Dicha base de datos se compone fundamentalmente de transcripciones y tiempos de referencia manual. Los resultados finales del sistema

(tiempos de alineamiento) que se desarrolla en este PFC se compararán con unas referencias manuales extraídas de dicha base de datos. Las transcripciones también serán extraídas de la misma base. En cuanto a los audios de las canciones y noticias, éstas serán extraídas del portal YouTube [14], ya que inglesdivino hace uso de los videos de dicho portal para su funcionamiento.

3.2.4 Fonemas considerados

El conjunto de fonemas que se considerará a lo largo de todo este proyecto contiene 40 fonemas ingleses, que representaremos con los siguientes símbolos:

aa ae ah ao aw ay b ch d dh eh er ey f g hh ih iy jh k l m n ng ow oy p r s sh t th uh uw v w y z zh

Adicionalmente se considera un símbolo para representar los silencios y pausas: **_**. También se consideran los símbolos **R1** y **R2** como indicadores de principio y final de frase respectivamente.

3.2.5 Herramienta de evaluación del alineamiento

Dado que a lo largo de toda la sección de experimentos de este proyecto será necesario evaluar continuamente alineamientos, se ha desarrollado una herramienta muy sencilla, que permite realizar estas evaluaciones de forma fácil y consistente a lo largo de todo el proyecto.

La herramienta toma como entrada un par de secuencias de tiempo. Una de las secuencias es el resultado del alineamiento automático arrojados por el sistema desarrollado, y la otra secuencia es la referencia manual. A partir de estos datos, y especificando una determinada tolerancia de error, la herramienta es capaz de determinar el porcentaje de acierto y fallo de la segmentación automática. El porcentaje de acierto se obtiene observando el número de tiempos correctos obtenidos inferiores al valor absoluto de la tolerancia especificada.

3.3. PRINCIPIOS GENERALES DEL ALINEAMIENTO BASADO EN HMMs

El alineamiento está primordialmente basado en los mismos mecanismos usados en el reconocimiento de voz. En dichos mecanismos hay dos etapas claramente involucradas: primeramente, un proceso de entrenamiento mediante el uso de fragmentos de audio con sus respectivas transcripciones y, en segundo lugar, un proceso de reconocimiento en el que se obtienen unas trascripciones a partir de un audio desconocido. En la Figura 3.1 se puede apreciar un esquema de estas dos etapas involucradas:

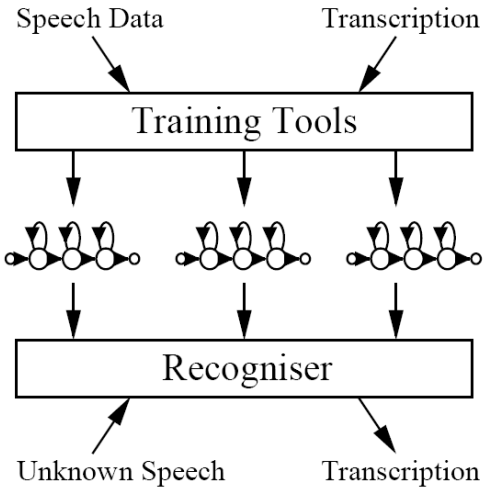


Figura 3.1: Proceso de entrenamiento y reconocimiento usando modelos HMM.

Como se verá más adelante en este proyecto, para la realización de los alineamientos se emplearán dos métodos: El primero (basado en modelos HMM preexistentes) estará centrado en la segunda etapa que se muestra en la Figura 3.1, y el segundo método (sin modelos), en la primera etapa (la de entrenamiento de modelos). En ambos casos se supondrá conocido el contenido de los audios, ya que no estamos interesados en el reconocimiento en sí, sino en las fronteras temporales de las palabras presentes en los audios.

A continuación se explica a grandes rasgos cómo funciona el reconocimiento mediante el uso de modelos ocultos de Markov (HMMs):

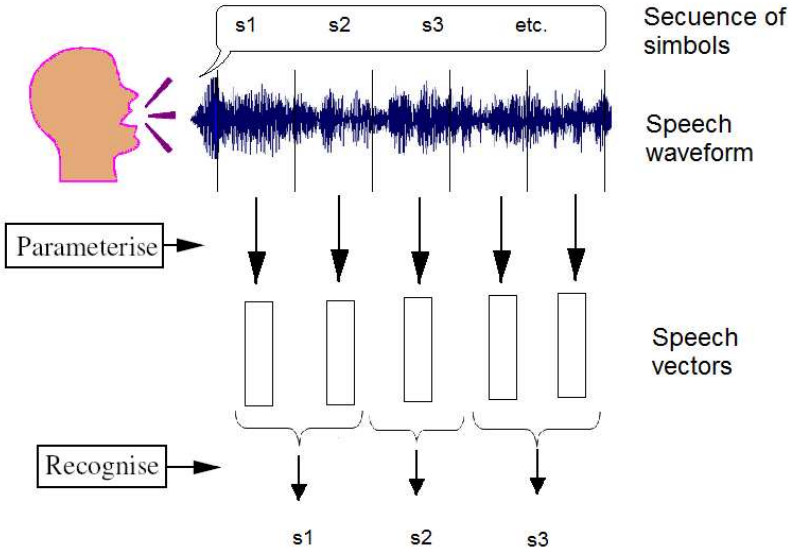


Figura 3.2: Codificación y decodificación del mensaje.

En reconocimiento de voz generalmente se asume que la señal de audio está compuesta por un mensaje codificado con uno o más símbolos (ver Figura 3.2). Para

reconocer esos símbolos detrás de ese audio es necesario convertir la señal en una secuencia discreta de vectores de parámetros igualmente espaciados en tiempo. Se asume que esta secuencia de vectores de parámetros forma una representación exacta de la señal de audio, basándose en que en la duración cubierta por un vector de parámetros (típicamente 10 ms más o menos), la señal se puede considerar estacionaria. Aunque esto no es estrictamente cierto, es una aproximación bastante razonable.

El papel de los reconocedores será la de realizar un mapeo entre la señal de audio y los símbolos ocultos en los mismos. Dos son los problemas que hacen esto muy complicado. Primero, el mapeo de audio a símbolos no es uno-a-uno, ya que distintos símbolos pueden dar lugar a sonidos iguales. Además hay una gran variación en las diferentes realizaciones de los audios, debido a la variabilidad del hablante, humor, ambiente, etc. En segundo lugar, las fronteras entre los símbolos no pueden ser identificadas explícitamente de la señal de audio. Por tanto no es posible tratar la señal como una secuencia de símbolos concatenados. El primer problema se consigue suavizar mediante varias técnicas como eliminación de ruido, utilización de bases de datos cada vez más sofisticadas, etc. Aunque en la actualidad sigue siendo uno de los retos más importantes a los que se enfrentan quienes se dedican a la investigación en este campo.

El segundo problema de no saber la localización de las fronteras de una palabra, puede ser evitado restringiendo la tarea al reconocimiento de palabras aisladas. Como se muestra en la Figura 3.3, esto implica que el audio se corresponde con un solo símbolo (ejemplo: el símbolo “w”), elegido de de un vocabulario preestablecido. Aunque este problema más simple es de alguna manera algo artificial, tiene sin embargo muchas aplicaciones. Además nos servirá como buena base para introducir las ideas del reconocimiento basado en HMMs, ideas en las que se basará el sistema Alineador de este proyecto. A continuación se explica más en detalle lo que es el reconocimiento de palabras aisladas para entender mejor cómo funcionan los modelos HMM en el reconocimiento de voz.

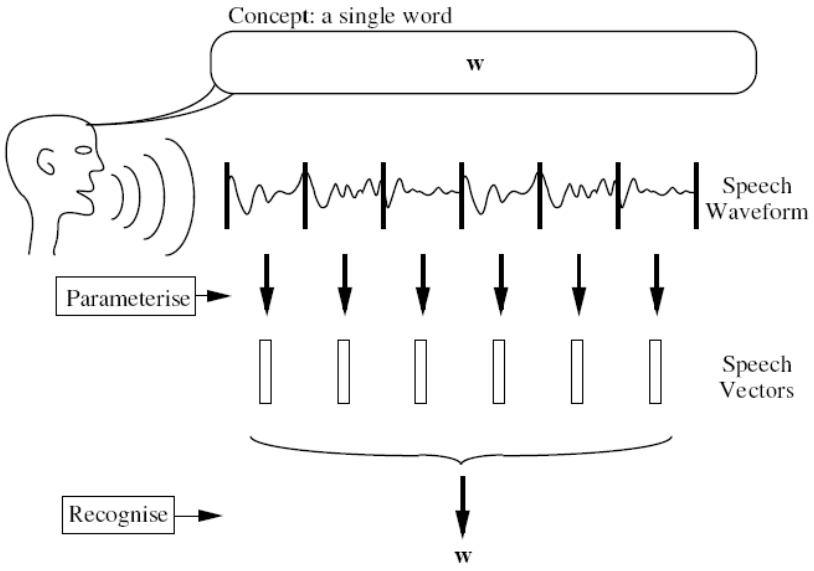


Figura 3.3: Reconocimiento de palabra aislada

3.3.1 Reconocimiento de palabra aislada

Supongamos que tenemos una observación ***O***, que puede ser la pronunciación de la palabra "hola", por ejemplo. Y supongamos también que las posibles palabras a reconocer son "hola" "que" "tal". Para realizar el reconocimiento de la palabra "hola" necesitamos tener un modelo HMM para cada una de las tres palabras que componen nuestro diccionario de posibles palabras, que en nuestro caso son "hola" "que" "tal". Llamaremos al modelo de cada palabra $\boldsymbol{M_i}$ Tratar de reconocer la observación ***O***, no es más que encontrar el modelo más probable de los tres, que posiblemente generó dicha observación. Matemáticamente esto se reduce a encontrar la probabilidad condicional $\boldsymbol{P(O|M_i)}$. Es aquí donde reside la elegancia y el poder de los modelos HMM. Dados un set de ejemplos de entrenamiento para un determinado modelo, los parámetros de ese modelo pueden ser estimados automáticamente por un robusto y eficiente método de reestimación. Por tanto, proporcionados un suficiente número de ejemplos representativos de cada palabra, se puede construir un modelo HMM que implícitamente modele todas las posibles fuentes de variabilidad inherentes al habla real. La Figura 3.4 resume el uso de los modelos HMM para el reconocimiento aislado de palabras. Primeramente, se entrena un HMM para cada palabra del vocabulario usando un determinado número de ejemplos para cada una de ellas. En este caso el vocabulario consiste sólo en tres palabras: "hola" "que" "tal". En segundo lugar, para reconocer una palabra desconocida, se calcula la probabilidad de que cada modelo genere esa palabra, y la probabilidad más alta identifica la palabra buscada. Para una detallada explicación matemática acerca de los Modelos ocultos de Markov, así como la descripción de los algoritmos utilizados para la reestimación de los mismos y para la

decodificación, el capítulo 1 del manual de HTK [17], contiene una estupenda explicación acerca de los modelos HMM y su uso en el reconocimiento de voz.

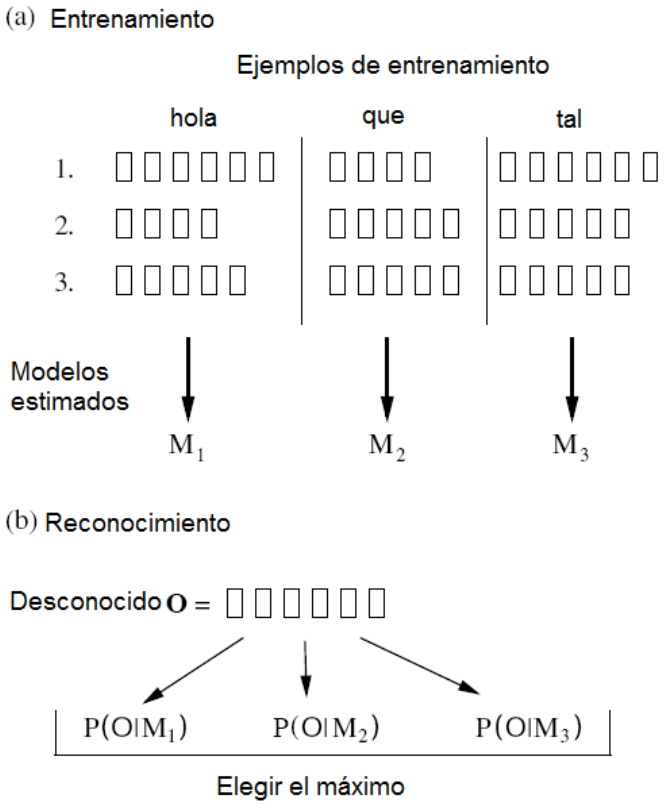


Figura 3.4: Uso de modelos HMM para el reconocimiento de palabra aislada.

3.3.2 Reconocimiento de palabras continuas

Volviendo al modelo conceptual de la reproducción del habla y el reconocimiento del mismo ejemplificado en la Figura 3.2, debería quedar claro que la extensión al audio continuo, es decir con más de una palabra, simplemente involucra conectar juntos modelos HMM en secuencia. Cada modelo de la secuencia se corresponde directamente con un símbolo presente en el audio. Estos símbolos podrían ser palabras enteras o sub-palabras como, por ejemplo, fonemas. Sin embargo hay una pequeña dificultad con la que tratar en este caso. Los datos de entrenamiento para el habla continua consisten en frases continuas (con más de una palabra) y, en general, las fronteras que dividen los segmentos del habla correspondientes a cada modelo de palabra en la secuencia es desconocido. En la práctica, suele ser más rentable marcar las fronteras de una pequeña cantidad de datos a mano. Todos los segmentos correspondientes a un modelo pueden ser entonces extraídos y el método de entrenamiento de palabra aislada, descrito anteriormente, puede ser aplicado. Sin embargo, la cantidad de datos obtenidos de esta forma es normalmente muy limitada

y los modelos resultantes será pobremente estimados. Además, aunque hubiese una gran cantidad de datos, las fronteras impuestas a mano podrían no ser tan buenas como lo requieren los modelos HMM. Por tanto, es necesario el uso de unas herramientas que permitan una inicialización de modelos desde cero. En HTK [17] se proporcionan dos herramientas para esta labor: HINIT y HREST.

En el caso de nuestro proyecto, los símbolos serán fonemas, ya que a partir de ellos se podrá formar cualquier palabra del idioma inglés. Crear modelos para cada palabra es inviable debido a la cantidad de palabras que hay en dicho idioma. En lugar de eso, se trabajarán con modelos de unidades menores que componen las palabras: los fonemas. En nuestro caso el número de fonemas con los que trabajaremos serán 40, que es una cantidad razonable de modelos con los que trabajar.

3.4. PREPARACIÓN DE LOS DATOS

En este apartado se describirá el tratamiento previo que se realiza sobre el audio y la transcripción correspondiente antes de ser procesados por el sistema Alineador.

3.4.1. Obtención del audio y la transcripción

El material multimedia proviene del portal YouTube [14]. Los audios sobre los que se realizarán los alineamientos automáticos son audios que están alineados manualmente en inglesdivino [1]. Dichos audios no se guardan en la base de datos de inglesdivino, lo único que se guarda son los alineamientos de dichos audios. Y serán estos tiempos los que se tomen de dicha base para la comparación con el alineamiento automático. Por esa razón es estrictamente necesario trabajar sobre audios que ya estén alienados en inglesdivino [1], para poder tener una referencia manual. Como es sabido, inglesdivino funciona con vídeos musicales, estos vídeos son propiedad del portal YouTube [14] y están en su base de datos, en inglesdivino lo único que se hace es referenciar a esas bases mediante un mecanismo de inserción de vídeo.

Para obtener los audios con los que se trabajará en este proyecto, lo que se hace primeramente es descargar los videos que contienen dichos audios desde YouTube [14]. A continuación se extrae el audio del vídeo y se convierte del modo estéreo al modo mono. El proceso de extracción se realiza con la ayuda del programa *MPlayer* y el proceso de conversión, mediante el programa *Sox*. Tras estos pasos, obtenemos un audio con una frecuencia de muestreo de 44100Hz en modo mono y en formato WAV. En la Figura 3.5 se pueden ver los pasos seguidos para la obtención del audio.

Figura 3.5: Proceso de obtención de audio

Para la obtención de la transcripción, el proceso es mucho más sencillo. Se podría transcribir a mano todo lo que se habla en el audio, o directamente ir a cualquier página web donde esté disponible la transcripción del vídeo descargado que posee el audio que estamos tratando. En nuestro caso, necesitamos que las transcripciones están libre de errores, y es por ello que recurrimos a las que están almacenadas en las bases de datos de inglesdivino [1], que al haberse usado para un alineamiento manual previo, han sido revisados meticulosamente y están prácticamente libre de errores. Como se verá en al siguiente sección, las transcripciones no se bajan tal cual, sino que será necesario hacer marcas en ellas para saber por dónde recortarlas posteriormente.

3.4.2 Marcaje manual de tiempos en el audio y de posición en las transcripciones.

En esta etapa lo que se hace es crear unas marcas temporales en el audio y unas marcas de posición en las transcripciones. Todo este proceso es realizado en un entorno creado en un apartado especial de desarrollo de inglesdivino [1].

El marcaje tiene como objetivo indicar al sistema Alineador los fragmentos de audio que ha de tomar para realizar el alineamiento y los fragmentos que ha de rechazar. Este marcaje se realiza de forma manual y se tarda en hacerlo aproximadamente el mismo tiempo que dura el audio (alrededor de 3 minutos y medio).

En la Figura 3.6, se puede ver como se realiza el proceso de marcaje temporal en una determinada señal de audio. Se parte del audio de una canción, en ella habrá partes en las que el locutor hable o cante (en caso de ser una canción) haciendo sus respectivas pausas cortas, y otras partes en las que habrá pausas muy largas, en las que o bien habrá silencio, ruido, o música instrumental (en el caso de ser una canción). Ésas zonas de vacío vocal son las que el sistema tiene que rechazar, y para ello se lo indicaremos fijando unas marcas (**Mi**) tanto en la señal de audio como en la transcripción. Dichas marcas **Mi,** en el caso de audio, son tiempos que se sitúan al principio de cada fragmento donde hay voz y al final de la misma, y en el caso de las transcripciones, son

asteriscos en la letra que delimitan los diferentes fragmentos de transcripción. De esta forma tendremos identificados qué partes del audio incluyen voz y cuáles no.

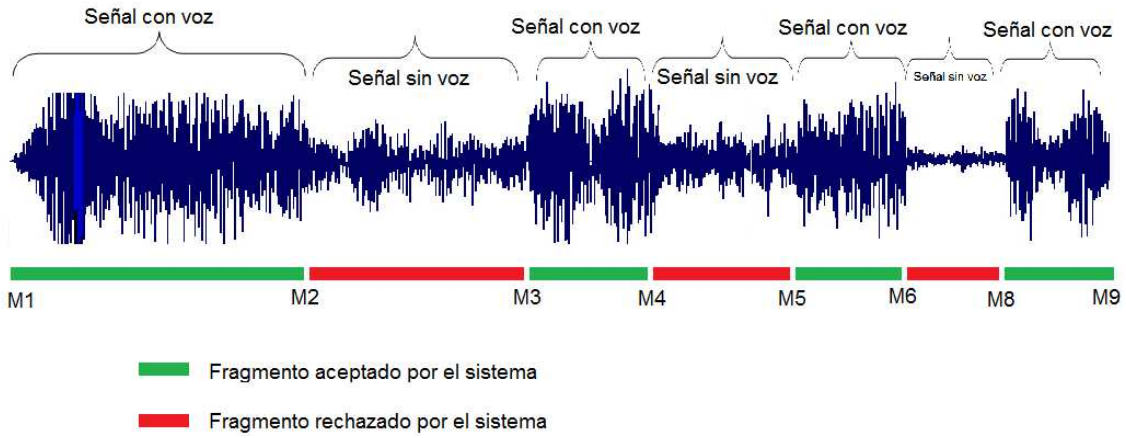


Figura 3.6: Marcas temporales sobre el audio

En la sección anterior vimos cómo se obtenía la señal de audio. Es esta misma señal de audio sobre la que se realiza el marcaje temporal en el entorno de desarrollo de inglesdivino. Una vez hecho dicho marcaje, se descargan de inglesdivino [1] las marcas temporales junto con la transcripción (también marcada). El sistema será el encargado de asociar dichas marcas y transcripción con el audio descargado previamente (como se explicó en la sección anterior).

Por su parte, las marcas de posición hechas en la transcripción le servirán al sistema para saber asociar a cada fragmento de audio su respectiva transcripción. En la Figura 3.7 se muestra la forma en que queda marcada una transcripción (que es lo que finalmente se descarga). Cada frase entre asteriscos pertenece a un fragmento de audio distinto. El sistema será el encargado más adelante de asociar cada audio con su respectiva transcripción basándose en los marcajes realizados.

Close enough to start a war*
All that I have* is on the floor*
God only knows what* we're fighting for*
All that I say,* you always say more*

I can't keep up with your turning tables*
Under your thumb,* I can't breathe*

Figura 3.7: Marcas de posición en las transcripciones

3.4.3. Troceado

En esta etapa, el sistema empieza a involucrarse en el tratamiento previo de los datos de entrada. Una vez que se disponen de los audios y transcripciones, marcados de la forma como se explicó en la sección anterior, se procede a trocear los audios basándose en los marcajes. Para ello se hace uso del programa *Sox* como herramienta de recorte de audio. En la Figura 3.8 se puede apreciar cómo, a partir de los datos de entrada marcados, el sistema es capaz de dar diferentes segmentos de audio (los que sólo incluyen voz) con sus respectivas transcripciones.

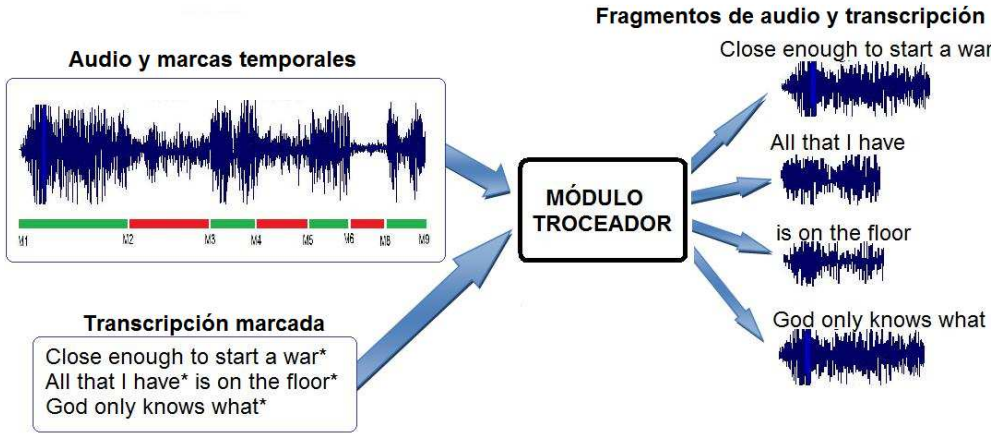


Figura 3.8: Proceso de troceado del audio y transcripción

3.5. Realización del alineamiento.

El alineamiento de texto y audio se realizará usando dos métodos: El primero está basado en modelos HMM preexistentes, y el segundo (sin modelos) está basado en el entrenamiento de los modelos HMM desde cero usando el audio a alinear (y posiblemente algunos audios complementarios similares). En ambos casos, los modelos serán usados o entrenados usando HTK [17]. En las siguientes secciones se explica con más detalles los métodos usados.

3.5.1. Alineamiento basado en modelos HMM preexistentes

En este método se usan modelos HMM de fonemas en inglés previamente entrenados con audios de 8 KHz (TIMIT corpus [18]). Dichos modelos han sido creados para cada fonema del idioma inglés con 40 gaussianas por estados y 3 estados por fonema. Para los experimentos que se realizarán en este proyecto, se usan los modelos sin ninguna modificación. A continuación se describen los pasos seguidos para la obtención de nuestros tiempos de interés.

1. Preparación de los datos de entrada. En el caso de nuestro proyecto dichos datos de entrada serán fragmentos de audio (canciones o noticias) con sus respectivas transcripciones (a nivel de palabra). La Figura 3.9 muestra una ilustración de los datos de entrada de un fragmento de audio en el que aparece la frase “Hello world”.

Figura 3.9: Fragmento de audio con su respectiva transcripción

2. Parametrización del audio. En esta etapa se convierten los ficheros de audio, inicialmente en formato *WAV,* al formato que manejan las herramientas de *HTK* [17]. Este nuevo formato contiene los MFCCs (*Mel Frecuency Cepstral Coefficients*) de los audios originales.

