

UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR



PROYECTO FIN DE CARRERA

**SEGMENTACIÓN DE AUDIO Y DE
LOCUTORES PARA RECUPERACIÓN DE
INFORMACIÓN MULTIMEDIA Y SU
APLICACIÓN A VIDEOS DE
INFORMACIÓN TURÍSTICA**

Ingeniería de Telecomunicación

José Antonio Morejón Saravia
Septiembre 2011

**SEGMENTACIÓN DE AUDIO Y DE
LOCUTORES PARA RECUPERACIÓN DE
INFORMACIÓN MULTIMEDIA Y SU
APLICACIÓN A VIDEOS DE
INFORMACIÓN TURÍSTICA**

AUTOR: José Antonio Morejón Saravia

TUTOR: Doroteo Torre Toledano

Área de Tratamiento de Voz y Señales (ATVS)
Dpto. de Tecnología Electrónica y Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Septiembre 2011



Este trabajo ha sido parcialmente co-financiado por la Comunidad de Madrid y el Fondo Social Europeo bajo el programa I+D en TIC MA2VICMR (2010-2013). El núcleo del consorcio está formado por un equipo multidisciplinar de ingenieros, científicos, lingüistas y documentalistas provenientes de grupos consolidados y empresas con clara vocación por la I+D+i, trabajando de forma conjunta y coordinada en el desarrollo de Sistemas de Acceso, Análisis y Tratamiento de la Información Multimedia y Multilingüe en la Red.

Agradecimientos

El proyecto fin de carrera pone fin a una etapa de formación académica, intelectual y de preparación previa antes de incorporarse al ejercicio de la vida laboral. Finaliza un largo período de orden, constancia, disciplina, hábitos de estudio, esfuerzo y sacrificio, cuya recompensa tiene un valor añadido incalculable en el futuro de una persona.

En primer lugar mostrar mi agradecimiento a mi tutor, Doroteo Torre Toledano, por darme la oportunidad de trabajar y aprender con él en la realización de este proyecto, y la posibilidad de pertenecer al Área de Tratamiento de Voz y Señales. Recordar también a todo el personal docente de la Escuela Politécnica Superior por aportar su granito de arena en mi formación.

Muchas personas son las que han caminado a diario por el duro y frío suelo, pero pocas las que se reflejan en él. A ellos agradecerles su apoyo durante los años de continua batalla tras días interminables, sin descanso y permaneciendo siempre de pie y con la cabeza alta. Al final ha merecido la pena y ahora veo como la gente no puede creer en lo que nos hemos convertido.

Un recuerdo agradable para aquellos años de colegio entre compañeros y profesores, donde la juventud e ignorancia de la vida, me cegaba ante la evidencia y a pesar de ello aprendí. Inolvidables las tardes en nuestros tres asientos del sector doscientos treinta y tres, ya en sus últimos años de vida.

Gracias a mi hermana Carmina, por ser tan alegre, decidida y sutil. Por compartir tantos momentos y vivencias durante los años, siempre tendrás una ayuda innegable en mí. Te queda un mundo apasionante por vivir, aprovéchalo.

Mi familia, fuente de realidad y experiencia que nunca abandona a pesar de la derrota. Gracias a mis padres, José Luis y María del Carmen por su paciencia, tesón y esfuerzo diarios, dedicados a la progresión de sus hijos. Siempre llevaré conmigo los valores de respeto, igualdad, responsabilidad, honestidad, generosidad, esfuerzo, empatía, entre otros, con los que nos habéis educado. Trataré de transmitirlos con la alegría característica de la tierra que os vio nacer.

Especial mención a mis abuelos y familiares más cercanos pues me han brindado una valiosísima experiencia de vida. Su demostración de humildad, audacia y valentía son un ejemplo diario de superación.

Muchas Gracias.

*José Antonio Morejón Saravia
Septiembre 2011*

"Para no fracasar en la vida, hay una premisa clara:
no esperar nada del amigo y esperarlo todo del enemigo"

*A mis Padres María del Carmen y José Luis
y mi Hermana Carmina*

Resumen

Resumen

En el presente proyecto se estudian los sistemas de segmentación de audio, identificación de locutores e identificación del idioma, con el fin de obtener información multimedia en videos de información turística.

En la actualidad, los sistemas de reconocimiento de locutores se están orientando a resolver problemas reales, como pueden ser grabaciones de voz con interferencias, diafonías, ruidos introducidos por el canal de transmisión, música de fondo, etc.

Tras un estudio del estado del arte en segmentación de audio y reconocimiento de locutores, se explicarán los sistemas que se han diseñado e implementado para el presente proyecto. En concreto se trata de dos sistemas usando diferentes tecnologías. Una de ellas permite la segmentación e identificación locutores mediante la adaptación y optimización a la base de datos del Sistema de Segmentación de Audio diseñado en el ATVS y presentado en la evaluación de Albayzin 2010. Por otra parte, se emplea un reconocedor fonético de audio para diferenciar y separar segmentos con voz de segmentos de no voz.

Los resultados que ofrecen ambos métodos se compararán, mostrando tasas de acierto y error tras la aplicación de los sistemas diseñados a la base de datos de vídeos de información turística española proporcionada por el proyecto MA2VICMR de la Comunidad de Madrid. Debido a la compatibilidad de los sistemas realizados, también se cruzarán resultados con el fin de obtener una mejora en el reconocimiento de segmentos, donde un sistema sea más débil que el otro.

Por último se obtendrá el idioma de cada segmento identificado como voz, usando un segundo reconocedor fonético, cuyos diccionarios y fonemas se han entrenado, diseñado y adaptado a los idiomas que se pueden encontrar en los vídeos: español ó inglés.

Palabras Clave

Audio, diarización, GMM, HTK, HMM, información multimedia, locutores, MA2VICMR, MLLR, reconocimiento fonético, score, segmentación, sistema de segmentación de audio, sistema de reconocimiento de idioma, solapamiento, tasa de acierto, tasa de fallo, turismo, música, voz y música.

Abstract

Abstract

This project studies the Audio-Segmentation systems, speaker diarization and language recognition, in order to obtain information on tourist videos.

Nowadays, speaker recognition systems are moving to solve real problems in our lives, such as voice recording with interference, crosstalk, noise introduced by the transmission channel, background music, etc..