3. Conversión de la transcripción a nivel de palabra a una gramática a nivel de fonema. Para ello usa un diccionario fonético en inglés derivado del diccionario de pronunciación CMU [19]. Esta gramática se usa para la creación de una red de modelos de fonema HMM. En la Figura 3.10, se observa cómo mediante el uso de la herramienta HDMan, proporcionada por HTK, se logra dicha red de fonemas.

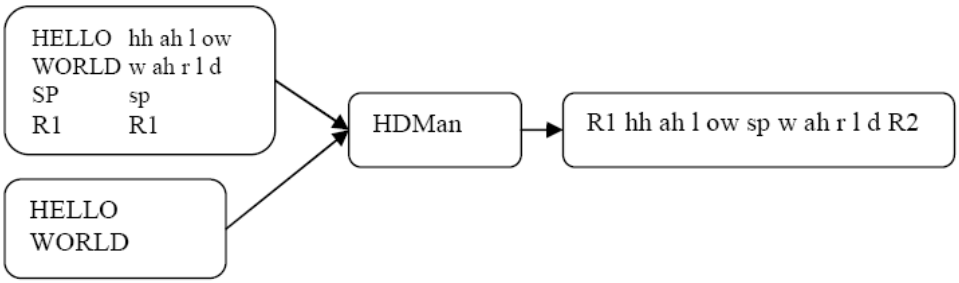


Figura 3.10: Creación de la gramática a nivel de fonema a partir de la transcripción a nivel de palabra y el diccionario.

4. Realización del alineamiento. El alineamiento es realizado por la herramienta HVite. Esta herramienta enfrenta el audio parametrizado con la red de HMMs creada, y da como resultado el tiempo de inicio y final para cada fonema y palabra. En la Figura 3.11 se muestra un ejemplo de la salida de HVite.

```
 7500000  8700000    f     FOUR
 8700000  9800000    ao
 9800000 10400000    r
10400000 10400000    -
10400000 11700000    s     SEVEN
11700000 12500000    eh
12500000 13000000    v
13000000 14400000    n
14400000 14400000    sp
```

Figura 3.11: Resultados arrojados por HVite

3.5.2. Alineamiento sin modelos: Alineando durante el entrenamiento

Este método está basado en el proceso de entrenamiento de los modelos HMM. Lo que hacemos aquí es entrenar modelos de fonemas a partir de los datos que queremos alinear. En nuestro caso estos datos serían canciones o noticias. Durante el proceso de entrenamiento, HVite y HERest son usados para realinear y reentrenar los modelos, dando como resultado un alineamiento a nivel de fonema de los datos de entrada. Comparado con el método anterior, éste tiene la ventaja de usar modelos acústicos completamente adaptados a los datos procesados con respecto al locutor, presencia de música, ruido de fondo, etc. Es bien sabido que los mejores resultados en reconocimiento se alcanzan cuando se intenta reconocer datos lo más parecidos a los usados en el entrenamiento. ¿Qué tal si intentásemos reconocer los mismos datos de entrenamiento?, el parecido sería del 100%. Esto es lo que precisamente se hace con este método. Normalmente usar datos de test para el entrenamiento no es justo, pero para esta aplicación en particular, esto es perfectamente válido. Usamos como datos de entrada (para el entrenamiento de los modelos) los datos (audio y texto) que queremos alinear, y luego como resultado del proceso de entrenamiento, obtenemos el alineamiento. Por supuesto, también hay ciertas desventajas. La principal es que usando solamente el audio y texto que queremos alinear estamos siendo muy ineficientes al utilizar una limitada cantidad de datos. Trataremos de aliviar esto añadiendo otros audios y textos del mismo locutor en similares condiciones (hasta donde sea posible) para mejorar el proceso de entrenamiento y alineamiento. A continuación se describen los paso seguidos en este método.

1. Preparar los datos de entrada como en el método anterior. Se preparan los audios y las transcripciones (transcripción a nivel de palabra). La principal novedad aquí es que puede que estemos interesados en preparar transcripciones y audios adicionales con similares características entre sí

(mismo locutor, condiciones acústicas, etc.) para ayudar al proceso de entrenamiento a mejorar sus resultados mediante la adición de más datos.

2. Parametrizamos los audios como en el método anterior, convirtiéndolos en MFCCs.

3. Se transforma la transcripción a nivel de palabra en una gramática a nivel de fonema, como en el método anterior, usando de nuevo un diccionario fonético en inglés derivado del diccionario de pronunciación CMU [19].

4. Con todos los datos necesarios preparados, procedemos a entrenar los modelos acústicos de cada fonema que aparece en la gramática que previamente hemos definido. Empezamos definiendo un prototipo de modelo y creando unos monofonemas "planos" usando la herramienta HERest de HTK. Luego, estos monofonemas "planos" son reestimados usando la herramienta HERest. El propósito de esto es cargar todos los monofonemas "planos" y reestimarlos usando los archivos MFCC generados a partir de nuestros datos de entrenamiento (audios de canciones o noticias) y crear así un nuevo set de modelos. Esta reestimación se hace cuatro veces.

5. En el paso final, se hace un realineamiento de los datos de entrenamiento usando la herramienta HVite. Esta herramienta es capaz de considerar todas las posibles pronunciaciones de cada palabra (en el caso en que una palabra tiene más de una pronunciación en la gramática), y luego dar como resultado la pronunciación que mejor encaje con el dato acústico. HVite nos da un primer alineamiento de los datos. Usamos este alineamiento para re-estimar los modelos y conseguir mejor precisión. Se hace la re-estimación (con HERest) cuatro veces más usando los resultados de HVite (el primer alineamiento). Después de este proceso, una vez hecho toda la re-estimación, ya tenemos los modelos listos y los usamos para alinear los datos de entrenamiento. A partir del alineamiento obtenido en este proceso, extraeremos los tiempos finales que serán comparados con la referencia manual.

3.6 Selección de los datos de interés

Como se ha insinuado en la explicación de alguna sección anterior, no todos los tiempos obtenidos en el alineamiento serán comparados con la referencia manual. Dado que el proceso de alineamiento manual se ha realizado a nivel de palabra, sólo de disponen de los tiempos, manualmente alineados, de comienzo y final de cada palabra. El alineamiento realizado por el sistema automático tiene un nivel de detalle mucho mayor en sus resultados, ya que los tiempos obtenidos son a nivel de fonema.

Es por eso que una vez realizado el alineamiento automático, procedemos a la extracción de los tiempos de inicio y fin de cada palabra. En el ejemplo de la Figura 3.12 estos tiempos serían el comienzo de los fonemas encerrados en un círculo rojo, que se corresponde justamente con el comienzo de la palabra de la que forman parte.

Tiempo inicio	Tiempo fin	Fonema	Palabra
7500000	8700000	f	FOUR
8700000	9800000	ao	
9800000	10400000	r	
10400000	10400000	-	
10400000	11700000	s	SEVEN
11700000	12500000	eh	
12500000	13000000	v	
13000000	14400000	n	
14400000	14400000	sp	

Figura 3.12: Tiempos de interés del alineamiento automático

Hay que tener claro una cosa, y es que al decir que se extraen los tiempos de inicio y fin de cada palabra, en realidad estamos extrayendo sólo los tiempos de inicio, que implícitamente supone también extraer los tiempos finales, ya que donde empieza una palabra, termina la anterior (dado que la mayoría de los fragmentos carecen de pausas entre medias de las palabras). En definitiva, únicamente los tiempos de comienzo de cada palabra serán los que se comparen con la referencia manual, ya que son los que realmente interesan en inglesdivino [1].

3.6 Fusión de este PFC con inglesdivino.

En este apartado se explicará la relación que tiene este Proyecto fin de Carrera con inglesdivino [1]. Aunque a estas alturas ya se tiene claro que el principal factor motivador de este proyecto ha sido la necesidad de un alineamiento automático en inglesdivino, en este apartado se entrará un poco más en detalle en cómo se ha solventado ese problema y cómo los resultados de este proyecto se aplican a dicho sitio web.

3.6.1 Datos en el servidor de inglesdivino

Como ya se ha mencionado anteriormente, para la realización de este proyecto, parte de los datos utilizados en los experimentos serán extraídos de la base de datos de inglesdivino [1]. Concretamente se extraerán las marcas temporales de los audios y las transcripciones (con marcas de posición). Estos datos se utilizarán en la etapa de preparación de los datos de entrada. En la fase final de los experimentos será necesario disponer de las referencias manuales para realizar las medidas del

alineamiento automático. Por eso será también necesario descargas dichos tiempos de la base de datos de inglesdivino.

Lo anterior se refiere a datos de inglesdivino que contribuyen al desarrollo de este proyecto. Pero no son únicamente datos lo que aporta el sitio web, sino también un entorno específico para el marcaje tanto en los audios como en la transcripción. El entorno permite a partir de un vídeo (musical o de noticas) y su transcripción obtener estas marcas mediante las herramientas que ahí se proporcionan. Los audios con los que se trabajará serán obviamente extraídos de esos vídeos que se utilizan para el marcaje. Es gracias al entorno también que se pueden obtener alineamientos manuales de una manera muy cómoda.

En la Figura 3.13 se ve la relación de todos los elementos citados anteriormente. En ella se ve como a partir de un vídeo musical y su transcripción y la ayuda del entorno desarrollado en inglesdivino, se obtienen las marcas temporales, la transcripción marcada y el alineamiento manual. Cabe recalcar que las tres salidas del entorno que aparecen en la figura se realizan todas de forma manual. El entorno únicamente proporciona las herramientas necesarias para realizar dichas labores.

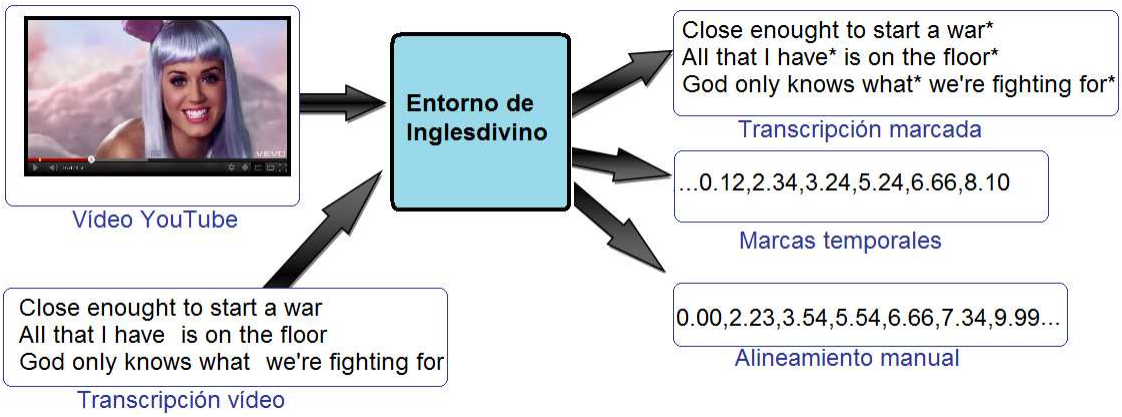


Figura 3.13: Entrada y salida del entrono de inglesdivino

3.6.2 Datos en el cliente: ordenador personal

Una vez que se descargan de inglesdivino las transcripciones marcadas, las marcas temporales y el alineamiento manual, se descargan también de YouTube el vídeo correspondiente a la transcripción, vídeo del que posteriormente se extraerá el audio en formato WAV con el que se trabajará (Ver sección 3.4).
Llegados a este punto, disponemos de los audios troceados y de sus correspondientes transcripciones. Estos serán los datos de entrada definitivos del sistema Alineador.

En la Figura 3.14 se puede observar todo el proceso que se realiza en la parte del cliente, o lo que es lo mismo una vez que se descargan los datos desde inglesdivino y desde el portal YouTube [14].

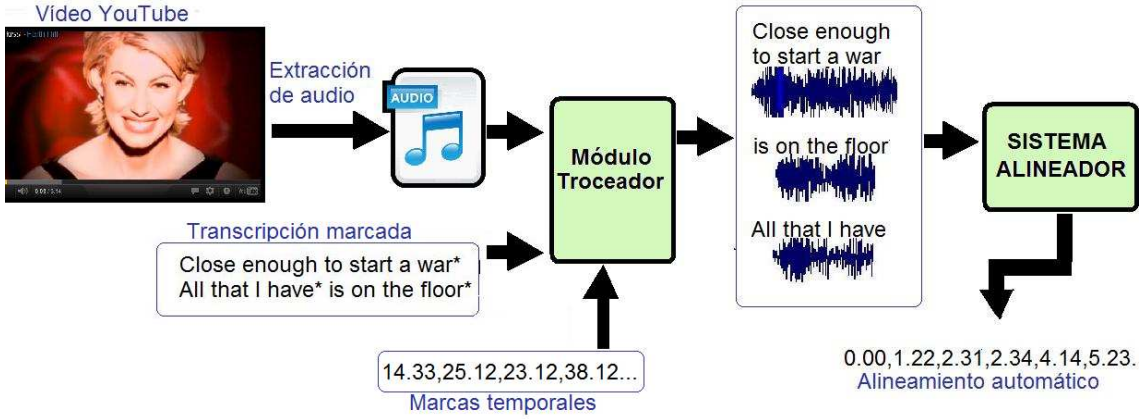


Figura 3.14: Proceso de alineamiento en el cliente

3.6.3 Sincronización entre cliente y servidor.

Una vez que se han descargado los datos desde inglesdivino y se ha realizado el alineamiento automático, es necesario volver a subir este alineamiento (secuencia de tiempos) al servidor de inglesdivino. Estos tiempos obtenidos automáticamente tendrán que pasar previamente por el entorno de inglesdivino para ser corregidos de los posibles errores que puedan tener. Una vez corregidos, se pasan a mostrarse en el sitio oficial de inglesdivino [1].

Todo el procedimiento de descarga y subida a inglesdivino se realiza mediante una conexión ftp desde el cliente. Se han desarrollado los scripts necesarios para que dada una lista de canciones o noticias (previamente tratadas en inglesdivino) el sistema automáticamente se capaz de descargar del servidor, tratarlo en el cliente, y volver a subirlo al servidor.

3.7 SISTEMA COMPLETO

En la siguiente Figura 3.15 se muestra todo el procedimiento llevado a cabo desde que se decide alinear una canción hasta obtener su resultado final. En la figura se incluyen etapas que ya han sido explicadas anteriormente con mayor detalle. En esta sección lo que se pretende es tener una visión global del sistema completo, así como de la relación que existe entre inglesdivino y este PFC.

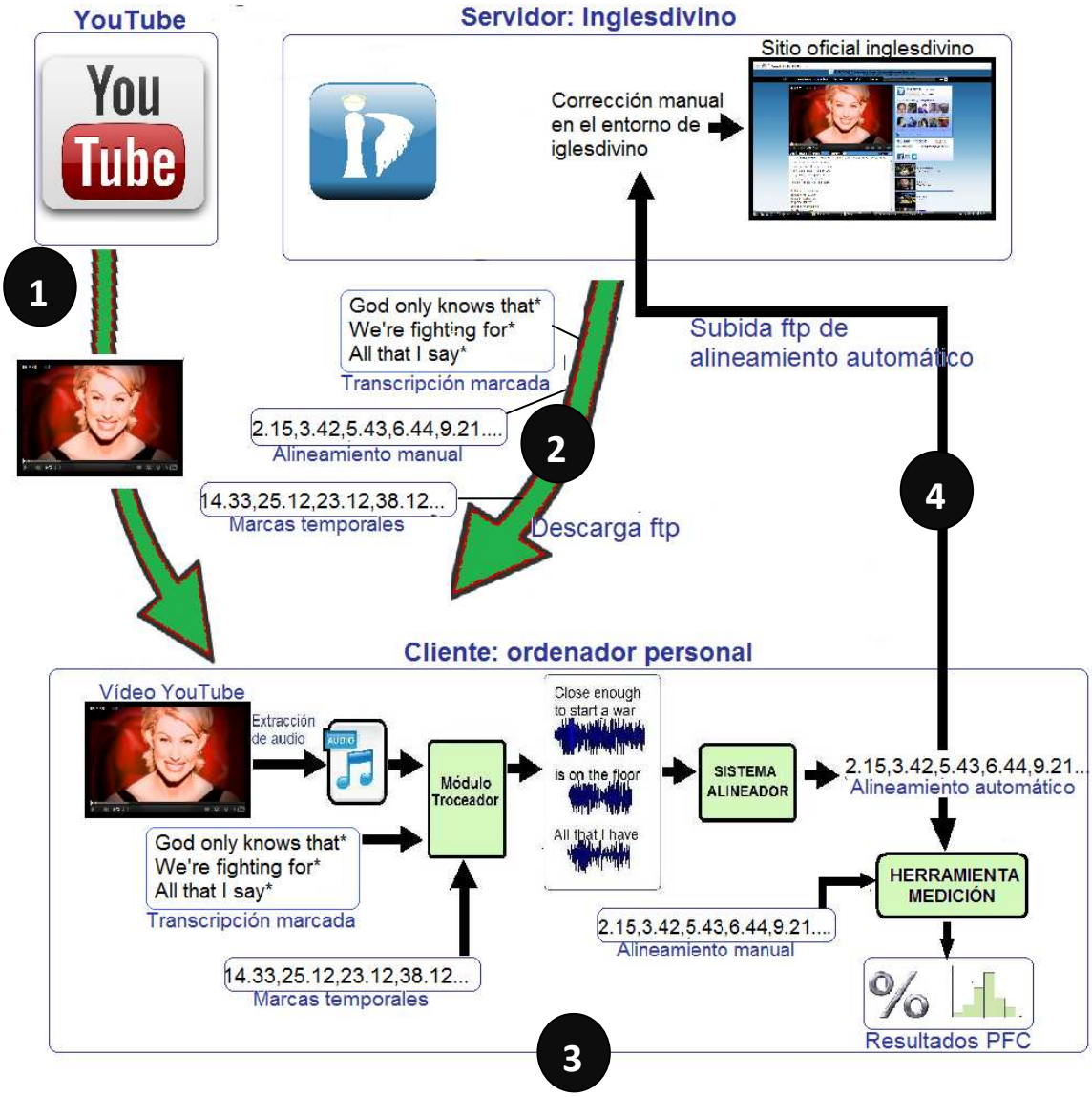


Figura 3.15: Funcionamiento del sistema completo

4

EXPERIMENTOS Y RESULTADOS

Los experimentos se realizarán con audios de canciones y noticias en inglés. Para los experimentos con noticias, se han elegido cuatro fragmentos de YouTube pertenecientes a cuatro locutores diferentes: dos mujeres y dos hombres. Los audios duran alrededor de un minuto y medio. Con respecto a las canciones, se han elegido 3 canciones que cubren diferentes estilos: La primera es una canción muy rápida (rap), la segunda tiene una velocidad normal (pop) y la última, una velocidad muy lenta (balada). Los experimentos para el método libre de modelos se realizarán con audios de frecuencia de muestreo igual a 44100 Hz y 8000 Hz, esta última nos será útil para comparar resultados con el primer método (dependiente de modelos). Dos serán los tipos de experimentos que se llevarán a cabo con este segundo método (libre de modelos). El primero consiste en utilizar como datos de entrada sólo la canción o fragmento de noticia que se desea alinear, y el segundo, en añadir audios adicionales que ayuden en el proceso de entrenamiento a mejorar los resultados del alineamiento; en decir, a parte del audio que queremos alinear, introduciremos más audios del mismo locutor o del mismo cantante. Estos audios adicionales son usados sólo para mejorar la precisión del alineamiento.

En el caso de las canciones, además, se hará un estudio del grado de influencia de la música instrumental de fondo en las canciones. Por ello, cada canción será primeramente tratada en su versión acapella, y, posteriormente, se incluirá la música instrumental.

4.1 RESULTADOS BASADOS EN EL ENTRENAMIENTO DE HMMs

Primeramente mostraremos los resultados obtenidos usando el método libre de modelos, y en la sección 4.2, los resultados del método dependiente de modelos. Los resultados se presentan mostrando el porcentaje de palabras con errores de segmentación menor que un cierto valor de tolerancia, que en este caso son los siguientes: 50, 100 y 200 ms. Se han elegido estos valores dado que la aplicación a la

que van dirigidos es relativamente robusta a ciertos errores. Estas métricas de evaluación son similares a las utilizadas en [3].

Tras un análisis subjetivo de la degradación que sufre el alineamiento automático ante el oído y el ojo humano, se ha visto que para errores menores de 100 ms dichos errores son prácticamente imperceptibles. Se ha visto también que esta percepción depende en gran medida de la velocidad de la canción, siendo más notorio en canciones lentas, y menos perceptibles cuanto más rápida es un audio. Si permitimos errores de valores menores a 100 ms, éstos son imperceptibles en ambos casos.

4.1.1. Alineamiento en noticias y canciones acapella

Las tablas 4.1 y 4.2 muestran una comparación de resultados obtenidos en los experimentos con noticias y canciones usando como datos de entrada sólo un audio y con audios adicionales. Estos resultados muestran que, aunque hay algunos casos en los que el método libre de modelos funciona muy bien incluso con un audio, es cuando se tiene acceso a otros audios adicionales del mismo locutor o del mismo cantante donde este método logra un mejor comportamiento. Para ilustrar esta mejora, la Figura 4.1 muestra un histograma del valor absoluto de los errores encontrados en el alineamiento del LOCUTOR 1 (de la Tabla 4.1) cuando no hay audios adicionales y cuando se añade un audio adicional más. Como se puede ver, el error en el alineamiento se reduce considerablemente al añadir más datos.

	Un solo audio			Con audios adicionales		
Tolerancia	50 ms	100 ms	200 ms	50 ms	100 ms	200 ms
LOCUTOR 1 (femenino)	(188 palabras y 1 audio)			(895 palabras and 4 audios)		
	7.45	19.68	19.68	29.79	50.53	93.09
LOCUTOR 2 (femenino)	(223 palabras y 1 audio)			(1181 palabras y 4 audios)		
	24.50	49.67	91.06	28.81	55.30	96.69
LOCUTOR 3 (masculino)	(319 palabras y 1 audio)			(1486 palabras y 3 audios)		
	1.57	3.13	10.97	35.11	57.68	96.87
LOCUTOR 4 (masculino)	(318 palabras and 1 audio)			(950 palabras y 3 audios)		
	2.52	5.03	8.18	3.46	5.35	9.12

Tabla 4.1: Porcentaje de palabras (%) en noticias con errores menores que tres valores de tolerancia (50, 100 y 200 ms) con sólo un audio y con audios adicionales.

	Un solo audio			Con audios adicionales		
Tolerancia	50 ms	100 ms	200 ms	50 ms	100 ms	200 ms
Cantante 1 (canción rápida)	(1 canción y 794 palabras)			(2 canciones y 1801 palabras)		
	27.71	56.55	88.54	28.72	59.07	92.44
Cantante 2 (canción normal)	(1 canción y 398 palabras)			(2 canciones y 767 palabras)		
	38.94	65.08	80.90	41.96	73.12	92.46
Cantante 3 (canción lenta)	(1 canción y 172 palabras)			(2 canciones y 412 palabras)		
	28.00	49.42	66.86	31.40	51.74	68.60

Tabla 4.2: Porcentaje de palabras (%) en canciones con errores menores que tres valores de tolerancia (50, 100 y 200 ms) con solo un audio y con audios adicionales.

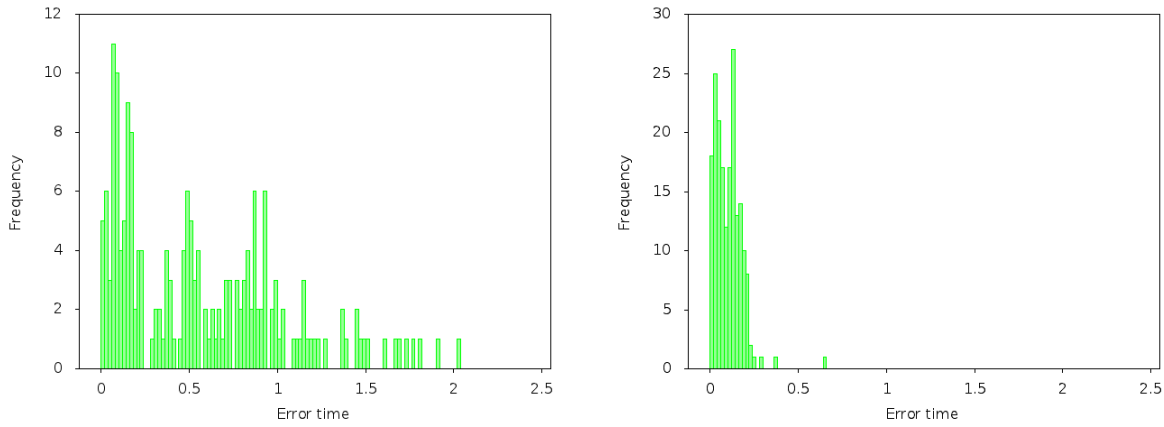


Figura 4.1: Comparación de los tiempos de error (en segundos) en los casos donde el alineamiento es realizado usando sólo un audio (izquierda), y cuando se añade un audio adicional más (derecha).

Es importante notar que hay noticias (locutor 4) y canciones (canción 3) que pueden resultar particularmente problemáticos para este método. En ambos casos se ha encontrado que el audio es lento, lo que parece ser particularmente problemático para este método.

4.1.2. Influencia de la longitud de los audios y la adición de locutores distintos

Dentro del análisis de las canciones, en la Tabla 4.2 se puede apreciar que es con las canciones lentas cuando este método arroja peores resultados. A pesar de ser los peores resultados en comparación con las demás, se pueden considerar aceptables para la aplicación a la que van destinadas, ya que hay un acierto de algo más del 50% (para errores menores de 100ms). No se puede decir lo mismo en el caso del locutor 4 en la Tabla 4.1, donde la diferencia entre el porcentaje de acierto de las "buenas alineaciones" y la del locutor 4 son muy notorias.

A continuación se intentará encontrar las causas del pésimo resultado del locutor 4. Para este análisis, primero resaltaremos las peculiaridades del audio 4 con respecto a los demás fragmentos de noticias:

- La media de las duraciones de las noticias está alrededor de un minuto y medio, salvo en el caso del locutor 4, donde la duración es de dos minutos y 7 segundos exactamente.
- En el caso del locutor 4, hay un momento de aproximadamente 20 segundos donde interviene un segundo locutor.

Teniendo en cuenta los puntos anteriores, surge la sospecha de que tiene que haber algún otro factor degradante, a parte de los audios lentos, que afecta en gran medida a los resultados del alineamiento. Empezaremos por analizar la influencia de la intervención de un segundo locutor. Lo que vamos a intentar es ver si la introducción de una segunda voz afecta considerablemente al alineamiento o no. Para ello, hacemos el alineamiento del mismo audio, pero esta vez eliminando la segunda voz.

	(950 palabras y 3 audios)		
Tolerancia	50 ms	100 ms	200 ms
LOCUTOR 4 (Sin voz secundaria eliminada)	3.46	5.35	9.12
LOCUTOR 4 (Con voz secundaria eliminada)	1.26	3.14	8.18

Tabla 4.3: Porcentaje de palabras (%) en el locutor 4 menores que tres valores de tolerancia (50, 100 y 200 ms) eliminando la voz secundaria.

En la Tabla 4.3 se puede ver el resultado de este nuevo alineamiento. Parece ser que la segunda voz no tenía nada que ver en los malos resultados, más bien al contrario, parece que ayudaba a mejorar los resultados al comportarse como dato adicional. Más adelante se hará el alineamiento con este mismo locutor, pero esta vez insertando locutores distintos de duración más larga. Pero antes de pasar a eso, analizaremos el otro posible factor degradante que nos queda, la duración del audio.

Los que hasta ahora se ha venido haciendo es introducir los fragmentos de noticias sin trocearlos, ya que no había la necesidad, dado que en los audios de noticias con los que trabajamos la voz es muy fluida y no hay pausas largas, por lo que no es necesario eliminar ningún trozo y seleccionar fragmentos. No ocurre lo mismo con las canciones, donde es frecuente encontrar pausas considerables después de cada frase o estrofa, es por ellos que es necesaria una fragmentación manual previa para eliminar dichas pausas largas. Normalmente los fragmentos resultantes tendrán una duración media de 3 segundos, por lo que no tenemos el problema de audios largos en el caso de las canciones. Sin embargo, en el caso de las noticias, los audios son muchos más largos, ya que no se ha troceado y solamente existe un único fragmento. Estos fragmentos, como se ha dicho antes, tienen una duración de 1 minuto y medio, salvo el del locutor 4, que dura 2 minutos y 7 segundos. Es por ello que trocearemos este audio aproximadamente por mitad con el objetivo de obtener dos fragmentos de una duración similar a los demás audios de noticias. En la Tabla 4.4 se pueden ver los resultados de este experimento.

	(950 palabras y 3 audios)		
Tolerancia	50 ms	100 ms	200 ms
LOCUTOR 4 (un fragmento)	3.46	5.35	9.12
LOCUTOR 4 (dos fragmentos)	22.33	43.40	81.13

Tabla 4.4: Porcentaje de palabras (%) en el locutor 4 menores que tres valores de tolerancia (50, 100 y 200 ms) y con audio de noticias fragmentado.

Parece que ahora sí hemos dado con la causa de tan malos resultados, y es que los audios no pueden ser tan largos, es necesaria una fragmentación en tal caso. A pesar de estas mejoras, los resultados siguen siendo ligeramente inferiores a los demás audios. Esto se debe a, como ya se ha dicho antes, la velocidad del audio. Con audios lentos, tiene un resultado ligeramente inferior comparado con los audios normales o rápidos.

A continuación se hará otro experimento que consistirá, como ya se anunció antes, en la adición de un audio sobre el audio del locutor 4. El audio que añadiremos será de un locutor diferente. Primero haremos el experimento añadiendo el audio de un locutor del mismo género (locutor 3), y luego con uno distinto (locutor 1). Los experimentos se harán partiendo de las condiciones expuestas en el apartado anterior; es decir, con el audio fragmentado y sin quitar la segunda voz presente en el audio. En la Tabla 4.5 se muestra el resultado del caso en el que se añade un audio de un locutor del mismo género, y en la Tabla 4.6, el caso en el que el locutor del audio adicional es de distinto género (femenino).

	(1269 palabras y 4 audios)		
Tolerancia	50 ms	100 ms	200 ms
LOCUTOR 4	22.33	43.40	81.13
LOCUTOR 4 (audio adicional: Locutor 3)	25.79	46.54	91.82

Tabla 4.5: Porcentaje de palabras (%) en el locutor 4 menores que tres valores de tolerancia (50, 100 y 200 ms) añadiendo un audio adicional del mismo género.

	(1138 palabras y 4 audios)		
Tolerancia	50 ms	100 ms	200 ms
LOCUTOR 4	22.33	43.40	81.13
LOCUTOR 4 (audio adicional: Locutor 1)	23.58	42.45	82.39

Tabla 4.6: Porcentaje de palabras (%) en el locutor 4 menores que tres valores de tolerancia (50, 100 y 200 ms) añadiendo un audio de distinto género.

Observando las tablas anteriores, vemos que los audios adicionales, ya sean del mismo locutor o no, contribuyen a mejorar el alineamiento. También se observa, que los resultados son mejores si los datos adicionales son de un locutor del mismo género.

4.1.3. Comparación entre canciones acapella y no acapella

Hasta el momento se ha experimentando con noticias y canciones acapella. Para un estudio comparativo esto está muy bien, pero en la práctica lo que nos interesa es alineamiento sobre canciones que incluyan música instrumental. En el siguiente apartado se analizarán las mismas canciones anteriores, pero esta vez ya no acapella, sino que con la música instrumental incluida. Los resultados se mostrarán de forma comparativa, es decir, los nuevos resultados de estos experimentos se mostrarán junto con los obtenidos anteriormente (canciones acapella) para poder hacernos una idea del grado de degradación que sufren. De antemano ya suponemos que los resultados serán peores, ya que al incluir música de fondo (que se podría interpretar como ruido de fondo) el reconocimiento de fonemas es mucho más complicado y, por tanto, el alineamiento también. En la Tabla 4.7 se muestra el resultado de haber realizado un alineamiento sobre canciones acapella y no acapella. El experimento se ha hecho considerando sólo un audio como dato de entrada. Como cabía esperar, los resultados en las canciones que incluyen la música instrumental son algo peores. En la siguiente sección veremos cómo podemos mejorar estos resultados gracias añadir más canciones del mismo cantante.

	Resultados (acapella y con música instrumental)			
	Tolerancia	50 ms	100 ms	200 ms
CANTANTE 1 (canción rápida)	Acapella	27.71	56.55	88.54
	No acapella	22.46	47.12	81.43
CANTANTE 2 (canción normal)	Acapella	38.94	65.08	80.90
	No acapella	26.88	43.97	68.84
CANTANTE 3 (canción lenta)	Acapella	28.00	49.42	66.86
	No acapella	19.77	36.05	54.07

Tabla 4.7: Porcentaje de palabras (%) en canciones acapella y no acapella menores que tres valores de tolerancia (50, 100 y 200 ms).

4.1.4 Influencia del número de canciones adicionales en canciones no acapella

Nuestro objetivo ahora es mejorar los resultados obtenidos sobre las canciones no acapella de la Tabla 4.7. Para el estudio de este caso, haremos el análisis sólo sobre la canción normal y lenta, ya que se sobreentiende que si se produce un mejoramiento para estos casos, en el caso de canciones rápidas también lo habrá, ya que este tipo de audios son los que menos inconvenientes presentan.

De la Tabla 4.8 podemos ver que un mayor número de canciones no necesariamente significa una mejora de resultados progresiva. Lo que sí está claro es que un alineamiento es mejor con audios adicionales, pero determinar el número adecuado de canciones o los estilos adecuados es algo muy complicado. En la Tabla 4.8 también se puede observar que hay canciones que al ser añadidas (canción 6, por ejemplo) con el propósito de mejorar el alineamiento, producen el efecto contario. El audio añadido puede resultar a veces “dañino” por razones como la variación del estilo musical, variación de su velocidad, presencia de más ruido en el audio adicional, etc.

	Tolerancia / Numero de canciones	50 ms	100ms	200ms
CANTANTE 2 (canción normal)	1	25.13	48.99	76.38
	2	27.64	52.01	80.40
	3	26.88	55.03	82.01
	4	28.14	56.28	85.93
	5	30.65	53.77	82.66
	6	29.65	51.76	79.40
	7	32.91	59.80	84.67
	8	31.16	48.47	84.17

Tabla 4.8: Comparación de resultados usando diferente número de audios (canciones) adicionales para el cantante 2. La tabla muestra el porcentaje de palabras (%) con errores menores a 3 valores de tolerancia (50, 100 y 200 ms)

A continuación se analiza el efecto que tiene la adición de audios adicionales sobre la canción 3. En este caso haremos un experimento para ver si el hecho de eliminar un audio "dañino" mejora el alineamiento.

Los resultados del alineamiento de la canción 3 sin añadir audios adicionales vimos que era lo que se muestra en la Tabla 4.9.

Tolerancia	50 ms	100 ms	200 ms
CANCIÓN 3	19.77	36.05	54.07

Tabla 4.9: Porcentaje de palabras (%) de la canción 3 menores que tres valores de tolerancia (50, 100 y 200 ms), incluyendo música instrumental y sólo un audio de entrada.

Ahora bien, si añadimos otra canción a la canción 3, obtenemos la Tabla 4.10. En esta tabla se puede ver que la adición de este audio adicional ha empeorado los resultados.

Tolerancia	50 ms	100 ms	200 ms
CANCIÓN 3	18.70	37.56	57.65

Tabla 4.10: Porcentaje de palabras (%) de la canción 3 menores que tres valores de tolerancia (50, 100 y 200 ms), incluyendo música instrumental y dos audios (canciones) de entrada.

Dado que no hemos tenido éxito añadiendo el audio anterior, añadiremos otro audio más (en total 3 canciones). El resultado se puede apreciar en la Tabla 4.11.

Tolerancia	50 ms	100 ms	200 ms
CANCIÓN 3	15.70	36.63	60.47

Tabla 4.11: Porcentaje de palabras (%) de la canción 3 menores que tres valores de tolerancia (50, 100 y 200 ms), incluyendo música instrumental y tres audios (canciones) de entrada.

En el caso de la Tabla 4.11, vemos que los resultados sí que han mejorado. Si nos dejamos llevar por la intuición es algo lógico pensar que si eliminamos el audio que empeoraba los resultados, los resultados de la Tabla 4.11 serán aún mejores. En la Tabla 4.12 se muestran los resultados de este experimento. Como se puede ver, nuestra intuición estaba equivocada, ya que los resultados son peores. Esto se explica debido a que en cierta manera al quitar una canción, estamos quitando fonemas con los que caracterizar mejor los modelos HMM, de ahí que los resultados empeoren. Es por esto que es difícil determinar qué canciones y qué parámetros son los más adecuados para mejorar los resultados de la forma más eficiente posible.

Tolerancia	50 ms	100 ms	200 ms
CANCIÓN 3	15.70	29.65	55.81

Tabla 4.12: Porcentaje de palabras (%) de la canción 3 menores que tres valores de tolerancia (50, 100 y 200 ms), eliminando audio "dañino".

En definitiva, según lo observado en esta sección, hasta que no se haga un estudio en profundidad de qué factores influyen en la degradación y en la optimización del alineamiento a la hora de añadir audios adicionales, lo más aconsejable es añadir cuanto más audios mejor, y, si es posible, muy parecidos y libres, en la medida de lo posible, de excesivo ruido. Esta es la metodología que se seguirá en inglesdivino [1] para alinear álbumes enteros de cantantes, por ejemplo. Si queremos alinear una canción de un determinado cantante, lo que se hará es reunir todas las posibles canciones de ese cantante y emplearlas como audios adicionales. Teniendo en cuenta que cuando se realiza el alineamiento de una canción en particular con audios adicionales, dichos audios adicionales también se alinean, tendremos todas las canciones que se han usado como entrada alineadas automáticamente.

4.2 EXPERIMENTOS BASADOS EN EL USO DE MODELOS HMM PREEXISTENTES

En esta sección se muestran los experimentos y resultados obtenidos con el método basado en modelos HMM preexistentes. También se mostrarán los resultados obtenidos en la sección anterior (método libre de modelos) a fin de establecer una comparación entre ellos. Dado que los resultados anteriores han sido obtenidos con audios de frecuencia igual a 44100 Hz, y teniendo en cuenta que para el método

basado en modelos preexistentes es necesario trabajar con audios de 8KHz (debido a la disponibilidad de los modelos entrenados en nuestro caso particular), necesitamos re-muestrear los audios con los que se ha experimentado para que la comparación entre el métodos sea justa.

En lo que sigue, se mostrarán los resultados del primer y segundo método. Para este último, tanto para audios de 8 KHz, como para audios de 44100Hz.

4.2.1 Alineamiento en audios de noticias

En la Tabla 4.13 se puede ver que para el caso de las noticias, si trabajamos con audios de 8KHz, el método basado en modelos preexistentes nos da unos resultados ligeramente mayores, pero éstos se ven sobrepasados por el segundo método en cuanto usamos audios de una frecuencia de muestreo mayor.

	Resultados			
LOCUTOR 1 (femenino)	Tolerancia	50 ms	100 ms	200 ms
	Modelos preexistentes (8000Hz)	28.72	59.57	90.96
	Libre de modelos (8000HZ)	28.72	48.40	80.32
	Libre de modelos (44100 Hz)	29.79	50.53	93.09

	Resultados			
LOCUTOR 2 (femenino)	Tolerancia	50 ms	100 ms	200 ms
	Modelos preexistentes (8000Hz)	30.46	65.23	86.42
	Libre de modelos (8000HZ)	26.49	54.30	95.70
	Libre de modelos (44100 Hz)	28.81	55.30	96.69

	Resultados			
LOCUTOR 3 (masculino)	Tolerancia	50 ms	100 ms	200 ms
	Modelos preexistentes (8000Hz)	34.80	67.40	94.36
	Libre de modelos (8000HZ)	34.80	56.43	95.92
	Libre de modelos (44100 Hz)	35.11	57.68	96.87

	Resultados			
LOCUTOR 4 (masculino)	Tolerancia	50 ms	100 ms	200 ms
	Modelos preexistentes (8000Hz)	26.73	55.66	88.05
	Libre de modelos (8000HZ)	21.18	41.98	80.21
	Libre de modelos (44100 Hz)	22.33	43.40	81.13

Tabla 4.13: Comparación de diferentes métodos y frecuencias de muestreo para noticias. La tabla muestra el porcentaje de palabras (%) con errores menores que tres valores de tolerancia (50, 100 y 200 ms). Para el método libre de modelos se usan audios adicionales.

4.2.2 Alineamiento en canciones acapella

A continuación se muestran resultados similares a los de la Tabla 4.13, pero esta vez para el caso de canciones acapella.

	Resultados (acapella)			
CANTANTE 1 (canción rápida)	Tolerancia	50 ms	100 ms	200 ms
	Modelos preexistentes (8000Hz)	17.23	32.15	76.57
	Libre de modelos (8000HZ)	27.23	57.98	90.78
	Libre de modelos (44100 Hz)	28.72	59.07	92.44

	Resultados (acapella)			
CANTANTE 2 (canción normal)	Tolerancia	50 ms	100 ms	200 ms
	Modelos preexistentes (8000Hz)	40.45	61.56	78.14
	Libre de modelos (8000HZ)	40.23	72.11	90.98
	Libre de modelos (44100 Hz)	41.96	73.12	92.46

	Resultados (acapella)			
CANTANTE 3 (canción lenta)	Tolerancia	50 ms	100 ms	200 ms
	Modelos preexistentes (8000Hz)	15.70	37.79	54.07
	Libre de modelos (8000HZ)	30.12	48.99	66.23
	Libre de modelos (44100 Hz)	31.40	51.74	68.60

Tabla 4.14: Comparación de diferentes métodos y frecuencias de muestreo para canciones acapella. La tabla muestra el porcentaje de palabras (%) con errores menores que tres valores de tolerancia (50, 100 y 200 ms). Para el método libre de modelos se usan audios adicionales.

En la tabla de arriba (Tabla 4.14) vemos que hay una notable diferencia con respecto a los resultados obtenidos con los auidos de noticias de la Tabla 4.13. En este caso, el segundo método (libre de modelos) es mejor tanto en el caso en el que el audio es de frecuencia igual a 8KHz como para el caso en que es igual a 44,1KHz.

4.2.2 Alineamiento en canciones no acapella

A continuación se hará un análisis similar al realizado en la Tabla 4.14 en canciones que incluyen audio instrumental. De antemano ya presuponemos que la relación en los resultados entre los distintos métodos y frecuencias serán similares a los que obtienen con canciones acapella, pero serán algo peores, debido a que esta vez las canciones incluyen música intrumental.

	Resultados (con música instrumental)			
CANTANTE 1 (canción rápida)	Tolerancia	50 ms	100 ms	200 ms
	Modelos preexistentes (8000Hz)	13.98	33.38	75.57
	Libre de modelos (8000HZ)	21.23	40.97	90.78
	Libre de modelos (44100 Hz)	24.76	42.31	91.18

	Resultados (con música instrumental)			
CANTANTE 2 (canción normal)	Tolerancia	50 ms	100 ms	200 ms
	Modelos preexistentes (8000Hz)	7.04	16.33	44.22
	Libre de modelos (8000HZ)	21.36	47.00	79.26
	Libre de modelos (44100 Hz)	27.64	52.01	80.40

	Resultados (con música instrumental)			
CANTANTE 3 (canción lenta)	Tolerancia	50 ms	100 ms	200 ms
	Modelos preexistentes (8000Hz)	13.56	26.98	54.09
	Libre de modelos (8000HZ)	13.21	31.92	57.23
	Libre de modelos (44100 Hz)	15.70	36.63	60.47

Tabla 4.15: Comparación de diferentes métodos y frecuencias de muestreo para canciones acapella. La tabla muestra el porcentaje de palabras (%) con errores menores que tres valores de tolerancia (50, 100 y 200 ms). Para el método libre de modelos se usan audios adicionales.

Comprobamos que las suposiciones iniciales son ciertas: los porcentajes son algo peores en general. Pero además notamos algo muy importante, y es que las diferencias entre el primer y el segundo método son muchos mayores, siendo mejores las del segundo método (libre de modelos).

5

CONCLUSIONES Y TRABAJO FUTURO

En esta sección se extraerán las conclusiones más importantes de los resultados observados en los experimentos de la sección anterior. También se propondrá un trabajo futuro acerca de las posibles líneas sobre las que se puede seguir investigando en este campo.

5.1 Conclusiones

Como se esperaba, los resultados en las noticias son mejores que los obtenidos en las canciones. Los resultados también muestran (Tabla 4.1) que en el caso de noticias de larga duración y velocidad lenta (caso del locutor 4), el método basado en modelos preexistentes es bastante más robusto que el método libre de modelos, que falla completamente en el alineamiento de este tipo de audios. Pero también se observa (Tabla 4.4) que esta precariedad puede ser solucionada gracias a la fragmentación de este tipo de audios en fragmentos de duración menor. Por lo general, la calidad del alineamiento de los dos métodos es similar para el caso de las noticias, siendo el segundo método (libre de modelos) ligeramente mejor gracias a que permite la utilización de audios con mayor frecuencia de muestreo. Aquí se debe aclarar que para hacer la comparación más justa, se ha reducido la frecuencia de los audios empleados en segundo método (libre de modelos) a 8KHz, ya que los modelos HMM utilizados en el primer método (basado en modelos preexistentes) han sido entrenados con audios a dicha frecuencia. Pero ha de quedar claro que con el segundo método no estamos limitados con la frecuencia de los audios, y, por tanto, se puede sacar ventaja de ello. Este hecho se puede observar perfectamente en la Tabla 4.13, donde gracias a utilizar audios de mayor frecuencia (44100Hz) se logran mejores resultados que el primer método. En definitiva, si se usan audios de la misma frecuencia, el primer método es mejor, si usamos audios de mayor frecuencia, como nos permite el segundo método, logramos que el segundo sea mejor. Nos queda la duda de si empleasen modelos entrenados con audios de mayor frecuencia, 44100Hz, por ejemplo, se seguiría

manteniendo esta relación, pero debido a que no se dispone de dichos modelos en este proyecto, no lo podemos comprobar.

Otra de las conclusiones que se puede observar en el alineamiento de noticias es que la intervención de un segundo locutor distinto al del audio que está siendo alineado es beneficiosa y no perjudicial. Este hecho se ve en la Tabla 4.5 y 4.6. En las mismas tablas, se puede apreciar también que es más beneficioso añadir locutores del mismo género que del género contrario.

En el caso de las canciones, todo parece indicar que el método libre de modelos funciona mucho mejor que el método basado en modelos preexistentes, y especialmente cuando se incluye música instrumental en las canciones. Los resultados parecen indicar que, cuanto más rápida es una canción, mejores son los resultados obtenidos. Las canciones que son lentas obtienen unos resultados peores, pero no están tan alejados de los resultados generales. También se ha visto, como era de esperar, que los alineamientos son mejores en las canciones acapella que las que incluyen música instrumental. Con respecto a los métodos empleados, se ve que el segundo método (libre de modelos) es sin duda bastante mejor para el caso de este tipo de audios (canciones) (ver Tabla 4.14). Pero es, sobre todo, cuando se trata de audios no acapella cuando las diferencias a favor del segundo método son más evidentes (Tabla 4.15). En el caso de las canciones lentas, el primer método falla notablemente con respecto al segundo, ya sea canciones acapella o no. Esto es lógico, ya que el método basado en modelos preexistentes está entrenado sólo con voz, mientras que el segundo método incorpora de forma natural la música y las condiciones ambientales durante el entrenamiento.

En la Tabla 4.8 se analiza en qué medida la introducción de audios adicionales mejora los resultados. Los resultados muestran que hay una tendencia hacia el mejoramiento de los mismos, sin embargo, esta tendencia no es monótona y hay máximos y mínimos que sugieren que algunos audios puede que ayuden mientras que otros realmente empeoran el comportamiento. En este caso en particular, encontramos que el primer máximo (en 100 y 200ms) con cuatro audios, pero en otros experimentos se ha encontrado el máximo con sólo un audio adicional.

5.2 Trabajo futuro

Como trabajo futuro nos gustaría profundizar en nuestro análisis, extendiendo los experimentos a un mayor número de canciones y noticias, y encontrar maneras de mejorar el alineamiento de los audios lentos en el método libre de modelos. También se podría extender el estudio a encontrar los factores ideales, como el número de canciones, estilo de música, número de canciones adicionales, etc. que maximizan los resultados.

Otras cosas más concretas que podrían llevarse a cabo, en el caso de las canciones en particular, es establecer un preprocesado de los audios similar al que se hace en [8], donde previamente al alineamiento se realizado una extracción de voz automática.