After art study about audio segmentation and speaker recognition, it will be explain the designed and implemented systems for this project. In fact, there are two systems using different technologies. One of them allows audio segmentation and speaker diarization by adapting and optimizing the segmentation system designed in the research group ATVS-UAM and presented in Albayzin 2010 evaluation. On the other hand, uses a phonetic recognizer to differentiate and separate speech segments of non-speech segments.

Results offered by both methods are compared, showing success and failure rates. Designed systems will be work with MA2VICMR video database (co-financed project by Madrid Community) that contains information about Spanish tourism. Due to systems compatibility, it will cross results in order to obtain an improvement in recognition segments, where one system was weaker than the other.

Finally we get the language of each speech segment identified like this, using a second phonetic recognizer, whose dictionaries and phonemes have been trained, designed and adapted to video languages that can be found: Spanish or English.

Key Words

Audio, audio-segmentation, diarization, failure rate, GMM, HTK, HMM, MA2VICMR, MLLR, multimedia information, speaker recognition, speech, phonetic recognition, score, segmentation, success rate, idiom speaker recognition system, overlap, tourism, music, speech y music.

Índice General

<i>Agradecimientos</i>	<i>I</i>
<i>Resumen</i>	<i>V</i>
<i>Abstract</i>	<i>VII</i>
<i>Índice General</i>	<i>IX</i>
<i>Índice de figuras</i>	<i>XII</i>
<i>Índice de tablas</i>	<i>XIII</i>
1. Introducción	14
1.1. Motivación del proyecto	14
1.2. Objetivos	15
1.3. Organización de la memoria	16
2. Estado del arte	18
2.1. Introducción	18
2.2. Sonido y procesamiento humano del habla	19
2.2.1. El sonido.....	19
2.2.2. Producción del habla	21
2.2.2.1. Cavidades infraglólicas.....	22
2.2.2.2. Cavidad laríngea	22
2.2.2.3. Cavidades supraglólicas.....	23
2.3. Percepción del sonido	26
2.4. Reconocimiento fonético	29
2.4.1 Fono y fonema.....	29
2.4.2 Modelos fonéticos	31
2.5. Reconocimiento automático de audio	33
2.5.1 Procesamiento de señales digitales de audio	34
2.5.1.1. Muestreo	35
2.5.1.2. Cuantificación.....	35
2.5.2 Parametrización de la voz.....	36
2.5.2.1. Filtrado de Preénfasis	37
2.5.2.2. Enventanado	37
2.5.2.3. Transformada discreta de Fourier	39
2.5.2.4. Transformada discreta del coseno.....	39
2.5.2.5. Análisis Espectral	40
2.5.2.6. Análisis Cepstral	42
2.6. Reconocimiento de voz usando HMMs	43

2.7.	Algoritmos de extracción de características	46
2.7.1	Mel Frequency Cepstral Coefficients	46
2.8.	Segmentación de Audio.....	48
2.9.	Diarización de Locutores	49
2.9.1	Sistema de Diarización de Locutores ATVS-UAM	49
2.9.2	Sistema de Diarización de Locutores empleando Reconocimiento Fonético	49
2.10.	Identificación de Idioma	50
2.10.1.	Introducción	50
2.10.2.	Niveles de distinción idiomática.....	50
2.10.3.	Aplicaciones	51
2.10.4.	Técnicas empleadas.....	51
2.10.4.1.	Sistemas basados en GMMs (Gaussian Mixture Models)	51
2.10.4.2.	Sistemas SVMs (Support Vector Machines)	53
2.10.4.3.	Sistemas de reconocimiento fonético: PRLM, PPRLM y PPR.....	53
3.	Diseño y Desarrollo.....	55
3.1.	Introducción.....	55
3.2.	Medios disponibles	57
3.2.1.	Bases de datos	57
3.2.2.	Software	60
3.2.3.	Hardware	61
3.3.	Sistemas desarrollados.....	62
3.3.1	Preparación de los datos	62
3.3.2	Sistemas de Segmentación de Audio y Diarización de Locutores.....	62
3.3.2.1.	Segmentación de Audio ATVS-UAM.....	63
3.3.2.2.	Reconocimiento fonético	65
3.3.2.3.	Reconocimiento fonético con adaptación al locutor	67
3.3.2.4.	Combinación de resultados de Segmentación de Audio y Reconocimiento fonético	67
3.3.3	Sistema de Identificación de Idioma	69
4.	Pruebas y Resultados.....	71
4.1.	Pruebas realizadas	71
4.2.	Resultados experimentales	73
4.2.1.	Forma de presentación de resultados.....	73
4.2.1.1.	Evaluación de los sistemas de Segmentación y Reconocimiento de locutores.....	73
4.2.1.2.	Evaluación del sistema de Identificación de Idioma.....	76
4.2.2.	Tablas de Resultados.....	76
4.2.2.1.	Sistema de Segmentación de Audio ATVS-UAM.....	77
4.2.2.2.	Sistema de Segmentación de Audio mediante rec. fonético.....	78
4.2.2.3.	Sistema de Segmentación de Audio mediante rec. fonético y adaptación al locutor.....	79
4.2.2.4.	Combinación de sistemas mediante función AND.....	80
4.2.2.5.	Combinación de sistemas mediante función OR.....	81
4.2.2.6.	Resultados Globales de Segmentación de audio y Reconocimiento de locutor.....	82
4.2.2.7.	Sistema de Identificación de Idioma mediante HMMs de 1 Gaussiana.....	83
4.2.2.8.	Sistema de Identificación de Idioma mediante HMMs de 2 Gaussianas.....	84
4.2.2.9.	Sistema de Identificación de Idioma mediante HMMs de 3 Gaussianas.....	85
4.2.2.10.	Sistema de Identificación de Idioma mediante HMMs de 4 Gaussianas.....	86
4.2.2.11.	Resultados Globales en Identificación de Idioma.....	87
5.	Conclusiones y Trabajo Futuro.....	88
5.1.	Conclusiones	88
5.2.	Trabajo futuro	89
	Bibliografía.....	91
	Glosario de Términos.....	94

<i>Anexo A: Base de Datos MA2VICMR</i>	<i>I</i>
<i>Anexo B: Tablas de Resultados</i>	<i>V</i>
<i>Anexo C: Presupuesto</i>	<i>XVI</i>
<i>Anexo D: Pliego de Condiciones</i>	<i>XIX</i>