Esto mismo se podría combinar con métodos de refinamiento local como los empleados en [2], en donde previamente al refinamiento se hace una segmentación basada en el uso de modelos HMM de forma similar a como se hace en el primer método descrito en este proyecto.

GLOSARIO DE TÉRMINOS

- **Alineador:** Sistema que consigue determinar el instante de inicio y fin de cada fonema de un archivo de audio.
- **Fonema:** Unidad teórica básica creada para estudiar una lengua a nivel fónico-fonológico.
- **HMM:** Hidden Markov Model. Modelo oculto de Markov, modelo estadístico en el que se asume que el sistema a modelar es un proceso de Markov de parámetros desconocidos. Nos ayudamos de ellos para realizar reconocimiento de voz.
- **HTK:** Hidden Markov Model Toolkit. Kit de herramientas para la creación y el tratamiento de HMMs.
- **Transcripción:** Texto escrito de todo lo que se dice en un determinado audio.
- **API: (Application Programming Interface),** es una especificación creada con la intención de ser usada como interfaz por componentes software para comunicarse entre ellos.
- **ID:** Identificador de algo, en el contexto de este proyecto significa identificador de un vídeo.
- **WAV:** Formato de audio digital sin compresión de datos.
- **CIS:** (Centro de Investigaciones Sociológicas). Es un centro de investigación que realiza estudios sobre la sociedad española.

BIBLIOGRAFÍA

[1] http://www.inglesdivino.com/

[2] Doroteo Torre Toledano, "*Segmentanción y Etiquetado Fonéticos Automáticos: Un Enfoque Basado en Modelos Ocultos de Markov y Refinamiento Posterior de las Fronteras Fonéticas"*, Tesis doctoral, Año 2001.

[3] Doroteo Torre Toledano, Luis A. Hernández Gómez, *Member IEEE*, Luis Villarubia Grander, "*Automatic Phonectic Segmentation"*, IEEE transactions on speech and audio processing, vol. 11, Nº. 6, november 2003.

[4] Alexander Haubold and John R. Kender, "*Alignment of Speech to Highly Imperfect Text Transcriptions*", Department of Computer Science, Columbia University.

[5] R. Schmidt and R. Neumann, "Automatic Text-to-Speech Aligment: Aspects of Robustification", Institute für deutsche

[6] Reichl W and Ruske G, "*Syllable Segmentation of Continuous Speech With Artificial Neural Networks*", In Proceedings EUROSPEECH 1993, pp. 1771-1774.

[7] Houben CGJ, "*Automatic Labelling of Speech Using an Acoustic-Phonetic Knowledge Base*", In Proceedings EUROSPEECH 1989, V 2, pp 104-107.

[8] Annamaria Mesaros and Tuomas Virtanen, "Automatic Alignment of Music Audio and Lyrics", Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08), Espoo, Finland, September 1-4, 2008.

[9] Kyogu Lee and Markus Cremer, "*Segmentation-Based Lyrics-Audio Alignment Using Dynamic Programming*", In Proc. ISMIR 2008, pp. 395-400.

[10] Hiromasa Fujihara, Masataka Goto, Jun Ogata, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "*Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals*", in Proceedings of the Eighth IEEE International Symposium on Multimedia (ISM'06).

[11] Hugo Meinedo, Alberto Abad, Thomas Pellegrini, Joao Neto, Isabel Trancoso, "*The L2F Broadcast News Speech Recognition System*", in Proc. FALA 2010: VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, pp. 93-96.

[12] A. Ortega, J. Garcia, A. Miguel, and E. Lleida, "*Real-time live broadcast news subtitling system for spanish*", in Proc. Interspeech 2009, Brighton, September 2009.

[13] Ye Wang, Min-Yen Kan, Tin Lay Nwe, Arun Shenoy, and Jun Yin, "*LyricAlly: automatic synchronization of acoustic musical signals and textual lyrics*", in *MULTIMEDIA '04:* Proceedings of the 12th annual ACM international conference on Multimedia, New York, NY, USA, 2004.

[14] http://www.youtube.com/

[15] http://www.google.com/

[16] http://www.lyricstraining.com/

[17] Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu X, Moore G, Odell J, Povey D, Valtchev V, Woodland P, "*The HTK Book"*, Version 3.4 March 2009.

[18] TIMIT Acoustic-Phonetic Continuous Speech Corpus, LDC Catalog Number LDC93S1, Available through the Linguistic Data Consortium, http://www.ldc.upenn.edu.

[19] CMU Pronouncing Dictionary, online, available on ftp://ftp.cs.cmu.edu/project/speech/dict/ (accessed 25 Jun 2012).

A. ALINEAMIENTOS E HISTOGRAMAS

A.1 ALINEAMIENTOS EN NOTICIAS

LOCUTOR 1 (Dana Ward)

Tiempos de referencia manual:

3.04 3.16 3.26 3.36 3.67 3.88 4.58 4.99 5.27 5.53 6.13 6.50 6.63 6.97 7.21 7.43 7.81 7.97 8.56 8.89 9.06 9.80 10.19 10.47 10.63 10.77 10.99 11.44 11.83 12.04 12.22 12.43 12.70 12.81 13.04 13.14 13.70 13.98 14.08 14.22 14.46 15.39 15.69 15.88 16.09 16.27 17.22 17.40 17.69 18.23 19.14 19.79 20.19 20.35 20.68 20.82 21.13 21.30 21.83 22.11 22.22 22.32 22.62 22.72 23.07 23.29 23.60 24.30 24.42 24.64 24.95 25.08 25.18 25.28 25.43 25.55 25.90 26.30 26.60 26.99 27.32 27.67 28.98 29.25 29.34 29.54 29.63 29.91 30.06 30.48 30.58 31.28 31.54 31.67 32.18 33.12 33.27 33.37 33.73 34.15 34.51 34.86 35.06 35.24 35.70 35.80 36.05 36.23 37.04 37.18 37.43 37.64 37.74 38.05 38.27 38.47 38.63 38.92 39.27 39.39 39.73 39.97 40.75 41.11 41.37 42.08 42.38 42.48 42.78 43.00 43.46 43.56 43.66 44.23 44.80 45.26 45.36 45.98 46.09 46.23 46.98 47.86 48.39 48.60 48.95 49.30 49.95 50.29 50.49 50.73 51.12 51.29 51.54 51.82 52.14 52.59 52.80 53.08 53.28 53.50 53.66 54.02 54.26 54.36 54.58 55.24 55.55 55.83 56.15 56.41 57.30 57.48 57.67 57.96 58.06 58.26 58.49 59.02 59.21 59.36 59.80 60.18 60.34 60.66 60.83 61.26 61.38 62.38

Alineamiento sin audios adicionales:

- Audio 1
  - Duración de audio: 00:01:05.83
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 188

Tiempos de alineamiento:

2.81 3.12 3.48 3.55 4.04 4.10 4.72 5.09 5.26 5.68 6.05 7.04 7.18 7.46 7.88 8.42 8.71 8.79 9.48 9.72 9.78 9.89 10.10 10.46 10.62 10.71 11.12 11.61 11.70 12.21 12.38 12.50 12.66 12.74 13.33 13.64 13.73 13.92 14.16 14.54 14.62 15.74 15.90 16.00 16.27 16.38 17.37 17.48 17.62 18.09 18.53 19.63 20.04 20.32 20.53 20.90 21.15 21.40 21.47 21.81 22.06 22.18 22.49 22.84 23.17 23.88 24.06 24.79 24.89 25.16 25.34 25.47 25.59 25.83 25.93 26.10 26.83 27.07 27.46 27.52 27.68 28.15 29.07 29.23 29.55 29.76 30.07 30.40 31.17 31.47 31.72 32.74 32.80 33.18 33.63 34.01 34.19 34.39 34.96 35.36 35.53 35.85 35.93 36.33 36.88 36.96 37.24 37.62 37.86 38.03 38.09 38.41 38.90 39.42 39.99 40.23 40.32 40.58 40.87 41.42 41.53 41.88 42.19 42.60 42.73 42.90 43.01 43.36 43.59 43.71 44.43 44.49 44.57 45.23 45.50 46.04 46.30 46.48 46.57 46.73 47.46 48.06 48.48 48.61 48.87 49.13 49.19 49.43 49.62 49.86 50.20 50.33 50.39 51.01 51.55 51.97 52.08 52.30 52.42 52.69 53.16 53.30 53.42 53.67 54.27 54.78 55.22 55.45 55.99 56.20 56.66 56.78 57.15 57.49 57.91 58.26 58.46 59.08 59.28 59.41 59.88 60.25 60.36 60.51 60.76 61.20 61.75 62.44

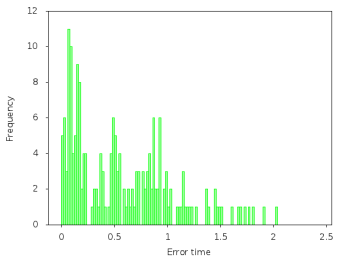
Alineamiento con 3 audios adicionales:

- Audio 1
  - Duración de audio: 00:01:05.83
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 188
- Audio 2
  - Duración de audio: 00:01:24.20
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 226
- Audio 3
  - Duración de audio: 00:01:28.51
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 253
- Audio 4
  - Duración de audio: 00:01:24.54
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 228

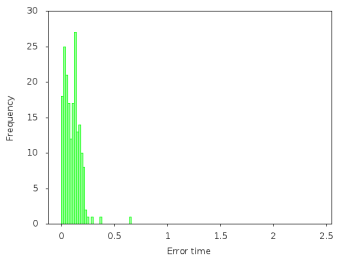
Tiempos de alineamiento:

2.97 3.16 3.31 3.34 3.81 3.98 4.71 5.02 5.25 5.67 6.26 6.61 6.69 7.07 7.16 7.57 8.01 8.18 8.63 9.18 9.24 9.93 10.29 10.45 10.63 10.79 11.11 11.61 11.78 12.19 12.36 12.62 12.71 12.83 13.03 13.20 13.74 13.96 14.18 14.43 14.60 15.42 15.68 15.93 16.23 16.38 17.35 17.60 17.81 18.24 19.00 19.96 20.28 20.52 20.66 20.91 21.19 21.38 21.94 22.06 22.20 22.43 22.62 22.81 23.04 23.45 23.69 24.37 24.54 24.77 24.90 25.03 25.16 25.23 25.51 25.68 26.08 26.37 26.80 27.03 27.35 27.88 29.10 29.24 29.47 29.56 29.76 29.92 30.07 30.53 30.75 31.35 31.55 31.64 32.35 33.10 33.33 33.42 33.81 34.34 34.57 34.83 34.99 35.49 35.83 35.91 35.99 36.33 37.16 37.35 37.56 37.65 37.78 38.21 38.40 38.55 38.78 38.91 39.25 39.57 39.77 40.10 40.85 41.47 42.02 42.19 42.38 42.59 42.75 42.97 43.58 43.75 43.82 44.42 44.95 45.23 45.29 45.94 46.14 46.35 47.20 48.05 48.53 48.73 49.12 49.46 50.14 50.31 50.50 50.80 51.24 51.42 51.53 51.93 52.30 52.72 52.96 53.14 53.33 53.51 53.79 54.11 54.23 54.47 54.78 55.26 55.59 55.93 56.29 56.62 57.52 57.64 57.76 58.04 58.25 58.38 58.47 59.06 59.27 59.49 59.95 60.27 60.33 60.62 60.88 61.36 61.50 62.44

Histogramas:



Sin audios adicionales



Con audios adicionales

Tabla A. 1: Alineamientos e histogramas de las pruebas con el locutor 1. Los tiempos se corresponden con el inicio de cada palabra de los audios.

LOCUTOR 2 (Josselyn Davis)

Tiempos de referencia manual:

3.00 3.26 3.78 4.04 4.36 4.63 5.11 5.77 6.01 6.13 6.25 6.50 6.80 7.09 8.00 8.29 8.39 8.57 9.06 9.58 9.85 10.71 11.14 11.67 12.12 12.58 12.80 13.21 14.29 14.54 14.64 14.90 15.63 15.83 16.36 16.74 17.48 17.87 18.10 18.98 19.26 19.43 19.94 20.09 20.19 20.46 20.69 20.99 22.06 22.37 22.69 22.88 23.10 23.29 23.50 23.76 24.21 24.54 24.74 25.22 25.61 25.74 26.54 26.64 26.76 27.47 28.05 28.29 28.39 28.68 29.37 29.71 31.12 31.80 32.02 32.24 32.42 32.68 32.87 33.47 33.73 34.59 35.01 35.11 35.23 35.39 35.92 36.03 36.45 36.71 37.00 37.10 37.25 37.78 38.16 38.49 38.71 39.08 40.19 40.43 40.58 40.91 41.09 41.35 41.56 41.90 42.19 43.05 43.30 44.05 44.48 44.78 44.90 45.04 45.26 45.41 45.54 45.82 45.95 46.39 46.66 47.11 47.95 48.13 48.33 48.60 49.25 49.57 49.67 49.77 50.03 50.21 51.03 51.13 51.63 51.73 51.97 52.41 52.76 53.03 53.21 53.70 53.80 54.06 54.28 54.49 54.59 55.87 56.26 56.62 57.15 57.43 57.76 58.08 58.42 58.74 59.50 59.60 59.70 59.84 60.11 60.22 60.52 60.75 61.17 61.79 62.07 62.18 62.42 63.64 64.05 64.15 64.45 64.99 65.09 65.23 65.60 66.41 66.75 66.84 67.10 67.24 67.50 67.59 67.82 68.18 68.76 69.05 69.52 69.65 70.16 70.39 70.58 71.14 71.99 72.22 72.57 72.71 73.28 73.39 73.48 73.59 73.83 74.02 74.17 74.49 74.89 75.00 75.32 75.61 75.89 76.08 77.11 77.44 77.62 77.82 77.95 78.25 78.50 78.80 79.07 79.51 79.88 80.29 80.96 81.60 82.00 82.53 82.75 82.86 83.11 83.23 83.51 83.82 83.97 84.87 84.97 85.17 85.47 85.56 85.88 86.30 86.53 86.99 87.28 88.67 88.84 89.11 89.21 89.43 89.55 89.89 90.14 90.48 91.02 91.40 91.71 92.43 92.87 93.28 93.85 94.42 94.80 95.17 95.39 95.49 95.98 96.22 96.33 96.48 96.79 96.98 97.17 97.58 98.46 98.69 98.79 98.96 99.06 99.17 99.59 100.22 100.31 100.57 100.79 101.63 101.91 102.17 102.59 102.92 103.13 103.34 103.71 103.94 104.03 104.13 104.34 104.53 104.76 105.07 105.44 105.73

Alineamiento sin audios adicionales:

- Audio 1
  - Duración de audio: 00:01:22.36
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 223

Tiempos de alineamiento:

3.13 3.44 4.00 4.15 4.22 4.78 5.27 5.90 6.02 6.26 6.44 6.52 6.88 7.15 8.07 8.26 8.48 8.64 9.15 9.68 9.85 10.74 11.01 11.81 12.21 12.78 12.93 13.40 14.21 14.67 14.77 14.89 15.63 15.87 16.46 16.86 17.64 17.74 18.10 19.16 19.37 19.49 19.90 19.96 20.29 20.48 20.83 20.89 22.01 22.49 22.65 23.00 23.22 23.43 23.46 23.87 24.32 24.54 24.83 25.38 25.68 25.75 26.54 26.74 26.84 27.64 28.25 28.31 28.55 28.82 29.44 29.82 30.76 31.92 32.10 32.43 32.51 32.55 32.97 33.64 33.74 34.51 35.01 35.13 35.37 35.56 35.96 36.17 36.55 36.64 37.12 37.31 37.55 37.89 38.34 38.61 38.79 39.13 40.04 40.35 40.71 41.06 41.22 41.54 41.65 42.00 42.33 42.96 43.30 43.92 44.22 44.53 44.71 44.85 45.10 45.56 45.69 45.72 46.07 46.50 46.85 47.28 47.86 48.24 48.42 48.50 49.32 49.73 49.80 49.90 50.12 50.26 51.05 51.15 51.76 51.85 51.91 52.61 52.89 53.06 53.38 53.74 53.85 53.97 54.11 54.39 54.63 55.54 56.06 56.68 57.22 57.45 57.80 58.15 58.58 58.78 59.56 59.71 59.79 59.94 60.24 60.34 60.64 60.79 61.37 61.79 62.25 62.95 63.37 63.88 64.04 64.19 64.47 65.04 65.10 65.17 65.65 66.16 66.59 66.94 67.08 67.35 67.44 67.56 67.92 68.25 68.89 69.17 69.65 69.78 69.87 70.30 70.77 71.22 72.15 72.21 72.52 72.70 73.29 73.39 73.54 73.60 73.71 73.98 74.15 74.62 74.79 75.19 75.31 75.75 76.02 76.20 77.11 77.46 77.81 77.94 78.10 78.23 78.41 78.97 79.63 79.98 80.50 81.02 81.22 81.56 82.14 82.62 82.70 83.00 83.15 83.38 83.66 83.88 84.08 84.45 84.69 84.92 85.39 85.62 85.93 86.39 86.67 86.99 87.40 88.30 88.80 89.08 89.24 89.58 89.61 89.87 90.30 90.61 91.05 91.57 91.81 92.55 93.00 93.43 94.06 94.53 94.87 95.35 95.46 95.51 96.03 96.38 96.54 96.64 96.89 97.00 97.14 97.67 98.35 98.67 98.88 98.91 99.15 99.24 99.63 100.17 100.40 100.47 100.71 100.95 101.84 102.27 102.54 102.99 103.08 103.47 103.65 103.86 104.05 104.19 104.27 104.58 105.01 105.34 105.53 105.73

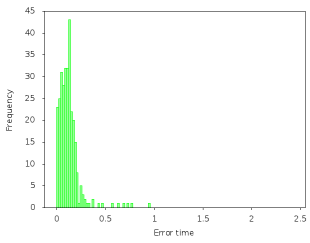
Alineamiento con 3 audios adicionales:

- Audio 1
  - Duración de audio: 00:01:22.36
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 223
- Audio 2
  - Duración de audio: 00:01:24.20
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 226
- Audio 3
  - Duración de audio: 00:01:28.51
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 253
- Audio 4
  - Duración de audio: 00:01:24.54
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 228

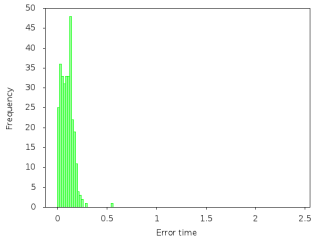
Tiempos de alineamiento:

3.13 3.44 3.95 4.16 4.23 4.79 5.27 5.90 6.04 6.26 6.34 6.44 6.88 7.11 8.06 8.27 8.47 8.64 9.23 9.68 9.90 10.83 11.12 11.81 12.17 12.72 12.93 13.43 14.19 14.39 14.73 14.89 15.59 15.87 16.52 16.87 17.63 17.86 18.12 19.21 19.38 19.49 19.90 19.97 20.28 20.43 20.83 20.91 22.11 22.49 22.65 22.98 23.20 23.44 23.47 23.87 24.32 24.54 24.83 25.38 25.64 25.71 26.53 26.74 26.85 27.62 28.18 28.27 28.55 28.84 29.44 29.83 31.23 31.92 32.15 32.42 32.51 32.55 32.99 33.60 33.74 34.49 35.13 35.19 35.38 35.51 36.12 36.18 36.55 36.64 37.10 37.25 37.31 37.89 38.34 38.61 38.82 39.13 40.21 40.34 40.70 41.06 41.22 41.54 41.65 42.07 42.27 42.88 43.28 44.18 44.51 44.72 44.85 45.07 45.25 45.55 45.70 45.73 46.04 46.52 46.86 47.28 48.08 48.19 48.44 48.51 49.34 49.64 49.71 49.79 50.08 50.26 50.48 51.01 51.67 51.91 51.97 52.59 52.89 53.06 53.34 53.74 53.86 53.98 54.40 54.49 54.55 55.73 56.25 56.68 57.23 57.43 57.81 58.15 58.57 58.72 59.56 59.69 59.79 59.88 60.24 60.39 60.64 60.79 61.35 61.91 62.06 62.23 62.53 63.80 64.00 64.19 64.43 65.05 65.11 65.20 65.65 66.12 66.56 66.95 67.04 67.36 67.42 67.53 67.92 68.26 68.82 69.16 69.64 69.87 70.26 70.41 70.77 71.23 72.12 72.27 72.52 72.71 73.28 73.38 73.55 73.70 73.95 74.15 74.18 74.62 75.03 75.18 75.34 75.55 75.73 76.10 77.25 77.46 77.79 77.92 78.08 78.23 78.42 78.90 79.12 79.61 79.95 80.43 81.16 81.56 82.15 82.60 82.70 83.00 83.14 83.20 83.66 83.81 83.93 84.85 85.01 85.20 85.41 85.67 85.91 86.39 86.61 86.99 87.40 88.76 88.95 89.12 89.28 89.46 89.56 89.84 90.31 90.60 91.03 91.57 91.83 92.53 92.98 93.43 94.05 94.53 94.75 95.35 95.47 95.52 96.03 96.35 96.47 96.60 96.89 96.99 97.14 97.68 98.34 98.56 98.88 98.93 99.15 99.23 99.62 100.15 100.40 100.47 100.78 101.70 101.90 102.28 102.75 102.99 103.08 103.46 103.77 103.94 104.05 104.20 104.28 104.62 105.01 105.31 105.54 105.73

Histogramas:



Sin audios adicionales



Con audios adicionales

Tabla A. 2: Alineamientos e histogramas de las pruebas con el locutor 2. Los tiempos se corresponden con el inicio de cada palabra de los audios

LOCUTOR 3 (Chris Matthews)

Tiempos de referencia manual:

3.32 3.63 3.74 4.03 4.17 4.40 4.56 4.74 5.15 5.42 5.55 5.79 6.04 6.27 6.70 7.04 7.23 7.44 7.74 8.04 8.66 9.14 9.44 9.67 9.90 10.01 10.30 10.40 11.22 11.40 11.51 11.63 11.82 11.92 12.21 12.48 12.74 12.91 13.05 13.69 13.82 13.99 14.15 14.53 14.63 14.91 15.55 15.88 16.17 16.72 16.82 17.07 17.46 17.82 17.95 18.14 18.57 19.17 19.45 19.66 19.90 20.11 20.21 20.80 21.03 21.16 21.27 21.38 21.92 22.15 22.39 22.53 23.28 23.68 24.06 24.47 24.59 25.26 25.36 25.50 25.83 25.98 26.35 26.71 26.92 27.14 27.25 27.46 27.56 28.11 28.48 29.36 29.46 29.89 30.13 30.33 30.52 30.81 30.94 31.05 31.15 31.35 31.94 32.77 32.87 33.13 33.53 33.70 33.82 34.03 34.13 34.61 34.71 35.50 36.22 36.90 37.12 37.33 37.73 37.83 38.05 38.34 38.68 39.16 39.27 39.96 40.08 40.18 40.42 40.52 40.85 40.98 41.27 41.43 41.87 42.22 42.41 43.06 43.28 43.50 43.75 43.95 44.05 44.32 44.59 44.89 45.26 45.41 45.71 45.83 46.54 46.63 46.77 47.08 47.40 47.58 48.15 48.39 48.79 49.08 49.35 49.60 49.70 49.90 50.08 50.18 50.40 50.62 51.01 51.29 51.97 52.43 52.53 52.73 52.91 53.50 53.85 54.26 54.44 54.59 55.16 55.26 55.36 55.72 56.26 56.52 56.72 56.94 57.07 57.38 57.51 57.63 57.92 58.41 58.59 59.12 59.27 60.02 60.19 60.72 60.83 61.30 61.51 61.94 62.07 62.17 62.36 62.58 62.77 63.05 63.56 63.89 64.48 64.70 64.89 65.14 65.32 65.55 65.76 65.86 66.00 66.41 67.16 67.26 67.38 67.48 67.65 67.75 68.03 68.94 69.19 69.59 69.81 70.10 70.27 70.64 70.87 71.12 71.22 71.55 72.07 72.25 72.77 73.02 73.20 73.58 74.31 74.45 74.55 74.83 75.21 75.32 75.47 76.07 76.41 76.52 76.90 77.19 77.44 78.46 78.56 78.66 78.77 78.94 79.08 79.19 79.28 79.68 79.83 79.98 80.18 80.50 80.69 80.91 81.40 81.54 82.13 82.48 83.03 83.13 83.23 83.33 83.43 83.65 83.92 84.20 84.31 84.58 85.05 85.42 85.55 85.72 86.34 86.66 86.98 87.08 87.55 88.44 88.64 88.80 89.09 89.27 89.49 89.81 89.97 90.11 90.90 91.05 91.25 91.39 92.22 92.34 92.89 93.14 93.58 93.86 94.36 95.04 95.82