Índice de figuras

ILUSTRACIÓN 1. MODELO DE COMUNICACIÓN ORAL.....	19
ILUSTRACIÓN 2. APLICACIÓN DE ENERGÍA PROVOCANDO COMPRESIÓN Y EXPANSIÓN DE MOLÉCULAS DE AIRE.	20
ILUSTRACIÓN 3. MAGNITUDES DE UNA SEÑAL SINUSOIDAL.....	20
ILUSTRACIÓN 4. VISTA TRANSVERSAL DE LA APERTURA Y CIERRE DE LAS CUERDAS VOCALES.....	22
ILUSTRACIÓN 5. FORMA DE ONDA TEMPORAL DE UN SONIDO SONORO (VOCAL A).....	23
ILUSTRACIÓN 6. FORMA DE ONDA TEMPORAL DE UN SONIDO SORDO (CONSONANTE S).....	23
ILUSTRACIÓN 7. ESQUEMA DE FUNCIONAMIENTO DEL APARATO FONADOR HUMANO.....	25
ILUSTRACIÓN 8. ESTRUCTURA DEL SISTEMA PERIFÉRICO AUDITIVO HUMANO.....	27
ILUSTRACIÓN 9. ESQUEMA DE FUNCIONAMIENTO DEL HABLA HUMANA.....	28
ILUSTRACIÓN 10. ESQUEMA CONCEPTUAL DE UN SISTEMA DE RECONOCIMIENTO DE VOZ.....	34
ILUSTRACIÓN 11. ESQUEMA DE DIGITALIZACIÓN DE UNA SEÑAL.....	34
ILUSTRACIÓN 12. PROCESO DE MUESTREO DE UNA SEÑAL ANALÓGICA.....	35
ILUSTRACIÓN 13. PROCESO DE CUANTIFICACIÓN DE UNA SEÑAL ANALÓGICA.....	36
ILUSTRACIÓN 14. PROCESO DE PARAMETRIZACIÓN DE UNA SEÑAL DE VOZ.....	37
ILUSTRACIÓN 15. ENVENTANADO Y SOLAPAMIENTO DE UNA SEÑAL DE VOZ.....	37
ILUSTRACIÓN 16. REPRESENTACIÓN TEMPORAL Y FRECUENCIAL DE LAS VENTANAS MÁS UTILIZADAS.....	38
ILUSTRACIÓN 17. PROPIEDAD DE COMPACTACIÓN DE LA DCT FRENTE A LA FFT.....	40
ILUSTRACIÓN 18. MODELO DIGITAL DE PRODUCCIÓN DE VOZ.....	41
ILUSTRACIÓN 19. EJEMPLO DE CADENA DE MARKOV CON SEIS ESTADOS.....	43
ILUSTRACIÓN 20. CAMINO DE MAYOR PROBABILIDAD MEDIANTE EL ALGORITMO DE VITERBI.....	45
ILUSTRACIÓN 21. ESQUEMA DE EXTRACCIÓN DE CARACTERÍSTICAS EN EL PROCESO DE HABLA HUMANA.....	46
ILUSTRACIÓN 22. BANCO DE FILTROS LOGARÍTMICOS PARA LA OBTENCIÓN DE COEFICIENTES MFCC.....	46
ILUSTRACIÓN 23. PROCESO DE EXTRACCIÓN DE COEFICIENTES MFCC.....	47
ILUSTRACIÓN 24. ESQUEMA DE FUNCIONAMIENTO DEL SISTEMA DE SEGMENTACIÓN DE AUDIO ATVS-UAM.....	48
ILUSTRACIÓN 25. ESQUEMA DE FUNCIONAMIENTO DEL SISTEMA DE DIARIZACIÓN DE LOCUTORES ATVS-UAM.....	49
ILUSTRACIÓN 26. DISCRIMINACIÓN DE CARACTERÍSTICAS USANDO LA CLASIFICACIÓN SVM.....	53
ILUSTRACIÓN 27. ESQUEMA PRLM DE VERIFICACIÓN DE IDIOMA.....	53
ILUSTRACIÓN 28. EJEMPLO DE RESULTADOS DE LA APLICACIÓN DE LOS SISTEMAS DEL PRESENTE PROYECTO.....	56
ILUSTRACIÓN 29. EJEMPLO DE FORMATO DE LAS TRANSCRIPCIONES ORIGINALES.....	58
ILUSTRACIÓN 30. PROBABILIDADES DE TRANSICIÓN DEL HMM DE CINCO ESTADOS GENERADO PARA LA EVALUACIÓN DE ALBAYZIN 2010.....	58
ILUSTRACIÓN 31. DICCIONARIO FONÉTICO DE HABLA EN ESPAÑOL Y LISTADO DE TRIFONEMAS GENERADOS PARA LA EVALUACIÓN DE ALBAYZIN 2010.....	59
ILUSTRACIÓN 32. DICCIONARIO FONÉTICO DE HABLA EN ESPAÑOL E INGLÉS RESPECTIVAMENTE.....	60
ILUSTRACIÓN 33. IMAGEN DEL SISTEMA DE SEGMENTACIÓN DE AUDIO Y DIARIZACIÓN DE LOCUTORES.....	63
ILUSTRACIÓN 34. ADAPTACIÓN DEL HMM DE 5 ESTADOS A UN HMM DE 2 ESTADOS.....	64
ILUSTRACIÓN 35. EJEMPLO DE ALMACENAMIENTO DE DATOS EN ARCHIVO DE TEXTO PLANO.....	64
ILUSTRACIÓN 36. ESQUEMA DE PROCESAMIENTO DE AUDIO Y HERRAMIENTAS HTK.....	65
ILUSTRACIÓN 37. EJEMPLO DE FICHERO TRAS RECONOCIMIENTO <i>HVITE</i> CON EXTENSIÓN <i>'MLF'</i>	66
ILUSTRACIÓN 38. ESQUEMA VISUAL DE RECONOCIMIENTO, COMBINANDO RESULTADOS.....	68
ILUSTRACIÓN 39. SISTEMA DE IDENTIFICACIÓN DE IDIOMA.....	69
ILUSTRACIÓN 40. SECUENCIA DE EJECUCIÓN EN CADENA DE LOS SISTEMAS IMPLEMENTADOS.....	70
ILUSTRACIÓN 41. EJEMPLO DE PRESENTACIÓN DE RESULTADOS CON <i>'WAVESURFER'</i>	75

Índice de tablas

TABLA 1. CLASIFICACIÓN DE LOS FONEMAS DEL CASTELLANO.....	30
TABLA 2. CLASIFICACIÓN DE LOS FONEMAS VOCÁLICOS DEL CASTELLANO.....	30
TABLA 3. VENTANAS MÁS UTILIZADAS PARA ENVENTANADO.....	38
TABLA 4. VÍDEOS DE LA BASE DE DATOS CORPUS MA2VICMR.....	57
TABLA 5. ESTADOS DEL HMM DEL SISTEMA DE SEGMENTACIÓN DE AUDIO ATVS-UAM.....	63
TABLA 6. VALORES DE CONFIGURACIÓN DE <i>HVITE</i>	66
TABLA 7. TABLAS DE APLICACIÓN DE LAS FUNCIONES LÓGICAS AND Y OR.....	67
TABLA 8. EJEMPLO DE PRESENTACIÓN DE LOS RESULTADOS OBTENIDOS EN SEGMENTACIÓN Y RECONOCIMIENTO DE LOCUTORES.....	73
TABLA 9. EXTRACTO DE TRANSCRIPCIONES ORIGINAL Y SEGMENTADA DEL AUDIO ' <i>ANDALUCIA_ESP.WAV</i> '.	74
TABLA 10. SIGNIFICADO DE LAS ETIQUETAS DE ' <i>WAVESURFER</i> '.	75
TABLA 11. EJEMPLO DE PRESENTACIÓN DE LOS RESULTADOS OBTENIDOS EN IDENTIFICACIÓN DE IDIOMA.....	76
TABLA 12. DURACIÓN DE TIEMPOS REAL DE LOS AUDIOS MOSTRADOS EN LAS PRUEBAS.....	76
TABLA 13. IDIOMAS REALES DE LOS AUDIOS MOSTRADOS EN LAS PRUEBAS.....	76
TABLA 14. RESULTADOS DEL SISTEMA DE SEGMENTACIÓN DE AUDIO ATVS-UAM.....	77
TABLA 15. RESULTADOS DEL SISTEMA DE SEGMENTACIÓN DE AUDIO POR REC. FONÉTICO.....	78
TABLA 16. RESULTADOS DEL SISTEMA DE SEGMENTACIÓN DE AUDIO POR REC. FONÉTICO Y ADAPTACIÓN AL LOCUTOR.....	79
TABLA 17. RESULTADOS DE LA COMBINACIÓN MEDIANTE FUNCIÓN AND DE LOS SISTEMAS DE SEGMENTACIÓN DE AUDIO Y RECONOCIMIENTO DE LOCUTORES.....	80
TABLA 18. RESULTADOS DE LA COMBINACIÓN MEDIANTE FUNCIÓN OR DE LOS SISTEMAS DE SEGMENTACIÓN DE AUDIO Y RECONOCIMIENTO DE LOCUTORES.....	81
TABLA 19. RESULTADOS DE LA IDENTIFICACIÓN DE IDIOMA MEDIANTE HMMS DE 1 GAUSSIANA.....	83
TABLA 20. RESULTADOS DE LA IDENTIFICACIÓN DE IDIOMA MEDIANTE HMMS DE 2 GAUSSIANAS.....	84
TABLA 21. RESULTADOS DE LA IDENTIFICACIÓN DE IDIOMA MEDIANTE HMMS DE 3 GAUSSIANAS.....	85
TABLA 22. RESULTADOS DE LA IDENTIFICACIÓN DE IDIOMA MEDIANTE HMMS DE 4 GAUSSIANAS.....	86

1

Introducción

1.1. Motivación del proyecto.

Si se busca pocas décadas atrás en el tiempo, el reconocimiento de palabras de una grabación de voz, se realizaba por personal especializado en lingüística. Tras pasar cientos de horas escuchando varias locuciones de una misma persona, se llegaba a conclusiones habitualmente acertadas. Este proceso requería un gran esfuerzo tanto físico como mental, pues los especialistas eran pocos y este arte requería mucha responsabilidad, de ellos dependía muchas veces que un hombre fuese declarado inocente o culpable.

Desde hace algunos años, con la aparición de los ordenadores más o menos evolucionados, algoritmos de tratamiento de señales y el estudio del funcionamiento fisiológico del cuerpo humano para emitir voz, se comenzó a trabajar con estas señales de audio a nivel informático.

Con ello nació el procesamiento de voz orientado hacia aplicaciones en situaciones reales. Este proyecto es una aportación más a éste amplio campo, tratando de dar soluciones a problemas que existen actualmente.

Las grabaciones de voz no siempre son claras, el audio puede provenir de diferentes fuentes de sonido, con efectos propios del medio de por el que se transmite la señal como pueden ser el ruido o la diafonía, o grabaciones con música de fondo.

Todos estos hechos, han servido de motivación para la participación del grupo ATVS-UAM en varias competiciones a nivel mundial. Recientemente, la participación en la evaluación de Albayzin 2010, con el sistema descrito en "ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation" [1], permitió obtener buenos resultados en este campo y se puede decir que es el punto de partida de éste proyecto.

1.2. Objetivos.

El objetivo a largo plazo que se persigue por este proyecto es la mejora de la recuperación de información multimedia aplicada a videos de información turística, en los que se encuentran mezcladas voces de personas, a los que se les denominará locutores, con música de fondo. Para ello se ha diseñado, implementado y realizado pruebas de un sistema de segmentación de audio, encargado de separar el audio de los videos en segmentos de distinta naturaleza, voz, música, ruido, etc.

Tras obtener los segmentos de voz se ha aplicado en cadena el sistema de diarización, encargado de la identificación de los diferentes locutores de las grabaciones. Y en el último eslabón de esta cadena, se ha aplicado un sistema de identificación de idioma a los segmentos de voz.

Como objetivos más concretos para éste proyecto se planificó la aplicación y evaluación del sistema ATVS-UAM en el contexto de videos de información turística (un contexto distinto a la última evaluación en la que participó en ATVS, Albayzin 2010) y la realización de mejoras tendentes a aumentar la robustez del sistema frente a distintos entornos.

El análisis de factores (una técnica utilizada para reducir el número de variables que se encuentran en la voz, a factores más influyentes según describió Najim [6]) es en la actualidad la tecnología de última generación en el reconocimiento de idioma y locutores que mejor funciona. Un esquema de ello es el propuesto por Castaldo [2], el cual utiliza un conjunto vectores de voz de baja dimensión y con un alto solapamiento. De esta forma se puede crear un nuevo espacio de locutores aplicables a la diarización con excelentes resultados. Por otra parte, Najim y Kenny [3] mejora el esquema clásico de análisis de factores mediante el modelado de la voz y la variabilidad del canal.