Alineamiento sin audios adicionales:

- Audio 1
  - Duración de audio: 00:01:33.19
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 319

Tiempos de alineamiento:

3.34 3.57 3.69 4.17 4.45 4.64 5.05 5.25 5.40 5.54 5.72 5.84 5.92 6.42 6.86 7.18 7.50 7.77 7.89 8.00 8.78 9.54 9.79 9.83 10.05 10.12 10.61 10.69 11.01 11.10 11.35 11.42 11.69 11.83 12.09 12.59 12.67 13.14 13.29 13.72 13.95 14.15 14.47 14.73 15.33 15.74 16.07 16.85 16.95 17.19 17.34 17.37 17.89 17.98 18.13 18.60 19.20 19.92 20.10 20.32 20.46 20.98 21.68 22.04 22.14 22.18 22.54 22.67 22.99 23.37 23.50 23.80 24.23 24.54 24.79 25.06 25.55 25.78 25.85 25.88 26.01 26.34 27.19 28.19 28.37 28.52 28.55 28.89 28.95 29.85 29.93 30.23 30.47 30.98 31.13 31.26 31.42 31.82 32.03 32.17 32.70 32.85 33.54 33.72 33.81 34.16 34.74 34.83 35.22 35.31 35.60 35.93 36.14 36.90 37.35 37.96 38.03 38.12 38.41 38.63 39.08 39.22 39.55 40.22 40.62 40.80 41.03 41.19 41.32 41.62 42.41 42.68 42.81 42.92 43.80 44.36 44.70 45.50 45.64 45.88 46.71 47.07 47.29 47.37 47.59 47.98 48.28 48.67 49.01 49.10 49.40 49.52 49.57 49.91 50.49 50.77 50.85 51.35 51.63 51.69 51.75 51.95 51.98 52.07 52.57 52.76 52.87 53.04 53.41 53.80 54.60 54.86 55.00 55.38 55.64 56.10 56.70 56.83 56.97 57.03 57.27 57.40 57.59 57.73 58.01 58.28 58.39 58.68 59.08 59.20 59.37 59.83 60.13 60.40 60.61 60.67 61.16 61.41 61.85 62.46 62.61 63.25 63.39 63.78 64.39 64.46 64.77 64.89 65.22 65.52 65.74 65.89 66.86 67.05 67.27 67.36 67.58 67.65 68.12 68.19 68.31 68.75 68.84 68.94 69.13 69.22 69.41 69.65 69.84 70.31 70.68 71.07 71.22 71.42 71.64 71.92 72.10 72.17 72.33 72.69 73.05 73.14 73.46 73.55 73.63 73.87 74.25 74.33 75.12 75.31 75.48 75.97 76.03 76.09 76.29 76.92 77.41 77.54 77.70 78.57 78.76 78.94 79.13 79.69 80.07 80.19 80.25 80.34 80.56 80.79 81.00 81.09 81.30 81.91 82.49 82.52 83.11 83.56 83.73 83.76 84.08 84.16 84.40 84.48 84.99 85.20 85.34 85.56 85.73 86.50 86.64 86.82 87.37 87.51 87.99 88.14 88.65 89.22 89.46 89.74 89.93 90.31 90.42 90.72 91.07 91.35 92.24 92.32 92.47 92.63 93.04 93.18 94.08 94.34 94.46 94.54 95.06 95.16 96.08

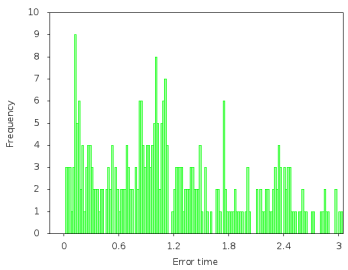
Alineamiento con 2 audios adicionales:

- Audio 1
  - Duración de audio: 00:01:33.19
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 319
- Audio 2
  - Duración de audio: 00:01:50.17
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 400
- Audio 3
  - Duración de audio: 00:02:28.77
  - Número de palabras: 465

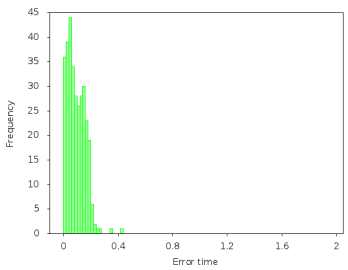
Tiempos de alineamiento:

3.50 3.61 3.68 4.09 4.31 4.44 4.70 4.86 5.26 5.38 5.66 5.72 5.95 6.43 6.85 6.93 7.19 7.60 7.93 8.01 8.77 9.35 9.58 9.67 10.05 10.14 10.42 10.56 11.31 11.42 11.52 11.56 11.80 12.06 12.15 12.57 12.66 13.12 13.22 13.71 13.81 14.01 14.28 14.56 14.76 15.06 15.71 16.06 16.13 16.82 17.00 17.04 17.57 17.82 17.93 18.05 18.52 19.24 19.54 19.70 20.00 20.21 20.36 20.91 21.04 21.07 21.41 21.52 21.91 22.30 22.38 22.70 23.48 23.79 24.23 24.61 24.75 25.33 25.48 25.53 25.77 25.89 26.39 26.86 27.02 27.08 27.13 27.55 27.63 28.18 28.56 29.36 29.54 29.86 30.26 30.51 30.65 30.77 30.92 31.01 31.41 31.50 32.07 32.71 32.87 33.38 33.56 33.74 33.94 34.06 34.30 34.77 34.85 35.68 36.20 36.96 37.10 37.37 37.87 37.96 38.21 38.33 38.67 39.22 39.36 39.99 40.10 40.34 40.47 40.69 40.98 41.17 41.29 41.46 42.03 42.31 42.42 43.12 43.35 43.50 43.80 44.05 44.25 44.37 44.67 44.88 45.24 45.56 45.75 45.87 46.58 46.70 46.73 47.03 47.45 47.77 48.26 48.56 48.95 49.19 49.41 49.55 49.58 49.85 50.21 50.27 50.33 50.71 51.05 51.38 52.04 52.39 52.53 52.88 52.96 53.51 53.79 54.27 54.56 54.64 55.27 55.39 55.47 55.90 56.23 56.59 56.77 57.06 57.24 57.39 57.48 57.78 57.93 58.37 58.64 59.06 59.47 60.06 60.26 60.71 61.00 61.49 61.56 61.96 62.09 62.16 62.58 62.64 62.88 63.05 63.64 63.93 64.51 64.83 65.04 65.17 65.42 65.51 65.80 65.89 66.18 66.60 67.27 67.36 67.49 67.55 67.67 67.90 68.02 68.89 69.38 69.74 69.83 70.19 70.42 70.65 70.91 71.10 71.25 71.59 72.08 72.28 72.86 73.05 73.32 73.74 74.10 74.32 74.74 75.00 75.31 75.44 75.61 76.18 76.41 76.60 77.09 77.34 77.45 78.11 78.14 78.47 78.58 79.07 79.21 79.37 79.43 79.68 79.78 80.01 80.33 80.42 80.72 80.97 81.41 81.46 82.30 82.52 82.80 83.02 83.23 83.38 83.54 83.60 84.00 84.26 84.43 84.58 85.15 85.58 85.68 85.91 86.46 86.58 87.02 87.08 87.64 88.44 88.62 88.95 89.01 89.33 89.43 89.88 89.96 90.26 90.94 91.03 91.22 91.30 92.28 92.38 93.05 93.20 93.72 93.85 94.48 94.98 95.82

Histogramas:



Sin audios adicionales



Con audios adicionales

Tabla A. 3: Alineamientos e histogramas de las pruebas con el locutor 3. Los tiempos se corresponden con el inicio de cada palabra de los audios

LOCUTOR 4 (Al Sharpton)

Tiempos de referencia manual:

1.64 1.94 2.33 2.83 3.07 3.55 3.66 3.83 4.79 5.09 5.20 5.30 5.54 5.78 6.18 6.42 6.62 6.75 7.04 7.49 7.66 8.31 8.57 8.92 9.83 9.97 10.23 10.90 11.13 11.34 11.63 12.10 13.55 13.66 13.85 14.06 14.16 14.58 14.86 15.35 16.07 16.61 17.24 17.77 18.66 19.10 19.37 20.35 20.55 20.79 21.46 21.65 21.75 21.93 22.13 22.92 23.26 23.75 24.07 24.21 24.81 24.96 25.24 25.55 25.66 26.27 26.50 26.91 27.13 27.96 28.23 28.43 28.67 29.01 29.47 29.87 30.52 31.09 31.41 32.55 32.73 32.83 32.93 33.40 33.65 33.87 34.64 34.84 35.05 35.18 36.37 36.67 36.96 37.41 38.51 38.61 38.90 39.05 39.23 39.33 39.54 39.64 39.97 40.88 41.15 41.51 42.75 42.90 43.42 43.61 43.87 44.01 44.65 45.05 45.26 46.21 47.07 47.23 47.49 47.82 47.98 48.30 48.70 49.86 50.10 50.32 50.57 50.85 51.24 51.47 51.78 52.37 52.77 53.29 53.53 54.02 54.80 55.01 55.70 55.97 56.38 56.75 57.08 58.34 58.49 58.82 58.95 60.18 60.28 60.66 61.06 61.43 61.75 62.11 63.07 63.38 63.50 63.66 64.11 64.75 65.01 65.51 66.27 67.61 68.31 68.66 68.96 69.21 70.11 70.44 70.83 71.00 71.52 72.54 73.50 73.83 73.92 74.14 74.24 74.41 74.79 74.93 75.52 75.80 76.04 76.61 77.49 77.92 78.02 78.13 78.67 79.20 79.60 79.85 80.01 80.28 80.64 81.31 81.42 81.73 82.01 82.22 83.43 83.81 84.47 84.79 85.09 85.52 85.86 86.72 87.06 87.43 87.86 88.35 88.49 88.76 89.49 89.91 90.49 90.98 91.26 92.19 92.42 92.93 94.06 94.41 94.56 94.66 94.76 95.03 95.24 95.45 95.55 96.06 96.24 96.77 97.00 97.35 97.55 97.65 98.36 98.66 98.75 99.05 99.14 99.75 101.08 101.18 101.34 101.61 102.53 102.88 103.16 103.32 104.16 104.29 104.58 105.48 105.81 106.10 106.35 106.95 107.11 107.28 107.61 107.88 108.15 108.44 108.55 108.92 109.16 110.17 110.38 110.58 110.76 111.55 111.70 112.05 112.96 113.11 113.29 113.56 113.96 114.24 114.34 114.74 115.06 115.16 116.53 116.65 116.95 117.12 117.66 118.55 118.74 119.10 119.23 119.50 120.47 120.85 121.05 121.16 121.53 121.96 122.83 123.13 123.30 123.51 124.35 124.53 124.91 125.17 126.12 126.41 126.52 127.44 127.56 127.82

Alineamiento sin audios adicionales:

- Audio 1
  - Duración de audio: 00:02:07.21
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 318

Tiempos de alineamiento:

1.69 2.03 2.45 3.00 3.13 4.40 4.50 4.53 5.18 5.33 5.40 6.26 6.34 6.68 6.95 7.14 7.36 7.78 8.04 8.28 8.84 9.87 9.93 10.13 10.89 11.04 11.10 11.37 11.85 12.15 12.33 12.75 13.96 14.31 14.80 15.02 15.21 16.13 16.23 16.66 17.01 17.22 18.20 18.94 19.17 19.83 20.50 20.76 21.34 22.03 22.51 23.51 23.87 24.45 25.42 25.77 26.04 26.17 26.33 26.43 26.61 26.81 26.95 27.09 27.31 27.63 28.67 29.02 29.22 29.44 29.52 29.66 30.24 30.70 31.02 31.75 32.55 33.02 33.81 34.45 34.48 34.96 35.02 35.31 35.50 35.80 36.60 36.83 37.24 37.60 37.85 37.91 39.14 39.33 39.48 39.54 39.79 39.98 40.23 40.44 40.56 41.25 42.05 43.40 43.52 44.30 45.22 45.47 47.21 47.30 47.39 47.42 47.51 47.81 48.00 48.41 48.54 48.67 48.81 50.05 50.14 50.54 50.95 51.34 51.40 51.83 52.45 52.70 53.77 54.19 54.74 55.17 55.99 56.25 56.71 58.37 59.02 59.23 60.32 60.78 61.55 61.85 62.18 63.57 63.77 64.01 64.15 64.95 65.05 65.13 65.31 65.40 65.83 66.22 66.68 66.96 67.83 68.06 68.70 70.04 70.27 70.75 71.07 71.65 72.61 73.10 73.16 74.01 74.13 74.20 74.38 74.53 75.61 75.92 76.08 76.14 76.53 76.65 76.79 77.03 77.22 77.44 77.73 78.01 78.80 80.92 82.01 82.24 82.64 83.48 83.83 84.46 84.82 84.88 85.19 85.27 85.49 85.82 85.88 86.07 86.17 86.96 88.27 88.45 88.56 88.71 88.90 89.66 89.72 89.81 89.87 90.15 90.56 91.29 91.35 91.43 91.77 92.74 94.08 94.36 94.80 94.96 95.36 95.89 96.21 96.48 96.51 97.07 97.28 97.46 97.57 98.06 98.09 98.53 98.64 98.97 99.09 99.38 99.47 99.53 99.86 100.00 100.24 100.30 101.01 101.23 101.42 101.61 101.70 101.91 102.57 102.70 102.86 103.14 103.83 104.25 104.78 105.64 105.90 106.07 106.25 106.59 106.73 107.63 107.85 108.42 108.68 108.90 109.41 109.58 110.08 110.31 110.95 111.14 111.56 111.81 112.07 113.14 113.29 113.46 113.69 114.10 114.24 114.30 114.57 114.87 115.45 115.81 116.85 117.06 117.20 117.26 117.86 118.36 118.72 119.18 119.26 119.73 120.14 121.04 121.16 121.46 121.87 122.18 122.60 122.66 122.75 123.51 123.99 124.18 124.95 125.10 126.31 126.64 126.93 127.64 127.73 127.79

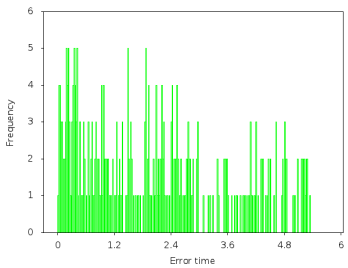
Alineamiento con 1 audio adicional:

- Audio 1
  - Duración de audio: 00:02:07.21
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 318
- Audio 2
  - Duración de audio: 00:02:25.31
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 330

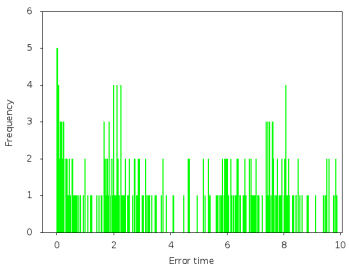
Tiempos de alineamiento:

3.50 3.61 3.68 4.09 4.31 4.44 4.70 4.86 5.26 5.38 5.66 5.72 5.95 6.43 6.85 6.93 7.19 7.60 7.93 8.01 8.77 9.35 9.58 9.67 10.05 10.14 10.42 10.56 11.31 11.42 11.52 11.56 11.80 12.06 12.15 12.57 12.66 13.12 13.22 13.71 13.81 14.01 14.28 14.56 14.76 15.06 15.71 16.06 16.13 16.82 17.00 17.04 17.57 17.82 17.93 18.05 18.52 19.24 19.54 19.70 20.00 20.21 20.36 20.91 21.04 21.07 21.41 21.52 21.91 22.30 22.38 22.70 23.48 23.79 24.23 24.61 24.75 25.33 25.48 25.53 25.77 25.89 26.39 26.86 27.02 27.08 27.13 27.55 27.63 28.18 28.56 29.36 29.54 29.86 30.26 30.51 30.65 30.77 30.92 31.01 31.41 31.50 32.07 32.71 32.87 33.38 33.56 33.74 33.94 34.06 34.30 34.77 34.85 35.68 36.20 36.96 37.10 37.37 37.87 37.96 38.21 38.33 38.67 39.22 39.36 39.99 40.10 40.34 40.47 40.69 40.98 41.17 41.29 41.46 42.03 42.31 42.42 43.12 43.35 43.50 43.80 44.05 44.25 44.37 44.67 44.88 45.24 45.56 45.75 45.87 46.58 46.70 46.73 47.03 47.45 47.77 48.26 48.56 48.95 49.19 49.41 49.55 49.58 49.85 50.21 50.27 50.33 50.71 51.05 51.38 52.04 52.39 52.53 52.88 52.96 53.51 53.79 54.27 54.56 54.64 55.27 55.39 55.47 55.90 56.23 56.59 56.77 57.06 57.24 57.39 57.48 57.78 57.93 58.37 58.64 59.06 59.47 60.06 60.26 60.71 61.00 61.49 61.56 61.96 62.09 62.16 62.58 62.64 62.88 63.05 63.64 63.93 64.51 64.83 65.04 65.17 65.42 65.51 65.80 65.89 66.18 66.60 67.27 67.36 67.49 67.55 67.67 67.90 68.02 68.89 69.38 69.74 69.83 70.19 70.42 70.65 70.91 71.10 71.25 71.59 72.08 72.28 72.86 73.05 73.32 73.74 74.10 74.32 74.74 75.00 75.31 75.44 75.61 76.18 76.41 76.60 77.09 77.34 77.45 78.11 78.14 78.47 78.58 79.07 79.21 79.37 79.43 79.68 79.78 80.01 80.33 80.42 80.72 80.97 81.41 81.46 82.30 82.52 82.80 83.02 83.23 83.38 83.54 83.60 84.00 84.26 84.43 84.58 85.15 85.58 85.68 85.91 86.46 86.58 87.02 87.08 87.64 88.44 88.62 88.95 89.01 89.33 89.43 89.88 89.96 90.26 90.94 91.03 91.22 91.30 92.28 92.38 93.05 93.20 93.72 93.85 94.48 94.98 95.82

Histogramas:



Sin audios adicionales



Con audios adicionales

Tabla A. 4: Alineamientos e histogramas de las pruebas con el locutor 4. Los tiempos se corresponden con el inicio de cada palabra de los audios

A.2 ALINEAMIENTOS EN CANCIONES (ACAPELLA)

CANCIÓN 1 (Lose Yourself – Eminem)

Tiempos de referencia manual:

33.04 35.74 35.97 36.18 38.03 38.27 40.08 40.22 40.45 42.73 42.86 43.05 43.51 43.70 43.98 45.55 45.84 47.94 48.10 48.25 48.53 49.88 50.06 50.25 50.39 50.53 51.36 52.57 52.85 53.03 53.31 53.87 54.24 54.47 54.85 55.08 55.40 55.59 55.84 56.08 56.22 56.54 57.31 57.63 58.20 58.38 59.02 59.21 59.35 59.45 59.70 59.93 60.21 60.53 60.72 61.00 61.23 61.51 62.25 62.44 62.53 62.95 63.20 63.90 64.04 64.23 64.50 64.92 65.03 65.30 65.81 66.08 66.27 66.64 66.87 67.25 67.48 67.80 68.03 68.17 68.45 68.78 69.01 69.47 69.66 70.12 70.73 71.33 71.80 72.21 72.45 72.72 72.91 73.40 73.70 74.21 74.67 75.14 75.51 75.84 76.07 76.81 77.14 77.32 77.60 78.11 78.53 78.72 78.95 79.37 79.55 79.88 80.11 80.34 80.67 80.85 80.99 81.27 81.46 81.70 82.00 82.34 82.76 82.89 83.17 83.41 83.59 83.78 84.20 84.33 84.57 84.87 85.00 85.22 85.54 85.73 86.05 86.38 86.52 86.84 87.13 87.40 87.57 87.72 88.00 88.38 88.51 88.75 89.21 89.35 89.63 89.73 89.85 90.14 90.37 90.60 90.74 91.25 91.53 91.91 92.14 92.35 92.72 92.86 93.00 93.34 93.58 93.95 94.13 94.37 95.02 95.20 95.39 95.85 96.04 96.36 96.55 96.71 96.85 97.08 97.20 97.30 97.41 97.78 98.03 98.73 98.85 99.03 99.56 99.66 100.22 100.36 100.73 100.91 101.05 101.47 101.75 101.98 102.17 102.49 102.77 103.00 103.24 103.38 103.56 103.98 104.21 104.58 104.70 104.84 105.21 105.44 105.81 105.91 106.77 107.14 107.37 107.56 107.69 108.48 108.62 108.99 109.27 109.83 110.00 110.18 110.71 110.81 111.37 111.51 111.88 112.06 112.20 112.62 112.90 113.13 113.32 113.64 113.92 114.15 114.39 114.53 114.71 115.13 115.36 115.73 115.85 115.99 116.36 116.59 116.96 117.06 117.92 118.29 118.52 118.71 118.84 119.54 119.77 120.05 120.19 120.51 121.44 121.63 121.77 122.00 122.28 122.37 123.16 123.34 123.76 123.86 124.13 124.32 124.50 125.20 125.43 125.69 126.34 126.48 126.71 126.94 127.31 127.55 127.65 127.92 128.48 128.71 129.26 129.54 129.82 130.38 130.48 131.43 131.70 131.89 132.07 132.75 132.85 133.31 133.54 134.24 134.61 134.98 135.34 135.63 136.00 136.14 136.37 136.84 137.00 137.35 137.53 137.76 138.18 138.41 138.74 138.85 139.18 139.50 139.78 140.02 140.15 140.25 141.22 141.64 142.04 142.31 142.50 142.87 143.06 143.34 143.80 143.94 144.24 144.55 144.80 145.47 145.59 145.89 146.17 146.35 146.63 147.15 147.33 147.66 148.14 148.30 148.47 148.70 148.93 149.20 149.58 149.81 150.05 150.47 150.98 151.16 151.44 151.86 152.09 152.18 152.37 152.69 152.88 153.21 153.69 154.02 154.37 154.51 154.70 154.88 155.20 155.40 155.57 156.04 156.18 157.04 157.15 157.31 157.97 158.15 158.38 158.80 159.31 159.41 159.78 159.88 160.47 160.66 161.17 161.31 161.63 161.87 162.10 162.40 162.68 162.91 163.28 163.47 163.65 163.90 164.05 164.44 164.61 164.98 165.37 165.86 166.10 166.28 166.77 166.91 167.51 167.61 167.98 168.16 168.30 168.72 169.00 169.23 169.42 169.72 170.02 170.25 170.49 170.63 170.81 171.23 171.46 171.83 171.95 172.09 172.46 172.69 173.06 173.16 174.02 174.39 174.62 174.81 174.94 175.60 175.87 176.24 176.60 177.00 177.17 177.35 177.88 177.98 178.54 178.68 179.05 179.23 179.37 179.79 180.07 180.30 180.49 180.81 181.09 181.32 181.56 181.70 181.88 182.30 182.53 182.90 183.02 183.16 183.53 183.76 184.13 184.23 185.09 185.46 185.69 185.88 186.01 186.71 186.94 187.22 187.34 187.66 188.03 188.31 188.73 188.90 189.06 189.38 189.89 190.17 190.40 190.91 191.17 191.50 191.73 191.91 192.28 192.79 192.98 193.21 193.54 193.68 193.86 194.33 194.56 194.77 195.07 195.53 195.81 196.00 196.23 196.42 196.56 196.88 197.11 197.30 197.58 197.81 198.44 198.62 198.76 199.04 199.48 199.60 200.02 200.34 200.30 200.76 201.50 201.78 202.18 202.73 203.10 203.29 203.66 204.31 204.45 204.64 204.96 205.73 206.48 206.66 207.08 207.40 207.54 207.84 208.17 208.26 208.72 208.91 209.28 209.47 209.79 210.21 210.40 210.58 210.91 211.51 211.65 211.93 212.28 212.65 212.90 212.97 213.37 213.72 214.00 214.32 214.51 214.81 215.11 215.25 215.67 215.85 216.04 216.83 217.11 217.34 217.57 217.94 218.22 218.50 219.01 219.75 219.85 220.08 220.45 220.73 220.87 221.01 221.33 221.52 221.70 222.28 222.47 222.57 222.94 223.14 223.57 223.80 223.98 224.12 224.33 224.54 224.77 225.14 225.74 226.40 226.63 226.77 227.28 227.60 227.74 228.11 228.34 228.48 228.81 229.13 229.51 229.88 230.30 230.57 230.71 230.85 231.13 231.50 231.62 231.75 231.94 232.07 232.50 232.66 232.85 233.31 233.64 233.96 234.29 234.60 235.22 235.50 235.68 235.82 236.10 236.20 236.30 236.43 236.84 237.03 237.26 237.54 238.24 238.33 238.56 238.84 238.98 239.17 239.35 239.63 239.86 240.28 240.60 240.88 241.02 241.37 242.07 242.76 243.28 243.88 244.10 244.20 244.39 244.58 244.74 244.88 245.48 245.81 245.99 246.27 246.55 246.80 247.20 247.48 247.66 248.31 248.64 248.82 249.01 249.15 249.34 249.47 249.71 250.22 250.50 250.87 251.05 251.33 251.63 251.82 251.96 252.15 252.47 253.26 253.45 253.63 254.00 254.19 254.51 254.84 255.40 255.65 255.83 256.36 256.46 257.02 257.16 257.53 257.71 257.85 258.27 258.55 258.78 258.97 259.29 259.57 259.80 260.04 260.18 260.36 260.78 261.01 261.38 261.50 261.64 262.01 262.24 262.61 262.71 263.57 263.94 264.17 264.36 264.49 265.19 265.42 265.80 266.17 266.63 266.77 266.95 267.48 267.58 268.14 268.28 268.65 268.83 268.97 269.39 269.67 269.90 270.09 270.41 270.69 270.92 271.16 271.30 271.48 271.90 272.13 272.50 272.62 272.76 273.13 273.36 273.73 273.83 274.69 275.06 275.29 275.48 275.61 276.41 276.64 278.71 278.94 279.08 279.17 279.50 279.64 279.78 280.01 280.35 280.47