Otros de los problemas afrontados durante el diseño del sistema de segmentación fue el uso de características que incluyen información temporal dependiendo de la estructura del habla, como son los coeficientes Cepstrales Delta Desplazados ó SDC (Shifted-Delta-Cepstral) [4] y el uso de MMI (Maximun Mutual Information). En los problemas de reconocimiento de locutor o de idioma de múltiples clases, los modelos GMM-MMI y HMM-MMI han sido los que mayor mejoras discriminativas han supuesto.

En cuanto a la base de datos utilizada, se ha elegido un conjunto de videos de información turística de diferentes series de documentales. Estos videos pertenecen al proyecto MA2VICMR [18], dedicado a la recuperación automática de información multimedia en la red.

Cabe destacar que los videos seleccionados presentan cualidades diversas en cuando a calidad de grabación, volumen de la música, diferentes locutores, ruidos, ecos, etc. por lo que se analizarán los resultados en función de éstos parámetros.

1.3. Organización de la memoria.

Capítulo 1. Introducción.

En el *capítulo 1* se presenta la motivación para el desarrollo del proyecto así como los objetivos que se pretenden conseguir durante la realización del mismo.

Capítulo 2. Estado del arte.

En el *capítulo 2* se habla del estado del arte de las tres principales líneas de trabajo, la segmentación de audio, la identificación de locutores (diarización) y la identificación de idioma.

Se entiende por segmentación de audio la separación de una grabación de voz, en zonas o fragmentos de voz y otras de no voz (silencios, música, ruido) siendo ésta la división más sencilla e intuitiva y la utilizada en el presente proyecto.

Cuando se habla de identificación de locutores, se hace referencia a la asignación de un segmento (de los que anteriormente se han reconocido como segmento de voz), a una persona de la cual se dispone de una base de datos de pronunciación de palabras propias, y por lo tanto la persona es conocida (aunque no en identidad, sí en características fonéticas).

Por último se hará una parada en la identificación de idioma, parte del proyecto que supone una mejora con respecto a lo inicialmente planificado. En este apartado se verán las diferentes aproximaciones que se han realizado en torno a la identificación automática del idioma de una locución.

Para este proyecto se ha limitado la cantidad de idiomas a dos, el español y el inglés, puesto que la base de datos de videos se encuentra en ambos idiomas. Para esta identificación se necesitarán adicionalmente las bases de datos propias de cada idioma, en las que se pueden diferenciar diccionarios de trifenemas más comúnmente utilizados,

Capítulo 3. Diseño y desarrollo del sistema.

En el *capítulo 3* se detalla el diseño y el desarrollo de los sistemas implicados. En un primer momento ha sido necesario preparar los datos a utilizar, pues proceden de videos grabados de varias naturalezas y es necesaria una estandarización. En el correspondiente capítulo se detallan las características de extracción del audio elegidas.

Los sistemas utilizados propiamente son:

- **Sistema de segmentación de audio y reconocimiento de locutores.**
Existen diversas formas de afrontar el problema. En este proyecto fin de carrera se verán dos aproximaciones:
 - o Aproximación mediante el Sistema de Segmentación de audio ATVS-UAM [1], basado en Modelos ocultos de Markov (HMMs) y que consiste en un sistema de separación de características mediante modelos de mezclas de gaussianas (GMMs) basado en cinco estados y la posterior obtención de la cadena de audio más probable mediante alineamiento con el algoritmo de Viterbi. Todo ello se detalla en apartados posteriores.

- Aproximación mediante reconocimiento fonético. Se entiende por reconocimiento fonético la diferenciación entre dos locutores por su manera de pronunciar las palabras y su pronunciación más o menos parecida a un conjunto de palabras estándar incluidas en un diccionario.
- **Sistema de identificación de idioma.**
Por último, para el sistema de identificación de idioma (tanto para español como para inglés), se ha empleado una técnica de reconocimiento fonético, que emplea el uso de un diccionario, una lista de fonemas y una gramática propia de cada idioma, para identificarlo y decidir cuál es el más aproximado a la locución probada.

Capítulo 4. Pruebas y resultados.

En el *capítulo 4* se muestra los resultados obtenidos tras la realización de las pruebas. En concreto se han realizado pruebas de:

- Segmentación de audio: comprobando si los segmentos reconocidos como voz encajan con los segmentos reales.
- Reconocimiento de locutores: observando si los segmentos de voz identificados con etiquetas de cada locutor, corresponden al propio locutor.
- Identificación de idioma: verificando el idioma correcto de cada segmento anteriormente identificado como voz.

Todos los datos obtenidos han podido ser probados gracias a que se dispone de las transcripciones manuales de todos los videos. En dichos archivos de etiquetado se detallan las transcripciones fonética y prosódica, el tiempo de inicio y fin de los segmentos de voz, palabras pronunciadas, idioma de habla, número de locutores participantes, etc. de cada video.

Capítulo 5. Conclusiones y trabajo futuro.

En el *capítulo 5* se darán las conclusiones obtenidas durante el desarrollo del presente proyecto así como unas pautas para posteriores trabajos futuros.

2

Estado del arte

2.1. Introducción

El reconocimiento automático de voz tiene como objetivo identificar rasgos o características propias de la persona mediante el análisis automático del habla humana. Esta tarea involucra diferentes campos de investigación como son la psicología, la medicina, las comunicaciones, la lingüística, el análisis de voz o el aprendizaje automático.

Es un campo en auge tras la creciente adaptación, de los sistemas de reconocimiento desarrollados por todo el mundo, a situaciones reales y útiles para el ser humano. Todo ello ha sido posible gracias a la continua mejora computacional de los sistemas informáticos.

En general, se puede entender que el reconocimiento de habla se basa en el establecimiento de una comunicación entre el hombre y una máquina (normalmente un ordenador con características especiales) para realizar algún tipo de tarea específica. Hoy en día este tipo de sistemas se encuentran en dispositivos móviles, tablets, móviles, ordenadores portátiles para el control por voz de los equipos. También se pueden encontrar en servicios telefónicos convencionales, realizando las labores de una centralita de atención automática o bien como método de validación en aplicaciones bancarias, hasta su uso por las redes de datos más extendidas internet como opción de dictado en traductores. Este alcance a situaciones de la vida real hace que estos sistemas tengan amplia aceptación por el usuario final, pues es una comunicación rápida y fácil, acercándose cada vez más a la forma natural de comunicación humana.

Pero esta labor no es sencilla, existen muchos factores que ponen límites al proceso de reconocimiento automático. Por ejemplo la variabilidad que se observa en la señal de voz, debido a situaciones físicas y psicológicas o diferencias existentes en pronunciación de una misma palabra requiere el uso de un sistema robusto. El vocabulario, la gramática y el entorno físico que rodea al locutor, son otros aspectos que no facilitan la labor perseguida.

Todo ello abre las puertas a un apasionante mundo de nuevas aplicaciones que marcará el futuro y porvenir de generaciones posteriores.

2.2. Sonido y procesamiento humano del habla

El habla constituye la forma más natural de comunicación entre las personas, de ahí el gran interés que tiene el desarrollo de sistemas informáticos capaces de procesar el habla y generarla de forma automática.

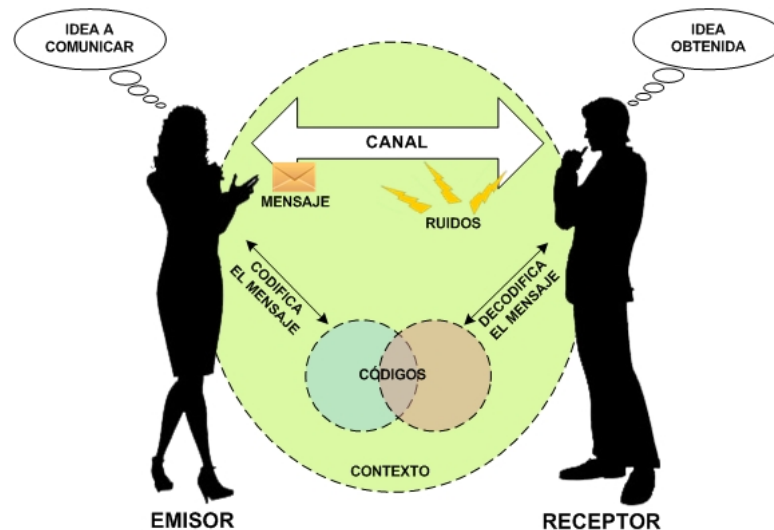


Ilustración 1. Modelo de comunicación oral.

El procesamiento del habla abarca un amplio abanico de métodos y técnicas que tienen como finalidad lograr que los ordenadores puedan comprender los mensajes pronunciados por los usuarios, y por otra, lograr que los usuarios puedan entender los mensajes generados por los ordenadores de forma oral.

Cada vez es más importante tener una interacción directa, clara y cómoda con todas las máquinas y ordenadores que se encuentran a nuestro alrededor y con los que los humanos se ven familiarizados desde edades muy tempranas. Los primeros sentidos que se desarrollan plenamente en los humanos son los que permiten la comunicación oral. Es por esto y por la posibilidad de acercar las máquinas al mundo de discapacitados, tanto físicos como motrices, que la comunicación oral con las máquinas ha cobrado una importancia vital en los últimos tiempos.

No obstante, si bien la voz es el medio de comunicación más usual, los humanos producen y perciben la misma voz con gran redundancia y de ella se extrae la información más relevante. Es por esto muy importante determinar cómo se produce y percibe la voz a la hora de realizar su tratamiento automático.

2.2.1. El sonido

Muchos matices que se aprenden del mundo que se levanta a nuestro alrededor, llega a través del sentido del oído. La capacidad de oír es importante no solamente para aprender del mundo, sino también para comunicarse con otros seres humanos. La voz humana es única en su habilidad de expresar ideas abstractas.

Un sonido es una onda de presión, formada por compresiones y expansiones del aire en dirección paralela a la aplicación de energía. Las compresiones son zonas donde las moléculas de aire han sido forzadas por la aplicación de energía, dando lugar a una mayor concentración de las mismas. Y las

expansiones son zonas donde la concentración de moléculas de aire es menor (zonas menos densas en la ilustración siguiente).

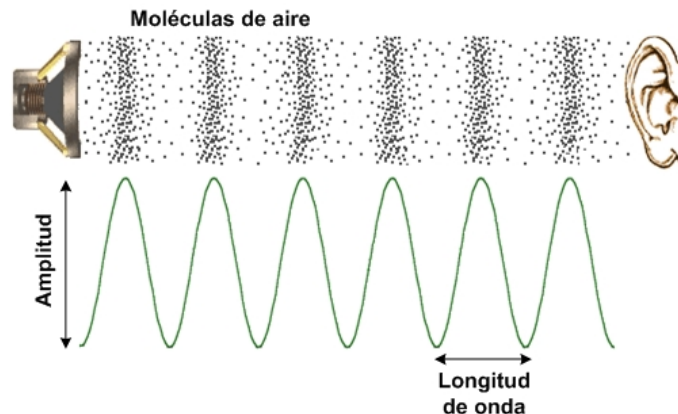


Ilustración 2. Aplicación de energía provocando compresión y expansión de moléculas de aire.

Cuando se hace referencia al sonido audible por el oído humano, se habla de una sensación percibida en el órgano del oído producida por la vibración que se propaga en un medio elástico en forma de ondas. El sonido audible para los seres humanos está formado por las oscilaciones de la presión del aire que el oído convierte en ondas mecánicas y finalmente, en impulsos nerviosos para que el cerebro pueda percibirlos y procesarlos.

El sonido se puede representar como una suma de curvas sinusoides con un factor de amplitud diferente que se puede caracterizar por las mismas magnitudes y unidades de medida que cualquier sinusoidal: longitud de onda (λ), frecuencia (f) o periodo (T) y amplitud (A). Cuando se considera la superposición de diferentes ondas es importante la fase que representa el retardo relativo en la posición de una onda con respecto a otra.