Alineamiento sin audios adicionales:

- Audio 1
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 794

Tiempos de alineamiento:

33.04 35.71 35.81 35.96 37.94 38.15 40.02 40.16 40.33 42.55 42.80 43.07 43.59 43.65 43.79 45.55 45.75 47.45 47.82 48.02 48.46 49.83 49.94 50.77 50.99 51.21 51.77 52.36 52.73 53.07 53.13 53.79 54.11 54.46 54.80 54.92 55.31 55.48 55.86 56.02 56.14 56.57 57.27 57.52 58.10 58.34 58.83 59.05 59.16 59.23 59.71 59.80 60.03 60.46 60.62 60.91 61.00 61.46 61.94 62.31 62.42 62.88 63.08 63.81 63.92 64.03 64.45 64.88 64.97 65.15 65.68 65.97 66.11 66.58 66.76 67.23 67.39 67.85 68.03 68.13 68.45 68.66 69.04 69.33 69.52 70.05 70.64 71.21 71.72 72.13 72.24 72.62 72.90 73.40 73.76 74.08 74.50 74.95 75.39 75.66 75.99 76.76 76.87 77.22 77.55 78.14 78.27 78.63 78.91 79.18 79.39 79.77 80.00 80.25 80.58 80.70 80.88 81.31 81.47 81.72 81.97 82.24 82.51 82.81 83.14 83.42 83.50 83.63 84.01 84.24 84.44 84.83 84.89 85.09 85.53 85.68 85.91 86.24 86.54 86.81 86.98 87.35 87.56 87.69 87.87 88.23 88.40 88.64 89.01 89.21 89.59 89.68 89.80 90.11 90.28 90.49 90.71 91.19 91.52 91.82 91.95 92.23 92.49 92.67 92.73 93.21 93.48 93.68 94.05 94.23 94.85 95.22 95.30 95.76 95.95 96.32 96.41 96.71 96.85 97.05 97.36 97.64 97.71 97.85 98.04 98.60 98.73 98.79 99.40 99.52 100.06 100.35 100.58 100.64 101.00 101.36 101.69 102.00 102.07 102.23 102.34 102.40 103.13 103.29 103.51 103.90 104.07 104.50 104.70 104.79 105.23 105.29 105.72 105.91 106.74 107.15 107.42 107.48 107.53 108.22 108.28 109.03 109.24 109.80 109.93 109.99 110.60 110.72 111.25 111.55 111.79 111.85 112.20 112.56 112.90 113.21 113.27 113.36 113.54 113.60 114.32 114.49 114.71 115.12 115.28 115.70 115.90 115.99 116.42 116.48 116.92 117.11 117.93 118.33 118.62 118.70 118.73 119.42 119.48 119.86 119.98 120.44 121.16 121.47 121.68 121.98 122.10 122.38 123.06 123.32 123.64 123.78 124.04 124.25 124.35 125.19 125.48 125.61 126.23 126.42 126.49 126.89 127.37 127.47 127.60 127.88 128.28 128.55 129.16 129.57 129.69 130.22 130.37 131.28 131.61 131.69 132.16 132.87 132.93 133.09 133.56 134.20 134.51 134.95 135.25 135.40 135.88 136.07 136.40 136.68 136.92 137.31 137.44 137.95 138.03 138.23 138.71 138.82 139.06 139.52 139.71 139.98 140.04 140.19 141.22 141.54 141.97 142.21 142.40 142.83 142.93 143.26 143.64 143.92 144.19 144.44 144.65 145.22 145.51 145.81 146.19 146.35 146.65 147.06 147.24 147.49 147.88 148.05 148.31 148.53 148.93 149.20 149.48 149.80 149.89 150.32 150.83 151.09 151.44 151.75 152.01 152.13 152.30 152.66 152.81 153.06 153.68 153.82 154.26 154.41 154.47 154.71 155.08 155.37 155.49 155.99 156.11 156.86 156.95 157.28 157.88 158.18 158.30 158.75 159.10 159.35 159.70 159.79 160.44 160.48 161.08 161.25 161.45 161.88 162.11 162.19 162.61 162.95 163.24 163.40 163.63 163.89 163.96 164.24 164.33 164.87 165.24 165.80 165.93 165.99 166.61 166.72 167.26 167.55 167.79 167.85 168.20 168.56 168.90 169.19 169.25 169.41 169.54 169.60 170.32 170.49 170.71 171.11 171.29 171.70 171.90 171.99 172.43 172.49 172.92 173.12 173.93 174.34 174.61 174.69 174.73 175.51 175.57 176.25 176.44 177.00 177.13 177.19 177.81 177.92 178.44 178.75 178.98 179.04 179.39 179.76 180.10 180.39 180.47 180.61 180.74 180.80 181.53 181.68 181.91 182.30 182.48 182.91 183.10 183.19 183.63 183.69 184.12 184.31 185.13 185.55 185.82 185.88 185.92 186.63 186.69 187.13 187.35 187.68 188.13 188.24 188.72 188.87 188.96 189.31 189.98 190.11 190.27 190.94 191.23 191.50 191.60 191.88 192.16 192.80 192.87 193.13 193.45 193.74 193.85 194.29 194.42 194.76 194.90 195.59 195.71 195.82 196.02 196.08 196.43 196.81 197.08 197.25 197.55 197.69 198.25 198.54 198.62 198.95 199.38 199.47 199.88 200.14 200.21 200.72 201.16 201.67 202.08 202.65 203.06 203.12 203.46 203.90 204.00 204.50 205.01 205.53 206.30 206.59 206.87 207.28 207.46 207.68 208.10 208.24 208.66 208.81 209.08 209.52 209.70 210.10 210.26 210.45 210.84 211.42 211.52 211.97 212.23 212.59 212.82 212.97 213.22 213.67 213.89 214.21 214.41 214.63 214.99 215.21 215.68 215.81 216.16 216.77 217.04 217.22 217.48 217.92 218.13 218.38 218.83 219.36 219.88 220.00 220.31 220.61 220.72 220.94 221.24 221.30 221.66 222.27 222.36 222.51 222.78 223.12 223.47 223.70 223.76 224.06 224.25 224.48 224.66 224.99 225.67 226.25 226.71 226.77 227.21 227.54 227.58 228.04 228.25 228.30 228.76 229.04 229.40 229.70 230.11 230.53 230.69 230.78 230.97 231.38 231.50 231.77 231.85 232.12 232.50 232.56 232.66 233.22 233.56 233.91 233.97 234.73 235.07 235.35 235.56 235.65 235.98 236.11 236.28 236.32 236.81 237.02 237.26 237.40 238.06 238.10 238.60 238.77 238.98 239.10 239.20 239.61 239.71 240.16 240.70 240.85 240.99 241.32 242.07 242.66 243.28 243.81 243.90 244.00 244.38 244.45 244.61 244.83 245.42 245.64 245.92 246.27 246.37 246.94 247.24 247.46 247.55 248.19 248.55 248.70 248.91 249.05 249.22 249.38 249.49 250.00 250.33 250.73 250.88 251.34 251.53 251.71 252.04 252.10 252.41 253.28 253.39 253.61 253.95 254.09 254.52 254.84 255.40 255.52 255.60 256.21 256.32 256.86 257.15 257.39 257.45 257.80 258.16 258.51 258.80 258.86 258.96 259.14 259.20 259.92 260.09 260.32 260.71 260.88 261.30 261.50 261.60 262.03 262.09 262.52 262.72 263.54 263.96 264.22 264.30 264.33 265.04 265.10 265.80 266.04 266.59 266.73 266.79 267.41 267.52 268.06 268.35 268.59 268.65 269.01 269.36 269.70 270.00 270.06 270.21 270.34 270.40 271.12 271.29 271.51 271.91 272.09 272.50 272.70 272.79 273.23 273.29 273.72 273.92 274.74 275.15 275.42 275.50 275.53 276.23 276.29 278.68 278.82 278.98 279.32 279.50 279.56 279.84 280.00 280.20 280.44

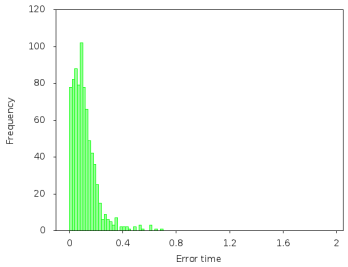
Alineamiento con una canción adicional:

- Audio 1
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 794
- Audio 2
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 1007

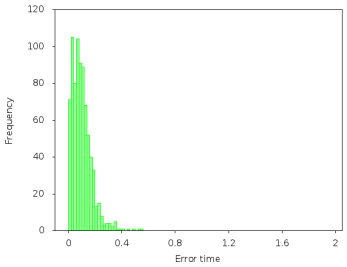
Tiempos de alineamiento:

33.13 35.76 35.94 36.09 37.94 38.19 40.03 40.18 40.31 42.55 42.80 43.07 43.48 43.64 43.85 45.55 45.73 47.78 47.92 48.00 48.53 49.84 49.93 50.77 50.94
51.02 51.30 52.20 52.72 53.08 53.14 53.79 54.14 54.46 54.81 54.98 55.31 55.48 55.86 56.01 56.21 56.56 57.27 57.55 58.17 58.35 58.83 59.06 59.17 59.23
59.74 59.80 60.03 60.47 60.64 60.93 61.02 61.36 61.94 62.32 62.42 62.88 63.09 63.80 63.93 64.03 64.45 64.79 64.90 65.16 65.64 65.97 66.16 66.59 66.75
67.23 67.40 67.71 67.85 68.01 68.46 68.67 69.03 69.28 69.52 70.05 70.63 71.27 71.69 72.02 72.24 72.68 72.89 73.29 73.76 74.09 74.50 75.05 75.39 75.74
76.01 76.76 76.88 77.15 77.52 78.14 78.28 78.58 78.95 79.29 79.39 79.82 80.02 80.27 80.59 80.70 80.83 81.21 81.48 81.64 81.98 82.22 82.51 82.81 83.14
83.43 83.50 83.63 84.02 84.23 84.46 84.84 84.90 85.09 85.53 85.67 85.91 86.21 86.52 86.82 86.98 87.35 87.57 87.70 87.88 88.16 88.40 88.62 89.16 89.22
89.60 89.69 89.75 90.12 90.40 90.49 90.72 91.18 91.53 91.84 91.93 92.12 92.59 92.69 92.78 93.22 93.50 93.71 94.06 94.25 94.91 95.17 95.31 95.76 95.96
96.33 96.42 96.71 96.79 97.09 97.35 97.64 97.73 97.89 98.05 98.61 98.74 98.80 99.39 99.52 100.20 100.34 100.68 100.89 101.01 101.32 101.69 102.00
102.06 102.36 102.61 102.76 103.09 103.32 103.52 103.98 104.07 104.50 104.69 104.78 105.22 105.28 105.73 105.91 106.74 107.16 107.42 107.48 107.59
108.22 108.28 108.97 109.24 109.81 109.94 110.00 110.59 110.72 111.40 111.55 111.89 112.11 112.20 112.51 112.88 113.20 113.26 113.56 113.81 113.96
114.31 114.50 114.71 115.24 115.30 115.70 115.89 115.98 116.42 116.48 116.93 117.11 117.94 118.36 118.61 118.67 118.79 119.43 119.49 119.86 120.04
120.44 121.15 121.48 121.68 121.99 122.11 122.35 123.11 123.34 123.63 123.79 124.05 124.24 124.35 125.19 125.48 125.61 126.24 126.43 126.49 126.88
127.24 127.47 127.61 127.99 128.54 128.60 129.18 129.56 129.71 130.15 130.34 131.28 131.69 131.83 132.15 132.73 132.79 133.14 133.58 134.20 134.54
134.96 135.26 135.41 135.88 136.05 136.38 136.68 136.93 137.31 137.43 137.78 138.03 138.22 138.63 138.71 139.08 139.54 139.72 139.95 140.08 140.14
141.21 141.61 141.95 142.26 142.41 142.84 142.94 143.26 143.65 143.91 144.27 144.44 144.65 145.16 145.43 145.81 146.20 146.35 146.64 147.07 147.24
147.55 147.89 148.27 148.50 148.60 148.93 149.22 149.44 149.82 149.90 150.35 150.86 151.10 151.44 151.74 151.94 152.21 152.30 152.51 152.82 153.07
153.68 153.82 154.25 154.46 154.59 154.69 155.09 155.38 155.50 156.01 156.12 156.87 156.96 157.29 157.89 158.19 158.30 158.75 159.10 159.35 159.70
159.79 160.43 160.48 161.08 161.24 161.50 161.88 162.11 162.20 162.49 162.95 163.19 163.35 163.66 163.80 163.98 164.24 164.33 164.91 165.25 165.80
165.94 166.00 166.59 166.72 167.40 167.54 167.89 168.09 168.21 168.52 168.87 169.17 169.24 169.58 169.80 169.95 170.31 170.51 170.72 171.24 171.30
171.70 171.89 171.98 172.43 172.49 172.93 173.12 173.94 174.37 174.62 174.68 174.86 175.50 175.56 176.14 176.44 177.01 177.14 177.20 177.79 177.92
178.60 178.75 179.09 179.30 179.40 179.72 180.07 180.36 180.44 180.76 181.01 181.16 181.49 181.69 181.91 182.44 182.50 182.90 183.09 183.18 183.63
183.69 184.13 184.31 185.13 185.56 185.82 185.88 186.03 186.63 186.69 187.16 187.26 187.59 188.13 188.24 188.72 188.87 188.95 189.38 189.82 190.12
190.27 190.97 191.23 191.50 191.61 191.83 192.18 192.79 192.85 193.13 193.46 193.64 193.76 194.31 194.43 194.72 194.91 195.58 195.72 195.82 196.14
196.24 196.43 196.79 197.10 197.25 197.51 197.75 198.27 198.54 198.62 199.01 199.39 199.48 199.92 200.13 200.33 200.71 201.16 201.68 202.08 202.70
203.06 203.16 203.49 203.90 204.01 204.41 204.99 205.59 206.32 206.59 206.86 207.29 207.45 207.68 208.07 208.25 208.61 208.81 209.08 209.53 209.69
210.11 210.25 210.45 210.85 211.44 211.53 211.87 212.23 212.59 212.86 212.97 213.22 213.68 213.89 214.22 214.40 214.75 215.01 215.22 215.66 215.81
216.06 216.81 217.08 217.22 217.48 217.93 218.13 218.38 218.85 219.36 219.88 220.01 220.28 220.62 220.73 220.94 221.24 221.30 221.67 222.15 222.29
222.48 222.83 223.12 223.44 223.72 223.78 224.07 224.20 224.48 224.66 225.01 225.56 226.25 226.53 226.59 227.21 227.54 227.59 228.05 228.27 228.30
228.69 229.04 229.41 229.73 230.11 230.52 230.71 230.80 230.99 231.33 231.50 231.74 231.83 232.12 232.50 232.57 232.66 233.23 233.58 233.91 233.97
234.71 235.06 235.26 235.56 235.66 235.98 236.10 236.29 236.32 236.81 237.01 237.30 237.40 238.06 238.11 238.59 238.73 238.94 239.11 239.20 239.58
239.72 240.13 240.70 240.85 241.00 241.32 242.06 242.61 243.28 243.80 244.05 244.08 244.37 244.46 244.63 244.86 245.41 245.79 245.89 246.26 246.37
246.86 247.16 247.47 247.55 248.19 248.54 248.71 248.90 248.99 249.22 249.38 249.59 249.99 250.33 250.73 250.88 251.34 251.54 251.72 252.05 252.11
252.41 253.29 253.42 253.55 253.95 254.10 254.52 254.85 255.40 255.54 255.60 256.20 256.32 257.00 257.14 257.48 257.69 257.80 258.11 258.49 258.78
258.84 259.16 259.41 259.53 259.89 260.11 260.32 260.86 260.92 261.30 261.49 261.58 262.03 262.09 262.53 262.72 263.54 263.97 264.22 264.31 264.45
265.02 265.08 265.78 266.04 266.60 266.74 266.80 267.40 267.52 268.20 268.35 268.69 268.90 269.01 269.32 269.67 269.97 270.04 270.38 270.61 270.77
271.11 271.32 271.52 272.04 272.10 272.50 272.69 272.78 273.23 273.29 273.73 273.92 274.74 275.16 275.42 275.50 275.60 276.23 276.29 278.68 278.81
278.92 279.09 279.45 279.53 279.84 279.93 280.23 280.45

Histogramas:



Sin audios adicionales



Con audios adicionales

Tabla A. 5: Alineamientos e histogramas de las pruebas con el cantante 1. Los tiempos se corresponden con el inicio de cada palabra de los audios

CANCIÓN 2 (You Belong With Me – Taylor Swift)

Tiempos de referencia manual:

8.31 8.77 8.95 9.07 9.50 9.78 9.93 10.92 11.32 12.17 12.36 12.95 13.16 13.64 14.47 15.04 15.25 15.88 15.98 16.45 17.38 17.80 18.25 19.40 19.64 20.21 23.11 23.54 23.71 24.01 24.46 24.70 24.82 25.75 26.30 26.90 27.22 27.56 27.89 28.02 28.40 28.61 29.00 29.39 30.22 30.69 30.83 31.19 32.34 32.55 32.84 33.70 34.17 34.97 37.50 37.67 38.17 38.53 39.02 39.73 39.99 40.51 41.39 41.90 42.37 42.91 43.17 43.61 44.01 44.24 45.16 45.67 45.86 46.08 46.52 46.83 46.95 47.32 47.62 47.72 48.18 48.46 48.71 49.11 49.50 50.04 50.29 50.75 50.94 51.02 51.42 51.92 52.15 52.29 52.43 53.04 53.32 54.20 54.41 54.61 54.84 55.84 56.25 56.46 57.16 57.60 58.21 58.55 59.00 59.53 59.82 62.29 62.48 63.09 63.22 66.09 66.16 66.38 67.07 71.08 71.66 71.78 72.44 72.66 72.99 73.45 73.63 74.42 74.86 74.89 75.22 75.62 76.06 76.36 76.60 76.89 77.30 77.94 78.18 78.98 79.39 79.63 79.66 80.25 80.72 81.50 81.98 82.87 83.08 83.95 84.47 85.33 85.78 86.13 86.47 86.53 87.17 87.47 87.70 87.93 88.34 88.84 89.02 89.79 89.98 90.33 90.69 90.99 91.14 91.19 91.71 91.91 92.12 92.74 92.93 93.68 93.74 93.86 93.98 94.41 94.61 94.73 95.00 95.48 95.73 97.14 97.21 97.61 97.73 98.20 98.55 98.68 99.44 99.61 100.42 100.99 101.54 101.83 102.47 102.70 103.19 104.26 104.69 105.11 106.02 106.17 106.61 106.79 106.97 107.93 108.36 108.63 108.84 109.10 109.37 109.73 110.08 110.35 110.52 110.94 111.08 111.48 111.90 112.32 112.71 112.95 113.18 113.66 113.75 114.30 114.70 114.91 115.07 115.23 115.78 116.05 116.92 117.16 117.37 117.59 118.60 119.02 119.30 119.90 120.45 120.97 121.32 121.65 121.84 122.24 125.08 125.26 125.93 126.38 130.08 131.01 131.24 131.52 132.40 132.70 133.17 133.40 133.78 134.25 134.54 135.15 135.41 136.13 136.51 136.61 137.09 139.89 140.01 140.63 140.76 143.36 143.71 144.39 144.65 157.73 158.23 158.48 159.20 159.57 160.08 160.27 160.52 161.10 161.27 161.35 162.09 162.19 162.30 162.83 163.11 163.27 163.45 163.77 164.11 164.31 164.77 165.00 165.14 165.39 165.68 166.00 166.15 166.79 167.08 167.17 167.23 167.51 167.88 168.46 168.67 168.81 169.06 169.30 169.65 169.76 170.24 170.56 170.65 170.94 171.13 171.39 172.05 172.39 172.47 172.88 173.06 173.30 175.68 176.39 176.52 176.71 176.86 176.94 177.25 177.70 177.91 179.54 180.04 180.45 180.74 181.44 181.93 182.30 182.53 183.24 183.47 185.93 186.17 186.85 187.27 191.01 191.48 192.18 192.58 193.31 193.62 193.78 194.29 194.70 195.16 195.40 196.35 196.77 197.07 197.42 197.54 197.94 200.75 200.95 201.57 201.67 204.57 204.85 205.30 205.78 208.09 208.51 208.96 209.50 211.00 211.14 211.36 211.82 212.26 212.70 215.56 215.70 216.39 216.73 219.41 219.63 220.25 220.36

Alineamiento sin audios adicionales:

- Audio 1
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 398

Tiempos de alineamiento:

5.13 5.59 5.77 5.89 6.32 6.60 6.75 7.74 8.14 8.99 9.18 9.77 9.98 10.46 11.29 11.86 12.07 12.70 12.80 13.27 14.20 14.62 15.07 16.22 16.46 17.03 19.93 20.36 20.53 20.83 21.28 21.52 21.64 22.57 23.12 23.72 24.04 24.38 24.71 24.84 25.22 25.43 25.82 26.21 27.04 27.51 27.65 28.01 29.16 29.37 29.66 30.52 30.99 31.79 34.32 34.49 34.99 35.35 35.84 36.55 36.81 37.33 38.21 38.72 39.19 39.73 39.99 40.43 40.83 41.06 41.98 42.49 42.68 42.90 43.34 43.65 43.77 44.14 44.44 44.54 45.00 45.28 45.53 45.93 46.32 46.86 47.11 47.57 47.76 47.84 48.24 48.74 48.97 49.11 49.25 49.86 50.14 51.02 51.23 51.43 51.66 52.66 53.07 53.28 53.98 54.42 55.03 55.37 55.82 56.35 56.64 59.11 59.30 59.91 60.04 62.91 62.98 63.20 63.89 67.90 68.48 68.60 69.26 69.48 69.81 70.27 70.45 71.24 71.68 71.71 72.04 72.44 72.88 73.18 73.42 73.71 74.12 74.76 75.00 75.80 76.21 76.45 76.48 77.07 77.54 78.32 78.80 79.69 79.90 80.77 81.29 82.15 82.60 82.95 83.29 83.35 83.99 84.29 84.52 84.75 85.16 85.66 85.84 86.61 86.80 87.15 87.51 87.81 87.96 88.01 88.53 88.73 88.94 89.56 89.75 90.50 90.56 90.68 90.80 91.23 91.43 91.55 91.82 92.30 92.55 93.96 94.03 94.43 94.55 95.02 95.37 95.50 96.26 96.43 97.24 97.81 98.36 98.65 99.29 99.52 100.01 101.08 101.51 101.93 102.84 102.99 103.43 103.61 103.79 104.75 105.18 105.45 105.66 105.92 106.19 106.55 106.90 107.17 107.34 107.76 107.90 108.30 108.72 109.14 109.53 109.77 110.00 110.48 110.57 111.12 111.52 111.73 111.89 112.05 112.60 112.87 113.74 113.98 114.19 114.41 115.42 115.84 116.12 116.72 117.27 117.79 118.14 118.47 118.66 119.06 121.90 122.08 122.75 123.20 126.90 127.83 128.06 128.34 129.22 129.52 129.99 130.22 130.60 131.07 131.36 131.97 132.23 132.95 133.33 133.43 133.91 136.71 136.83 137.45 137.58 140.18 140.53 141.21 141.47 154.55 155.05 155.30 156.02 156.39 156.90 157.09 157.34 157.92 158.09 158.17 158.91 159.01 159.12 159.65 159.93 160.09 160.27 160.59 160.93 161.13 161.59 161.82 161.96 162.21 162.50 162.82 162.97 163.61 163.90 163.99 164.05 164.33 164.70 165.28 165.49 165.63 165.88 166.12 166.47 166.58 167.06 167.38 167.47 167.76 167.95 168.21 168.87 169.21 169.29 169.70 169.88 170.12 172.50 173.21 173.34 173.53 173.68 173.76 174.07 174.52 174.73 176.36 176.86 177.27 177.56 178.26 178.75 179.12 179.35 180.06 180.29 182.75 182.99 183.67 184.09 187.83 188.30 189.00 189.40 190.13 190.44 190.60 191.11 191.52 191.98 192.22 193.17 193.59 193.89 194.24 194.36 194.76 197.57 197.77 198.39 198.49 201.39 201.67 202.12 202.60 204.91 205.33 205.78 206.32 207.82 207.96 208.18 208.64 209.08 209.52 212.38 212.52 213.21 213.55 216.23 216.45 217.07 217.18

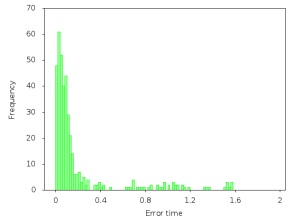
Alineamiento con una canción adicional:

- Audio 1
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 398
- Audio 2
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 369

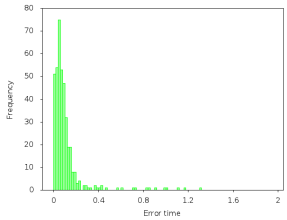
Tiempos de alineamiento:

8.38 8.61 9.07 9.16 9.81 10.01 10.24 11.02 11.46 12.22 12.53 13.01 13.27 13.78 14.64 15.15 15.35 15.90 16.02 16.62 17.54 17.92 18.33 18.91 19.53 20.24 23.16 23.65 23.84 24.13 24.56 24.83 24.86 25.81 26.55 27.12 27.29 27.72 27.99 28.10 28.52 28.61 29.09 29.48 30.24 30.82 30.91 31.28 32.35 32.59 32.94 33.78 34.27 35.05 37.67 37.81 38.29 38.66 39.16 39.78 40.19 40.65 41.54 42.17 42.37 42.68 42.86 42.92 42.98 43.04 44.93 45.76 46.05 46.22 46.48 46.94 47.10 47.46 47.67 47.83 48.41 48.59 48.82 49.30 49.64 50.09 50.35 50.55 51.07 51.29 51.62 51.97 52.24 52.42 52.59 53.18 53.42 54.15 54.21 54.36 55.06 55.97 56.33 56.86 57.85 58.03 58.35 58.58 59.03 59.66 59.90 62.39 62.52 62.64 63.34 63.59 66.07 66.32 67.02 67.29 71.08 71.81 71.87 72.82 72.91 73.27 73.49 73.85 74.68 75.13 75.16 75.56 75.76 76.23 76.51 76.73 77.04 77.35 77.70 78.26 78.85 79.14 79.49 79.89 80.30 80.89 81.31 81.55 82.89 83.17 83.72 84.76 85.72 85.98 86.46 86.63 86.66 87.31 87.47 87.67 88.06 88.44 88.89 89.30 89.57 89.67 90.46 90.81 91.27 91.37 91.40 91.83 92.09 92.28 92.80 92.97 93.57 93.63 93.93 94.12 94.49 94.66 94.87 95.08 96.62 96.71 97.18 97.37 97.74 97.83 98.39 98.88 98.97 99.24 99.78 100.54 101.13 101.65 101.97 102.55 102.95 103.31 104.29 105.02 105.19 105.79 106.23 106.71 106.93 107.13 108.09 108.52 108.82 108.94 109.46 109.74 109.89 110.21 110.45 110.59 111.14 111.35 111.60 112.09 112.39 112.87 113.13 113.29 113.82 114.08 114.40 114.75 115.00 115.17 115.34 115.92 116.16 116.84 116.94 117.54 117.78 118.74 119.11 119.65 120.60 120.80 121.13 121.38 121.81 122.46 122.74 125.16 125.41 126.13 126.42 129.98 130.92 131.37 131.78 132.56 132.84 132.96 133.52 133.84 134.39 134.63 135.34 135.51 136.04 136.21 136.48 137.19 139.94 140.13 140.90 141.18 143.45 143.84 144.56 144.84 157.70 157.76 158.58 159.26 159.64 160.21 160.47 160.77 161.21 161.37 161.43 162.13 162.31 162.44 163.04 163.23 163.36 163.75 163.95 164.24 164.47 164.86 165.12 165.23 165.54 165.78 166.17 166.24 166.90 167.11 167.36 167.54 168.00 168.61 168.82 168.92 169.28 169.44 169.78 169.87 170.40 170.71 170.76 171.07 171.35 171.50 172.06 172.52 172.58 172.97 173.19 173.46 175.83 176.21 176.27 176.90 177.11 177.71 177.88 178.55 178.64 179.69 179.99 180.60 180.83 181.48 182.05 182.35 182.68 183.34 183.60 185.72 186.32 187.09 187.42 190.92 191.87 192.37 192.76 193.50 193.76 193.87 194.45 194.78 195.33 195.56 196.29 196.43 197.02 197.14 197.41 198.11 200.87 201.11 201.84 202.13 204.50 204.72 205.50 205.76 208.11 208.45 209.14 209.40 209.87 211.21 211.39 211.84 212.44 212.79 215.48 215.88 216.58 216.88 219.13 219.57 220.29 220.55

Histogramas:



Sin audios adicionales



Con audios adicionales

Tabla A. 6: Alineamientos e histogramas de las pruebas con el cantante 2. Los tiempos se corresponden con el inicio de cada palabra de los audios

CANCIÓN 3 (My Heart Will Go On – Celine Dion)

Tiempos de referencia manual:

20.39 21.36 22.08 22.65 23.21 24.58 24.80 25.75 27.11 27.26 28.59 30.00 30.87 31.16 31.72 32.22 32.86 34.12 34.88 39.43 40.60 41.53 41.98 43.95 44.33 46.40 48.07 49.47 50.17 50.56 51.10 51.54 52.32 53.95 54.32 59.04 61.30 63.34 65.58 66.11 67.58 68.34 69.88 70.71 71.03 71.78 73.08 73.66 78.45 80.85 82.66 83.35 85.06 85.33 86.97 87.49 88.07 89.30 89.92 90.42 91.19 92.47 92.77 93.55 94.68 95.31 96.64 97.89 107.59 108.49 108.73 109.30 109.76 110.43 111.86 112.39 113.04 114.33 114.83 117.31 118.14 118.73 119.55 120.22 121.75 122.09 126.93 127.84 128.20 128.78 129.04 129.87 131.08 131.70 132.19 133.70 134.02 135.35 136.83 137.48 137.83 138.28 138.88 140.80 141.82 146.24 148.51 150.46 153.03 153.38 154.63 155.68 157.16 158.00 158.44 159.06 160.47 161.14 165.69 168.15 169.84 170.45 172.37 172.77 174.29 174.72 175.30 176.47 177.13 177.66 178.47 179.64 180.22 180.83 181.99 182.54 183.98 185.08 204.34 206.86 208.71 209.30 211.16 211.46 212.86 213.66 213.92 215.48 215.88 216.28 217.07 218.59 219.06 223.95 226.03 228.15 230.53 230.86 232.36 232.90 233.29 234.84 235.37 235.90 236.65 237.70 238.31 238.96 240.23 240.63 241.99 243.06

Alineamiento sin audios adicionales:

- Audio 1
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 172

Tiempos de alineamiento:

20.39 21.36 22.08 22.65 23.21 24.58 24.80 25.75 27.11 27.26 28.59 30.00 30.87 31.16 31.72 32.22 32.86 34.12 34.88 39.43 40.60 41.53 41.98 43.95 44.33 46.40 48.07 49.47 50.17 50.56 51.10 51.54 52.32 53.95 54.32 59.04 61.30 63.34 65.58 66.11 67.58 68.34 69.88 70.71 71.03 71.78 73.08 73.66 78.45 80.85 82.66 83.35 85.06 85.33 86.97 87.49 88.07 89.30 89.92 90.42 91.19 92.47 92.77 93.55 94.68 95.31 96.64 97.89 107.59 108.49 108.73 109.30 109.76 110.43 111.86 112.39 113.04 114.33 114.83 117.31 118.14 118.73 119.55 120.22 121.75 122.09 126.93 127.84 128.20 128.78 129.04 129.87 131.08 131.70 132.19 133.70 134.02 135.35 136.83 137.48 137.83 138.28 138.88 140.80 141.82 146.24 148.51 150.46 153.03 153.38 154.63 155.68 157.16 158.00 158.44 159.06 160.47 161.14 165.69 168.15 169.84 170.45 172.37 172.77 174.29 174.72 175.30 176.47 177.13 177.66 178.47 179.64 180.22 180.83 181.99 182.54 183.98 185.08 204.34 206.86 208.71 209.30 211.16 211.46 212.86 213.66 213.92 215.48 215.88 216.28 217.07 218.59 219.06 223.95 226.03 228.15 230.53 230.86 232.36 232.90 233.29 234.84 235.37 235.90 236.65 237.70 238.31 238.96 240.23 240.63 241.99 243.06

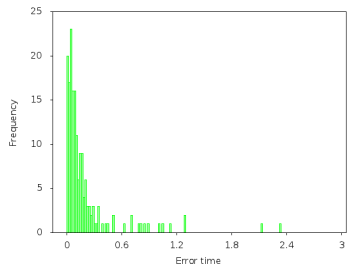
Alineamiento con una canción adicional:

- Audio 1
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 172
- Audio 2
  - Frecuencia de audio: 44100Hz
  - Número de palabras: 240

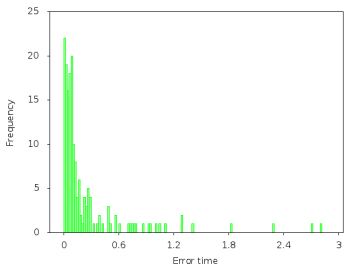
Tiempos de alineamiento:

20.38 21.52 22.15 22.86 23.23 24.61 24.84 26.07 26.72 27.31 29.15 30.08 31.16 31.33 32.43 32.49 33.00 34.71 35.90 39.53 40.36 41.55 42.11 44.58 45.00 46.38 48.31 50.07 50.30 50.71 52.28 52.40 52.49 54.01 54.93 59.13 61.40 63.47 65.76 66.22 67.61 67.75 69.78 70.32 70.46 71.86 73.29 75.69 78.59 81.08 82.83 83.45 85.17 85.47 87.56 88.22 89.18 89.73 89.93 89.99 91.27 92.64 93.59 94.82 94.95 96.37 96.69 97.97 107.62 108.42 108.79 109.42 110.10 110.59 112.10 112.50 114.31 114.85 114.90 117.30 117.44 119.16 119.80 120.10 121.78 122.21 127.08 127.88 128.43 129.42 129.47 130.06 131.39 131.71 132.27 133.67 134.05 135.47 136.74 137.65 137.98 138.55 139.12 140.99 142.19 146.76 148.65 151.09 153.10 153.42 154.87 154.98 157.36 158.16 158.58 159.12 160.75 161.58 166.25 168.16 170.11 170.69 171.84 172.45 174.27 174.90 175.28 177.28 177.36 177.99 178.82 179.92 180.88 182.61 185.82 187.20 187.93 188.10 204.58 206.93 209.34 210.39 211.26 211.44 213.19 213.67 213.72 215.19 215.56 215.91 217.35 218.72 220.07 225.62 226.12 228.21 230.03 230.94 232.63 232.97 233.36 234.57 235.51 235.84 239.13 240.69 241.00 242.11 244.62 244.86 247.00 247.65

Histogramas:



Sin audios adicionales



Con audios adicionales

Tabla A. 7: Alineamientos e histogramas de las pruebas con el cantante 2. Los tiempos se corresponden con el inicio de cada palabra de los audios.

B. TRANSCRIPCIONES EMPLEADAS

A continuación se muestran las transcripciones empleadas en los experimentos tanto para los audios de noticias, como para las canciones. Los asteriscos (*) presentes en las transcripciones son marcas de posición (Ver sección 3.4.2.) que indican dónde deben partirse las transcripciones. Como se verá, las noticias sólo tienen un asterisco al final, lo que indica que no han sido troceados (por tanto tampoco el audio), sino que se ha tomado un único fragmento de transcripción en la entrada del Alineador.

Por el contrario, en las transcripciones de las canciones vemos que hay muchas marcas de posición (asteriscos), lo que indica que, al haber muchas pausas largas en el audio, ha sido necesario trocear el audio varias veces eliminando las partes de pausas largas. Esto troceado lleva implícito la fragmentación de las transcripciones, de tal manera que cada fragmento de audio queda asociado con su correspondiente transcripción.

B.1 TRANSCRIPCIONES DE NOTICIAS

We've got a close up of vampires Bella and Edward Cullen, thanks to Summit for passing along the official photo to us!! Along with one of Taylor Lautner as Jacob, now that one is coming up soon, so just hold the moment! This is all in anticipation of the November 16, 2012 release date for Breaking Dawn Part 2 and the best part about these photos is arguably the fact that we get to see the bright red eyes of Bella post transformation, in fact, it looks like she's already pretty comfortable in her vampire form. She and Edward closely look on at something together in the still photo. And all of you Jacob fans out there, you're also gonna see another serious photo of him. It looks like he's outside in the shot, seemingly about to pounce or do something, exactly what, we just don't know. Put your best guesses in the comments section below as to what is going on in these newly released breaking down part 2 photos. I'm your host Dana Ward in Hollywood, and as always thanks so much for tuning into ClevverTV, bye! *

Tabla B. 1: Transcripción del audio del LOCUTOR 1 con marcas de posición

Zac Efron is the latest Hollywood A-lister to head to the Cannes film festival, where he's been spotted posing for pictures and yes even revealing his latest love. Zac and his co-stars Nicole Kidman, Matthew McConaughey, and John Cusack are in Cannes to promote their film "The Paperboy" Today they sat down for a panel session to answer questions about the film, which is actually included in this year's prestigious Cannes competition. Obviously, this movie is a huge opportunity for Zac, especially they have the opportunity to share the screen with an iconic Oscar winner like Nicole Kidman. In fact Zac was quick to admit his feelings saying QUOTE: "I've been in love with her for a long time, since Moulin Rouge. It was the loveliest time in the world for me." "The Paperboy" is an independent film that takes place in the south in the 1960's. Zac Efron's character is working towards becoming a writer and in the film he helps his journalist brother (played by Matthew McConaughey) look into some mishaps in the justice system. Now over the course of the film, Zac's character finds himself in love with Nicole Kidman's character, who just so happens to be in love with a jailed man, who is played by John Cusack. And here is what you should all get very very excited about, Zac apparently spends the bulk of the movie in his underwear. We don't have too many photos of those scenes unfortunately, but we do have a full gallery of great photos from Cannes featuring Zac Efron handsome as ever in a suit, along with the rest of his "Paperboy" cast, to take a look at those photos click the link below. I'm Joslyn Davis, thanks for stopping by and we'll see you next time right here on ClevverTV. *

Tabla B. 2: Transcripción del audio del LOCUTOR 2 con marcas de posición

Let me finish tonight with this: I fear, I really do, the evil influence of big bad money in politics. Democracy isn't a joke, it shouldn't be anyway. And when a few guys with billions of dollars or influence with big money people can throw enormous amounts of money into a political race, they can destroy democracy. Yes, they can. Just think about what a dump of million dollars in TV advertising really nasty stuff can do to a race for Congress somewhere. All of a sudden, the airwaves are saturated with negative TV ads. They're all over the place on entertainment tonight, on sports, wherever you go; you can't escape the relentless, nasty assault on some candidates. Well, think this doesn't matter? Good luck.
We saw what Willie Horton did, the one Democratic candidate for president, so was swift boating did to another. Just imagine someone out there saying what a bad person you are, telling everyone you know what a bad person you are. OK, imagine someone spending tens of millions of dollars buying TV time to say what an evil SOB you are. Don't think that will change some opinions about you? Think again. Not everybody pays attention to politics like the people who watch this show; some hardworking people just have time to check in around election time. And what do they see on TV? Endless streams of bad news about some guy they once thought was OK, but now can't hear a good word about him or her, just this endless spew of negativity. I don't know how we're going to live with this stuff. We're still calling ourselves a democracy out there. I don't know when the Supreme Court is gonna look again at that monster they created in Citizens United and kill it before it destroys the very democracy on which the Congress, the presidency, and yes, the courts now precariously rest.*

Tabla B. 3: Transcripción del audio del LOCUTOR 3 con marcas de posición

And finally tonight, the power of a photograph. Take a look at this picture, it was snapped three years ago inside the Oval Office and it shows five year old Jacob Philadelphia, the son of a former White House staffer, touching President Obama's head. After three years, this picture still hangs in the west wing while others come and go. So how did this moment become stuck in time? The "New York Times" says, Jacob quietly told President Obama, "I want to know if my hair is just like yours. "The President told Jacob, why don't you touch it and see for yourself. When Jacob hesitated, the President gave him a nudge, "touch it dude" "Yes, it does feel the same" Jacob said. We talk about the importance of President Obama breaking barriers and stereo tapes by becoming the country's first black president and those are important. But we must keep breaking those barriers. Not only for African Americans but women, gays, everybody. Every barrier we break down, we make the country bigger, better, we make it live up to its creed. It's about breaking barriers and it's about making every child, no matter who they are, no matter what their background, know that they can reach the top, the most powerful position in the world, because somebody just like them has achieved that. I'm a boy that grew up without a father at home. I looked up to others to try to be like them. They were the ones that in many ways, was like me. If you can conceive it, you can achieve it. If you can see it, you can be it. And every time we make that clear to our young people in America, we make the country bigger, we make the country better because we make our young people know that their dreams can become a reality. Thanks for watching, I'm Al Sharpton. *

Tabla B. 4: Transcripción del audio del LOCUTOR 4 con marcas de posición

B.2 TRANSCRIPCIONES DE CANCIONES

Look,* if you had,* one shot,* or one opportunity*
To seize everything you ever wanted,*
one moment* Would you capture it?*
Or just let it slip, yo*

His palms are sweaty, knees weak, arms are heavy
There's vomit on his sweater already, mom's spaghetti
He's nervous, but on the surface he looks calm and ready
To drop bombs, but he keeps on forgetting
What he wrote down, the whole crowd goes so loud
He opens his mouth, but the words won't come out
He's choking how, everybody's joking now
The clock's run out, times up, over, blaw!*

Snap, back to reality, oh, there goes gravity
Oh, there goes Rabbit he choked, he's so mad but he won't
Give up that easy, no, he won't have it he knows
His whole back's to these ropes, it don't matter he's dope
He knows that but he's broke, he's so stagnant he knows
When he goes back to his mobile home, that's when it's
Back to the lab again, yo
This whole rhapsody better go capture this moment
And hope it don't pass him*

You better lose yourself in the music

The moment, you own it, you better never let it go, oh
You only get one shot, do not miss your chance to blow
This opportunity comes once in a lifetime*

You better lose yourself in the music
The moment, you own it, you better never let it go, oh
You only get one shot, do not miss your chance to blow
This opportunity comes once in a lifetime, you better.*

His soul's escaping, through this hole that is gaping
This world is mine for the taking, make me king
As we move toward a new world order, a normal life is boring
But superstardom's close to post mortem
It only grows harder, only grows hotter
He blows it's all over, these hoes is all on him
Coast to coast shows, he's known as the Globetrotter lonely roads

God only knows he's grown farther from home, he's no father
He goes home and barely knows his own daughter
But hold your nose 'cos here goes the cold water
These hoes don't want him no more, he's cold product
They moved on to the next schmoe who flows
He nose-dove and sold nada,
so the soap opera is told it unfolds
I suppose it's old partner, but the beat goes on
Da da dum, da dum*

You better lose yourself in the music
The moment, you own it, you better never let it go, oh
You only get one shot, do not miss your chance to blow
This opportunity comes once in a lifetime*

You better lose yourself in the music
The moment, you own it, you better never let it go, oh
You only get one shot, do not miss your chance to blow
This opportunity comes once in a lifetime, you better*

No more games, I'ma change what you call rage
Tear this motherfucking roof off like two dogs caged
I was playing in the beginning, the mood all changed
I've been chewed up and spit out and booed off stage
But I kept rhyming and stepped, writing the next cipher
Best believe somebody's paying the pied piper
All the pain inside amplified by the fact
That I can't get by with my 9 to 5

And I can't provide the right type of life for my family
'cos man, these god damn food stamps don't buy diapers*
And it's no movie, there's no Mekhi Phifer, this is my life
And these times are so hard and it's getting even harder
Trying to feed and water my seed, plus teeter-totter

Caught up between being a father and a prima donna
Baby mama drama's screaming on and too much for me to wanna*
Stay in one spot, another day of monotony's
Gotten me to the point I'm like a snail
I've got to formulate a plot or end up in jail or shot
Success is my only motherfucking option, failure's not
Mom, I love you, but this trailer's got to go
I cannot grow old in Salem's lot, so here I go it's my shot
Feet fail me not, this may be the only opportunity that I got*

You better lose yourself in the music
The moment, you own it, you better never let it go, oh
You only get one shot, do not miss your chance to blow
This opportunity comes once in a lifetime*

You better lose yourself in the music
The moment, you own it, you better never let it go, oh
You only get one shot, do not miss your chance to blow
This opportunity comes once in a lifetime
You better*

You can do anything you set your mind to, man*

Tabla B. 5: Transcripción del audio del CANTANTE 1 con marcas de posición

You're on the phone with your girlfriend, she's upset*
She's going off about something that you said*
'Cause she doesn't* get your humor like I do*
I'm in the room,* it's a typical Tuesday night*
I'm listening to the kind of music she doesn't like*
And she'll never* know your story like I do*

But she wears short skirts,* I wear T-shirts*
She's Cheer Captain and I'm on the bleachers*
Dreaming about the day* when you wake up and find
That what you're* looking for has been here the whole time*

If you could see that I'm the one who understands you*
Been here all along, so why can't you see?*
You belong with me,* you belong with me*

Walking the streets with you and your worn-out jeans*
I can't help thinking this is how it ought to be*
Laughing on a park bench, thinking to myself*
Hey, isn't this easy?*
And you've got a smile that could light up this whole town*
I haven't seen it in a while since she brought you down*
You say you're fine, I know you better than that*
Hey, what ya doing with a girl like that?*
She wears high heels, I wear sneakers*

She's Cheer Captain and I'm on the bleachers*
Dreaming about the day when you wake up and find
That what you're* looking for has been here the whole time*

If you could see that I'm the one who understands you*
Been here all along, so why can't you see?*
You belong with me*
Standing by and waiting at your back door*
All this time how could you not know?*
Baby,* you belong with me,* you belong with me*

Oh, I remember you driving to my house in the middle of the night*
I'm the one who makes you laugh* when you know you're 'bout to cry*
And I know your favorite songs and you tell me 'bout your dreams*
Think I know where you belong,* think I know it's with me*

Can't you see that I'm the one who understands you?*
Been here all* along, so why can't you see?*
You belong with me*
Standing by and waiting at your back door*
All this time, how could you not know?
Baby,* you belong with me,* you belong with me*

You belong with me*
Have you ever thought just maybe*
You belong with me*
You belong with me*

Tabla B. 6: Transcripción del audio del CANTANTE 2 con marcas de posición

Every night in my dreams* I see you,* I feel you*
That is how I know you* go on.* Far across the distance*
And spaces between us.* You have come to show you* go on*
Near,* far,* wherever you are* I believe that the heart does* go on*
Once more* you open the door,* and you're here in my heart
And* my heart will go on and on*

Love can touch us one time* And last for a lifetime*
And never let go till* we're gone* Love was when I loved you*
One true time* I hold to* In my life we'll always* go on*
Near,* far,* wherever you are* I believe that the heart does* go on*
Once more* you open the door* And you're here in my heart
And* my heart will go on and on*

You're here,* there's nothing I fear*
And I know that my heart will* go on.* We'll stay* forever this way*
You are safe in my heart,* and my heart will go on and on*

Tabla B. 7: Transcripción del audio del CANTANTE 3 con marcas de posición

C. PRESUPUESTO

1. Ejecución Material

   Compra de ordenador personal (Software incluido)..... 1200 €
   Material de oficina ..... 150 €
   Total de ejecución material ..... **1350 €**

2. Gastos generales

   16% sobre Ejecución Material ..... **216€**

3. Beneficio Industrial

   6% sobre Ejecución Material ..... **81 €**

4. Honorarios Proyecto

   1500 horas a 15 €/hora ..... **22500 €**

5. Material fungible

   Gastos de impresión ..... 180 €
   Encuadernación ..... 120 €
   Total de material fungible ..... **300 €**

6. Subtotal del presupuesto

   Subtotal Presupuesto ..... **24447 €**

7. I.V.A. aplicable

   18% Subtotal Presupuesto ..... **4400,46 €**

8. Total presupuesto

   Total Presupuesto ..... **28847,46 €**

Madrid, Septiembre 2012
El Ingeniero Jefe de Proyecto

Fdo.: Darwin Patricio Córdova Lucero
Ingeniero Superior de Telecomunicación

D. PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un sistema de *ALINEAMIENTO DE TEXTO Y AUDIO PARA EL APRENDIZAJE DEL IDIOMA INGLÉS.*

En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

CONDICIONES GENERALES

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.
2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.
3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.
4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.
5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.
6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.
7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.
9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.
10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.
11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.
12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.
13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.
14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.
15. La garantía definitiva será del 4% del presupuesto y la provisional del 2 %.
16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.
17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.
18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.
19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.
21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.
22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.
23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad “Presupuesto de Ejecución de Contrata” y anteriormente llamado “Presupuesto de Ejecución Material” que hoy designa otro concepto.