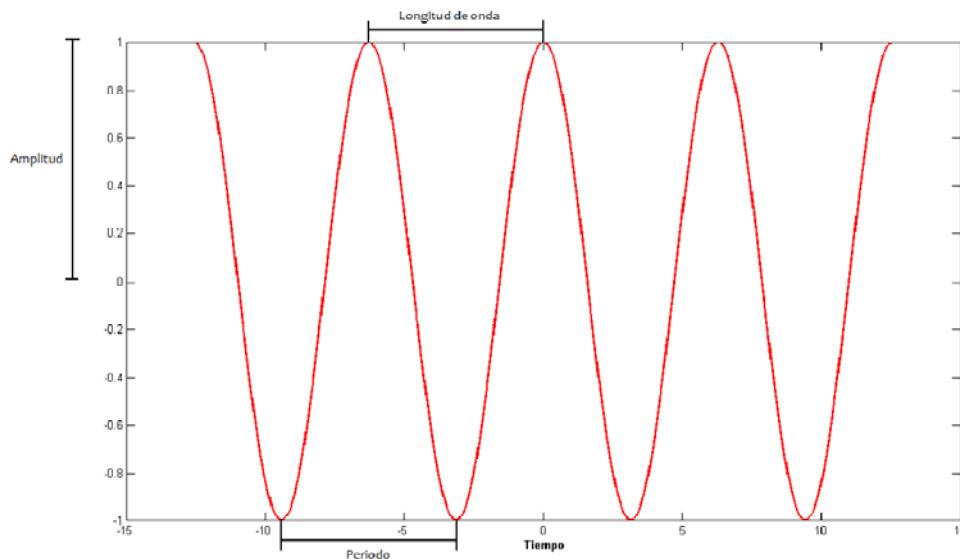


Ilustración 3. Magnitudes de una señal sinusoidal.

Sin embargo, un sonido complejo cualquiera no está caracterizado por los parámetros básicos (tiempo, amplitud, longitud de onda y período) ya que, en general, un sonido es una combinación de múltiples señales sinusoidales con diferentes valores en sus parámetros. La caracterización de un

sonido arbitrariamente complejo implica analizar tanto la energía transmitida como la distribución de dicha energía entre las diversas frecuencias, para ello resulta útil investigar:

- **Potencia acústica:** es la cantidad de energía por unidad de tiempo (potencia) emitida por una fuente determinada en forma de ondas sonoras. La potencia acústica viene determinada por la propia amplitud de la onda, pues cuanto mayor sea la amplitud de la onda, mayor es la cantidad de energía (potencia acústica) que genera.
- **Espectro de frecuencias:** que permite conocer en que frecuencias se transmite la mayor parte de la energía.

Los sonidos de los que consta el habla se pueden clasificar por la forma en que se produce el sonido básicamente en tres tipos:

- **Sonoros.** Son aquellos sonidos que hacen vibrar las cuerdas vocales. Esta vibración es cuasi-periódica y su espectro es muy rico en armónicos, que son múltiplos de la frecuencia de vibración de las cuerdas. A esta frecuencia de vibración de las cuerdas se le llama frecuencia fundamental y depende de la presión ejercida al pasar el aire por las cuerdas y de la tensión de éstas. En un hombre la frecuencia fundamental se encuentra en el rango 50-250 Hz, mientras en la mujer el rango es más amplio, encontrándose entre 100 y 500 Hz.
- **Fricativos.** En los sonidos fricativos se produce un estrechamiento del tracto vocal por el que se hace pasar el aire, lo que proporciona como resultado una excitación de ruido aleatorio.
- **Plosivos.** Estos sonidos se producen por la existencia de una obstrucción temporal al paso del aire. El sonido se produce al abrirse la obstrucción temporal produciéndose una liberación brusca de energía en forma de una pequeña explosión.

2.2.2. Producción del habla

El sistema de producción del habla no forma parte estricta del sistema sensorial humano, pero su importancia es tal indudable. La comunicación humana surgió para dar respuesta a los instintos de nuestros antepasados, y transmitir sentimientos, impresiones y emociones en la lucha por la supervivencia.

Para determinar las operaciones de un sistema automático de reconocimiento de voz y hablante, es fundamental conocer y determinar los mecanismos que han producido un mensaje hablado, para a continuación, poder reproducirlos automáticamente. Es por ello que se van a repasar algunos conceptos fundamentales y básicos en el mecanismo de producción del habla, tanto en el órgano físico que soporta dichos mecanismos, como la producción propia del mensaje.

El habla, como señal acústica, se produce a partir de las ondas de presión que salen de la boca y las fosas nasales del emisor. El proceso comienza con la generación de la energía suficiente (flujo de aire) en los pulmones, la modificación de ese flujo de aire en las cuerdas vocales, y su posterior perturbación por algunas constricciones y configuraciones de los órganos superiores.

Así, en el proceso fonador intervienen distintos órganos a lo largo del llamado tracto vocal, zona comprendida entre las cuerdas vocales y las aberturas finales: los labios y las fosas nasales.

El conjunto de órganos que intervienen en la fonación se puede dividir en tres grupos bastante bien delimitados: cavidades infraglóticas, laríngea y supraglóticas.

2.2.2.1. Cavidades infraglólicas

Las cavidades infraglólicas (sistema sub-glotal) u órgano respiratorio, consta de los órganos propios de la respiración (pulmones, bronquios y tráquea), que son la fuente de energía para todo el proceso de producción de voz.

En el proceso de inspiración, los pulmones toman aire, bajando el diafragma y agrandando la cavidad torácica. En el momento de la fonación, la espiración, provocada por la contracción de los músculos intercostales y del diafragma, aporta la energía necesaria para generar la onda de presión acústica que atravesara los órganos fonadores superiores.

2.2.2.2. Cavidad laríngea

La cavidad laríngea u órgano fonador, es la responsable de modificar el flujo de aire generado por los pulmones y convertirlo en una señal capaz de excitar adecuadamente las posibles configuraciones de las cavidades supraglólicas.

El ultimo cartílago de la tráquea, el cricoides, forma la base de la laringe, cuyo principal órgano son las cuerdas vocales que son dos pares de repliegues compuestos de ligamentos y músculos. El par inferior son las llamadas cuerdas vocales verdaderas, que pueden juntarse o separarse mediante la acción de los músculos circo-aritenoides lateral y posterior, y que están protegidas en su parte anterior por el cartílago tiroides, el más importante de la laringe, abierto por su parte posterior. Finalmente, la parte superior de la laringe esta unida al hueso hioides.

En la figura siguiente se muestra una vista transversal simplificada de la zona en la que se encuentran las cuerdas vocales, en sus posiciones extremas: abiertas y cerradas. A la apertura que queda entre las cuerdas vocales se le denomina glotis.