CONDICIONES PARTICULARES

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.
2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.
3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.
4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.
5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.
7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.
8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.
9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.
10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.
11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.

E. PUBLICACIONES

Durante el desarrollo de este proyecto se ha realizado y enviado un artículo científico al Congreso de *IberSPEECH2012,* el cual ha sido aceptado para ser presentado oralmente. El artículo también ha sido seleccionado para su publicación en *Springer Communications in Computer and Information Science* (CCIS).

Título: Preliminary Results of Alignment of Text and Audio in News and Songs

Autor: Darwin Patricio Córdova Lucero.

Coautores: Doroteo Torre Toledano

Referencia bibliográfica: D.P. Córdova Lucero, D. Torre Toledano, "*Preliminary Results of Alignment of Text and Audio in News and Songs", IberSPEECH2012* Conference, November 2012.

– Preliminary Results of Alignment of Text and Audio in News and Songs

Darwin Patricio Córdova Lucero, Doroteo Torre Toledano

ATVS, Escuela Politécnica Superior, Universidad Autónoma de Madrid, SPAIN
dar.cordova@estudiante.uam.es, doroteo.torre@uam.es

Abstract. This paper addresses the problem of forced alignment in news and songs in order to get the times where every word of the transcriptions begins and ends. For this purpose two methods are used. The first one is basically a forced alignment process of the audio and text based on pre-existent models. The second one is a model-free method in which new models are trained on the audio to align producing as a result the aligned text and audio. For analysis of the songs, we have considered two versions of the same song: one is an *a capella* song (only voice with no music) and the other, the full song (with instrumental music included). Three songs have been selected from different singers and different styles. Regarding news, we have analyzed four speakers (2 females and 2 males). Analyzing all the results, we observe that news is better aligned than songs, as expected. The two methods work similarly in both *a capella* songs and news, but in the case of songs that include the instrumental part, the model-free method is much better.

Keywords: Alignment, Songs and Lyrics, Language Learning, Broadcast News

Introduction

This paper presents preliminary experiments on alignment of songs and lyrics and texts and audio news. One of the purposes of this paper is to analyze the difference in the behavior of the forced alignment in songs and broadcast news. To that end, for the analysis of the news, we will be using four speakers (two females and two males), and for the analysis of the songs, we will consider three English songs from three different styles of music: the first one is a very fast-speed song (rap), the second one, a normal-speed song (pop), and, finally, a very slow-speed song (ballad). Two versions of these songs will be considered, one including instrumental music and one *a capella*.

Other of the purposes of this paper is to compare two different ways of producing the alignments. One way is based on pre-existent Hidden Markov Models (HMMs), and another a model-free approach based on training HMM models from scratch using only the audio to align, or this audio complemented by a set of similar audios.

Our main goal is the alignment of songs and lyrics to feed new songs and aligned lyrics into a web-based system (www.inglesdivino.com) that plays songs and videos and shows each word pronounced aligned in real time, among many other possibilities. This system tries to help students to learn and improve their English in less tedious ways. Having songs and lyrics aligned is very useful, for example, for students who are beginning to learn a new language, because they normally get lost when they try to follow the lyrics as they listen to the audio recordings. With this technology that wouldn't happen, since every word will be highlighted as accurately as possible while it's pronounced. All the experiments so far have been done in English, but in the future we plan to expand them to more languages. In general, this system will be useful for learning any language.

This problem is closely related to other similar problems that share in common the need to have audio and text aligned: TV subtitling, entertainment (i.e. karaoke), design of games based on synchronized audio, etc.

The issue of song and lyrics alignment has found some interest in the research community in the last years. Good examples are [1] where pre-existent models are used, [2] where dynamic programming and a model-free method is used and [3] where music and speech try to be first segregated and then pre-existent models are adapted to speech with music. On the other hand, the issue of broadcast news subtitling has been more studied due to its clear application, in particular to allow deaf people to access the content of the news. Broadcast news subtitling can be faced in two different ways, by using speech recognition and obtaining transcription and alignment from audio, as done in [4], or by exploiting knowledge from the news transcription used by the speakers to align text and audio as done in [5]. In this paper we will always use the text for the alignment. We will be comparing the problem of songs and lyrics and the one of text and news alignment and finally we will compare two methods for performing the alignment.

Proposed Methods

The alignments of text and audio will be performed using two methods: the first one is based on pre-existing English phonetic HMM models, and the other one (model-free method) is based on training HMM models from scratch using the audio to align (and possibly some similar complementary audios). In both cases, models will be used or trained using HTK [6]. Next subsections explain in more detail these methods.

Using Pre-Existing Models

In this method, we use English phonetic HMM models previously trained with 8 KHz English audio (TIMIT corpus [7]). The models have been created for each phone of English, with 40 Gaussians per state and 3 states per phone. For these experiments we use the models without any modification. In order to obtain our times of interest, we perform the following steps:

Prepare input data. In this case we need the audio recording and its transcription (a word level transcription).

Audio recording

HELLO WORLD

Transcription

Fig. 1. Audio recording and its transcription

Parameterize the audio. Here we convert the audio file to MFCCs (Mel Frequency Cepstral Coefficients).

Convert the word level transcription into a phone level grammar. For this, we use an English phonetic dictionary derived from the CMU pronouncing dictionary [8]. This phone level grammar will be used to create a network of HMM phone models.

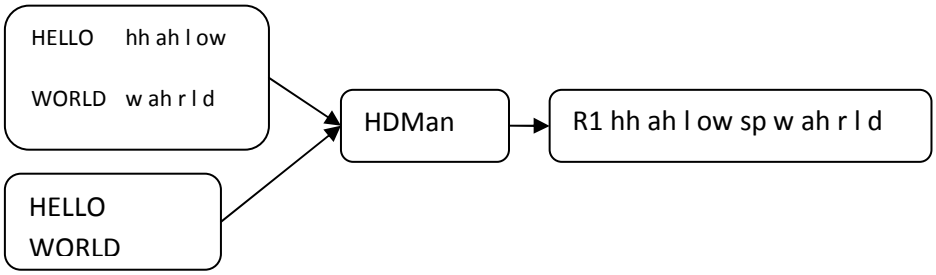


Fig. 2. Creation of phone-level grammar from word-level transcription and dictionary.

Perform the alignment. The alignment is performed by the HVite tool. It will match the parameterized audio against the created network of HMMs and output the beginning and ending time for each phone and word.

Once the alignment is performed, we extract the beginning and ending time of each word, and then make the comparison with the manual reference. We will show the results in Section 3.

Model-Free Alignment: Aligning During Training

This method is based on the model training process. What we do here is to train phone models from the data we want to be aligned. In our case this data could be songs or news recording. During the training process, HVite and HERest are used for realigning and retraining the models, giving as a result a phone-level alignment of the input data. Compared to the previous method, this one has the advantage that it uses acoustic models that are completely adapted to the data to process with respect to speaker, presence of music, noises, etc. It is well known that the best results in recognition are achieved when we try to recognize the data used for training. That is precisely what we do in this method. Normally using test data for training is not fair, but in this particular application it is perfectly valid. We use as input data for model training the data (audio and text) we want to align, and then as a result of the training process we obtain the alignment. Of course, there are also disadvantages. The main one is that using only the audio

and text to align means using a very limited amount of data. We will try to alleviate this by adding other audios and texts from the same speaker and in similar conditions (to the extent that it is possible) to improve the training and alignment process. The following describes the steps of this method:

1. Prepare input data as in the previous method. We prepare the audio recording and its transcription (a word level transcription). The main novelty here is that we may be interested in preparing additional transcriptions and audios with similar features (speaker, acoustic conditions, etc.) to help in the training and alignment process by adding more data.
2. Parameterize the audio as in the previous method, converting it to MFCCs.

   Convert the word level transcriptions into phone level grammar, as in the previous method, using again an English phonetic dictionary derived from the CMU pronouncing dictionary [8].

   With all the necessary data prepared, we proceed to train the acoustic models of each phone appearing in the grammar we have previously defined. We start defining a prototype of a model and creating "flat start" monophones using the HTK HCompV tool. Then, these "flat start" monophones are re-estimated using the HERest tool. The purpose of this is to load all the "flat start" monophones and re-estimate them using the MFCC files generated from our training data (audios of our songs or broadcast news) and create a set of new models. We do this re-estimation four times.

   In the final step a realignment of the training data is performed using the HVite tool. This tool can consider all pronunciations for each word (in the case where a word has more than one pronunciation in the grammar), and then output the pronunciation that best matches the acoustic data. HVite gives us a first alignment of the data. We use this alignment to re-estimate the models and get more accuracy. We re-estimate (with HERest) four more times using the output of the HVite (the first alignment). After this process, once all the re-estimation has been done, we have the models ready and use them to realign the training data. From the alignment obtained in this process, we will extract the final times for comparing with the manual reference.

Results

Experimental Data

For the experiments with broadcast news we have chosen four segments from YouTube containing four speakers: two females and two males. The duration of the audios is around a minute and a half. Regarding songs, three songs have been selected to cover different styles: The first one is a very fast-speed song (rap), the second one is a normal-speed song (pop) and the last one a very slow-speed song (ballad). The experiments for the model-free method will be performed with audios with a sampling rate of 44100 Hz. Two experiments will be carried out, the first one consists of introducing as input data only the song or piece of news to be aligned, and the second one consists of adding extra audios to help in the training process that produces the alignment. In other words, in the last case, apart from the audio we want to align we introduce more audios from the same speaker or the same singer. These extra audios are only used to improve the accuracy of the alignment.

Results with the Model-Free Method

We will first show the results obtained with the model-free method. Results referring to the method based on pre-existing models will be shown in Section 3.3.

Results are presented showing the percentage of words with segmentation errors smaller than certain values of tolerances, which were chosen to be 50, 100 and 200 ms, because the target application is relatively robust to segmentation errors and most probably errors of 100 ms could remain unnoticeable. These evaluation metrics are similar to those used in [9]. Tables 1 and 2 show a comparison of results obtained on the experiments using only a single audio or additional audios for broadcast news speakers and singers.

These results show that, although there are some cases where the model-free method works well even with one audio, it is when we have access to other audios from the same speaker or from the same singer where the method reaches better performance. To illustrate this improvement when we add more input data in song and lyrics alignment, Figure 3 shows an example histogram of the absolute value of the errors found in the alignment when no added data is used and when only two additional audios are used. As it can be seen, the error in the alignment reduces considerably when adding more data.

It is important to note that there are speakers (speaker 4) and songs (song 3) that are particularly problematic for this method. In both cases we found that speech was slow, which seems to be particularly problematic for this method.

Table 1. Percentage of words (%) in broadcast news with errors smaller than three values of tolerance (50, 100 and 200 ms) with only one audio and with additional audios.

	Single Audio			Additional Audios		
Tolerance	50 ms	100 ms	200 ms	50 ms	100 ms	200 ms
SPEAKER 1 (female)	(188 words and 1 audio)			(895 words and 4 audios)		
	7.45	19.68	19.68	29.79	50.53	93.09
SPEAKER 2 (female)	(223 words and 1 audio)			(1181 words and 4 audios)		
	24.50	49.67	91.06	28.81	55.30	96.69
SPEAKER 3 (male)	(319 words and 1 audio)			(1486 words and 3 audios)		
	1.57	3.13	10.97	35.11	57.68	96.87
SPEAKER 4 (male)	(318 words and 1 audio)			(950 words and 2 audios)		
	2.52	5.03	9.12	3.46	5.35	8.18

Table 2. Percentage of words (%) in songs with errors smaller than three values of tolerance (50, 100 and 200 ms) with only one audio and with additional audios.

	Single Audio			Additional Audios		
Tolerance	50 ms	100 ms	200 ms	50 ms	100 ms	200 ms
Singer 1 (fast-speed song)	(1 song and 794 words)			(2 songs and 1801 words)		
	27.71	56.55	88.54	28.72	59.07	92.44
Singer 2 (normal-speed song)	(1 song and 398 words)			(2 songs and 767 words)		
	38.94	65.08	80.90	41.96	73.12	92.46
Singer 3 (slow-speed song)	(1 song and 172 words)			(2 songs and 412 words)		
	1.14	1.71	4.00	0.00	0.57	1.71

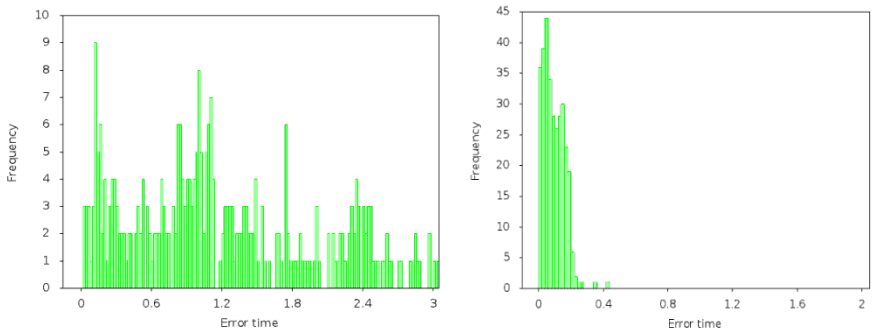


Fig. 3. Comparison of time errors (in seconds) in the cases where there the alignment is performed using only one audio (left), and when two more audios are added as the input data (right).

Comparison of methods

Now we compare the results obtained with the model-free method with the results obtained using the method based on pre-existing models. Since the previous results have been obtained with audios with a sampling rate of 44100 KHz, and taking into account that for the method based on pre-existing models it is necessary to work with audios of 8 KHz (due to availability of trained models in our particular case), we need to resample our audios to 8000Hz to compare their alignments in a fair way. Now we will show the results obtained with the method based on pre-existing models, the results obtained with the model-

free method with 8000Hz audios, and results obtained also with the model-free method, but with a sampling rate of 44100Hz.

Table 3. Comparison of different methods and sampling frequency for broadcast news. Table shows percentage of words (%) with errors smaller than three values of tolerance (50, 100 and 200 ms). For the model-free method we use always additional audios.

SPEAKER 1 (female)	Results			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	28.72	59.57	90.96
	Model-free method (8000Hz)	28.72	48.40	80.32
	Model-free method (44100 Hz)	29.79	50.53	93.09

SPEAKER 2 (female)	Results			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	30.46	65.23	86.42
	Model-free method (8000 Hz)	26.49	54.30	95.70
	Model-free method (44100 Hz)	28.81	55.30	96.69

SPEAKER 3 (male)	Results			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	34.80	67.40	94.36
	Model-free method (8000Hz)	34.80	56.43	95.92
	Model-free method (44100 Hz)	35.11	57.68	96.87

SPEAKER 4 (male)	Results			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	26.73	55.66	88.05
	Model-free method (8000Hz)	0.00	1.26	7.55
	Model-free method (44100 Hz)	3.46	5.35	8.18

Table 4. Comparison of different methods and sampling frequency in *a capella* songs. Table shows percentage of words (%) with errors smaller than three values of tolerance (50, 100 and 200 ms). For the model-free method we use always additional audios.

Singer 1	Results (*a capella*)			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	13.98	33.38	75.57
	Model-free method (44100 Hz)	28.72	59.07	92.44

Singer 2	Results (*a capella*)			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	40.45	61.56	78.14
	Model-free method (44100 Hz)	41.96	73.12	92.46

Singer 3	Results (*a capella*)			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	0.57	1.14	1.14
	Model-free method (44100 Hz)	0.00	0.57	1.71

The results obtained above for the singers are from *a capella* songs. We have performed a comparison of *a capella* songs with those that include instrumental music as well for one particular singer.

Table 5. Comparison of different methods and sampling frequency in songs with music. Table shows percentage of words (%) with errors smaller than three values of tolerance (50, 100 and 200 ms). For the model-free method we use always additional audios.

Singer 2	Results (with music and *a capella*)			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz, with music)	7.04	16.33	44.22
	Model-free method (44100 Hz, with music)	27.64	52.01	80.40

	Model-free method (44100 Hz, *a capella*)	41.96	73.12	92.46

Finally, we perform an analysis on how the number of songs (all with instrumental music) used as input data improves the final result. Again we perform this with only one song.

Table 6. Comparison of results using different number of additional audios (songs) for singer 2. Table shows percentage of words (%) with errors smaller than three values of tolerance (50, 100 and 200 ms).

Singer 2 (44100 Hz, with music)	Tolerance / Number of songs	50 ms	100ms	200ms
	1	25.13	48.99	76.38
	2	27.64	52.01	80.40
	3	26.88	55.03	82.01
	4	28.14	56.28	85.93
	5	30.65	53.77	82.66
	6	29.65	51.76	79.40
	7	32.91	59.80	84.67
	8	31.16	48.47	84.17

Discussion

As expected, results in broadcast news are better than those obtained in songs. Results show also (Table 3) that in the case of broadcast news pre-existing models are quite robust even in the case of very slow speech (as in speaker 4). On the other hand, the model-free approach completely fails at aligning very slow speech, while its results for other speakers are similar as those obtained with the pre-existing models method. Therefore it seems that the model-free method is not a good alternative to the pre-existing models method for broadcast news. We must point out that, in order to make the comparison fairer we report results using 8 kHz for both the pre-existing models approach and the model-free method. While using 8 kHz is required (due to the models available in our case) in the pre-existing method, it is not necessary in the model-free method. Table 3 shows that the model-free method can take advantage of this extended bandwidth yielding results slightly better than the pre-existing method with limited bandwidth for the three first speakers. We can see (in Table 1 and 3) that the results for the *speaker 4* are very poor due to the slow speech, as mentioned before. In this case, when we add an additional audio, the results get even worse. When we analyzed why we found that in the added audio, there was a small segment of around 20 seconds in which the voice of a different speaker appears. This example points out a real danger that we must deal with in a real-life scenario with the model-free method.

Although our results are still very preliminary, they seem to indicate that, unless the speech to align is very slow (as in the cases of speaker 4 and song 3), the model-free method tends to work better than the method based on pre-existing models in songs and particularly when music is included. Results seem to indicate that, the faster is a song, the better results we obtain. Songs which are very slow have very bad results. In our experiments we have made several experiments with songs: first we have performed the alignment of *a capella* songs, then we have compared with the case in which the instrumental music is included, and finally we have studied to what extent the introduction of additional songs improves the results on audios with instrumental music included.

With respect to the behaviour with songs with different types of audio (*a capella* or not), in the case of *a capella* songs, the model-free method performs better, but it is when we introduce instrumental music, when the difference is more evident in favour of the model-free method. It is logical since the pre-existing models are trained with speech only, while in the model-free method the music and environmental conditions are naturally incorporated during the training process.

Table 6 analyzes how much the introduction of additional audios improves results. Results show that there is a tendency towards improvement of results, however, this tendency is not monotonic and there are maximums and minimums suggesting that some audios may help while others actually decrease performance. In this particular case we find the first maximum (in 100 an 200ms) with four audios, but in other experiments we have found that maximum with only one additional audio.

Conclusions and Future Work

One of the main conclusions of the paper is that the use of the model-free method can be an alternative for performing alignments to the method using pre-existing models, particularly in the case of songs. This method is more robust to audio and speaker particularities and could benefit from the possibility of adding more similar data for training. This possibility, however, has some risks that have to be dealt with in the future such as the risk of including speech from other speaker or including songs from the same singer, but very different from the one being aligned. This model-free method is particularly interesting when instrumental music is present in the song to align. One curiosity that we found is that results tended to be better for fast songs than for slow songs, which may be counterintuitive. Our results, however, should be taken with cares since they are still preliminary and to be more conclusive more experimentation is required.

As future work we would like to deepen our analysis, to extend the experiments including a larger number of songs and news fragments and to find ways to improve the alignment of slow-speed songs and speech in the model free-method. We would also like to extend our study of the influence of the number of songs to be included in the model-free method.

Acknowledgements

This research has been partially supported by the Ministry of Education of Spain under project TEC2009-14719-C02-02 (PriorSpeech) project and by the Regional Government of Madrid under MA2VICMR project.

References

Annamaria Mesaros and Tuomas Virtanen, 'Automatic Alignment of Music Audio and Lyrics', Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08), Espoo, Finland, September 1-4, 2008.

Kyogu Lee and Markus Cremer, 'Segmentation-Based Lyrics-Audio Alignment Using Dynamic Programming', In Proc. ISMIR 2008, pp. 395-400.

Hiromasa Fujihara, Masataka Goto, Jun Ogata, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals, in Proceedings of the Eighth IEEE International Symposium on Multimedia (ISM'06).

Hugo Meinedo, Alberto Abad, Thomas Pellegrini, Joao Neto, Isabel Trancoso, The L2F Broadcast News Speech Recognition System, in Proc. FALA 2010: VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, pp. 93-96.

A. Ortega, J. Garcia, A. Miguel, and E. Lleida, "Real-time live broadcast news subtitling system for spanish," in Proc. Interspeech 2009, Brighton, September 2009.

Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu X, Moore G, Odell J, Povey D, Valtchev V, Woodland P, "The HTK Book", Version 3.4 March 2009.

TIMIT Acoustic-Phonetic Continuous Speech Corpus, LDC Catalog Number LDC93S1, Available through the Linguistic Data Consortium, http://www.ldc.upenn.edu.

CMU Pronouncing Dictionary, online, available on ftp://ftp.cs.cmu.edu/project/speech/dict/ (accessed 25 Jun 2012).

D.T. Toledano, L.A. Hernández, L. Villarubia Grande "Automatic Phonetic Segmentation", IEEE transactions on speech and audio processing, 11(6), November 2003.

Conference Management Toolkit <cmt@microsoft.com> 19 de Julio de 2012 08:07

Para: Darwin Córdova <dar.cordova@estudiante.uam.es>

Dear Darwin Patricio Córdova Lucero,

We are very pleased to announce that your paper:

22 - Preliminary Results of Alignment of Text and Audio in News and Songs

has been ACCEPTED for ORAL presentation at the conference.

Your paper has also been selected for publication in Springer Communications in Computer and Information Science (CCIS). This selection has been based on reviewer scores and only a small subset of the papers has been selected.

IMPORTANT INSTRUCTIONS FOR Springer CCIS PUBLICATION: You will shortly receive further instructions for this publication. We will need additional material such as:
1) The PDF camera-ready manuscript.
2) The sources of your camera-ready manuscript: full LaTeX package including Tex and Bib files, or Microsoft Word or RTF documents.
3) A filled-in, signed and scanned, Springer copyright form that we will send you soon.

Please, follow this link (https://cmt.research.microsoft.com/IS2012/Protected/Author/) to access the comments from the reviewers and submitting the camera ready paper and additional materials (Deadline July 30, 2012), addressing the improvements proposed by reviewers if possible.

Please, remember that at least one of the authors is required have a full registration to the conference and expected to present the work.

Thanks very much for your contribution,

IberSPEECH2012 Program Chairs.