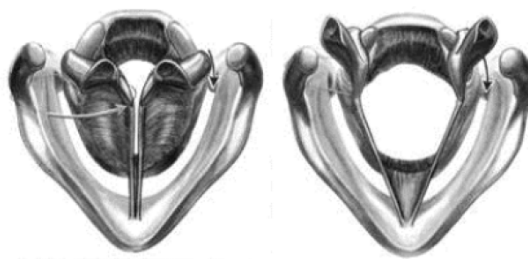


Ilustración 4. Vista transversal de la apertura y cierre de las cuerdas vocales.

La cavidad laríngea termina en la epiglotis, un cartílago en forma de cuchara que permite cerrar la apertura de la laringe en el acto de la deglución.

La distinción fundamental entre los sonidos se basa en su característica de sonoridad. En los sonidos sonoros, incluyendo las vocales, se observa un patrón regular tanto en su estructura temporal (ilustración 5) como en su estructura frecuencial, patrón del que carecen los aleatorios sonidos sordos (ilustración 6).

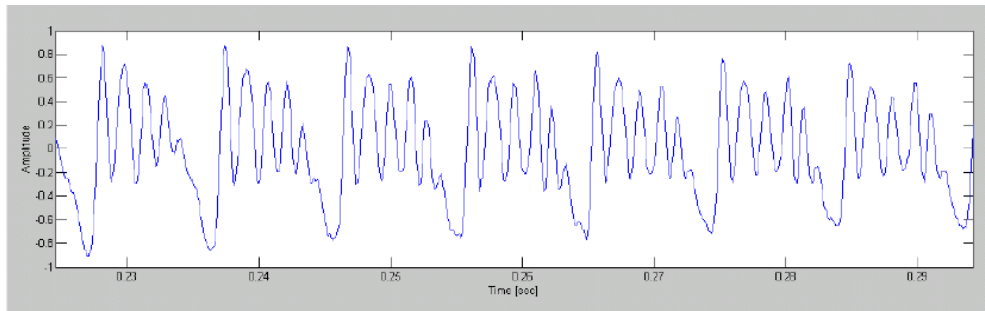


Ilustración 5. Forma de onda temporal de un sonido sonoro (vocal a).

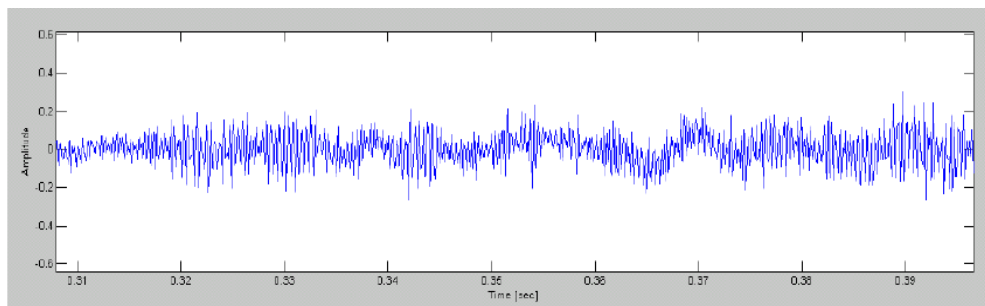


Ilustración 6. Forma de onda temporal de un sonido sordo (consonante s).

La cualidad de sonoridad de los sonidos sonoros se produce por la acción vibradora de las cuerdas vocales. El mecanismo de vibración se produce de la siguiente forma: si se supone que inicialmente las cuerdas vocales están juntas, la presión subglotal se incrementa lo suficiente para forzar a las cuerdas vocales a separarse. Al separarse, el aire pasa a través de ellas y la presión subglotal disminuye, momento en el que la fuerza de los músculos hace que las cuerdas vocales vuelvan a juntarse. Cuando las cuerdas vocales se juntan, el flujo de aire disminuye y la presión subglotal aumenta de nuevo, con lo que se vuelve a reproducir el ciclo y esta vibración de las cuerdas vocales produce pulsos casi periódicos de aire que excitan el sistema por encima de la laringe.

A esta frecuencia de vibración se la denomina frecuencia fundamental, y sus valores típicos oscilan entre los 60 Hz para un hombre voluminoso, y los 300 Hz para una mujer o un niño. La señal generada en las cuerdas vocales puede variar en frecuencia e intensidad según varíe la masa, la longitud y la tensión de las mismas.

2.2.2.3. Cavidades supraglóticas.

Las cavidades supraglóticas están constituidas por la faringe, la cavidad nasal y la cavidad bucal. Su misión fundamental de cara a la fonación es perturbar adecuadamente el flujo de aire procedente de la laringe, para dar lugar finalmente a la señal acústica generada a la salida de la nariz y la boca.

La faringe es una cavidad en forma tubular que une la laringe con las cavidades bucal y nasal, y que suele dividirse en tres partes: faringe laríngea, faringe bucal (boca) y faringe nasal, las dos últimas separadas por el velo del paladar. El volumen de la faringe laríngea puede ser modificado por los movimientos de la laringe, la lengua y la epiglotis mientras que el volumen de la faringe bucal se modifica por el movimiento de la lengua.

La faringe nasal y las restantes cavidades nasales forman, desde el punto de vista de su acción sobre el flujo de aire procedente de la faringe, un resonador que puede o no conectarse al resonador bucal

mediante la acción del velo del paladar. Según el resonador nasal este o no conectado, el sonido será nasal u oral, respectivamente.

Si se hace una descripción de la cavidad bucal, se deben de señalar las siguientes partes:

- Los labios.
- Los dientes.
- La zona alveolar, entre los dientes y el paladar duro.
- El paladar, en el que a su vez, y de forma simplificada, se distinguirán el paladar duro y el paladar blando o velo.

La raíz de la lengua forma la pared frontal de la faringe laríngea, y sus movimientos le permiten modificar la sección de la cavidad bucal (movimiento vertical), adelantar o retrasar su posición frente a la de reposo (movimiento horizontal), así como poner en contacto su ápice o la parte trasera con alguna zona del paladar.

El movimiento de los labios también interviene en la articulación, pudiendo ser de apertura o cierre y de protuberancia, alargando en este último caso la cavidad bucal.

De los movimientos de los órganos supraglotales surgen los distintos modos de articulación de los posibles sonidos emitidos por un locutor. En la mayor parte de los casos es un órgano el que se mueve (activo) y otro contra el que se efectúa la articulación (pasivo). Según la pareja de órganos activos/pasivos que se tengan, habrá una serie de posibles combinaciones de articulaciones.

A continuación se muestra un esquema completo de los órganos que forman parte del sistema de habla humano.

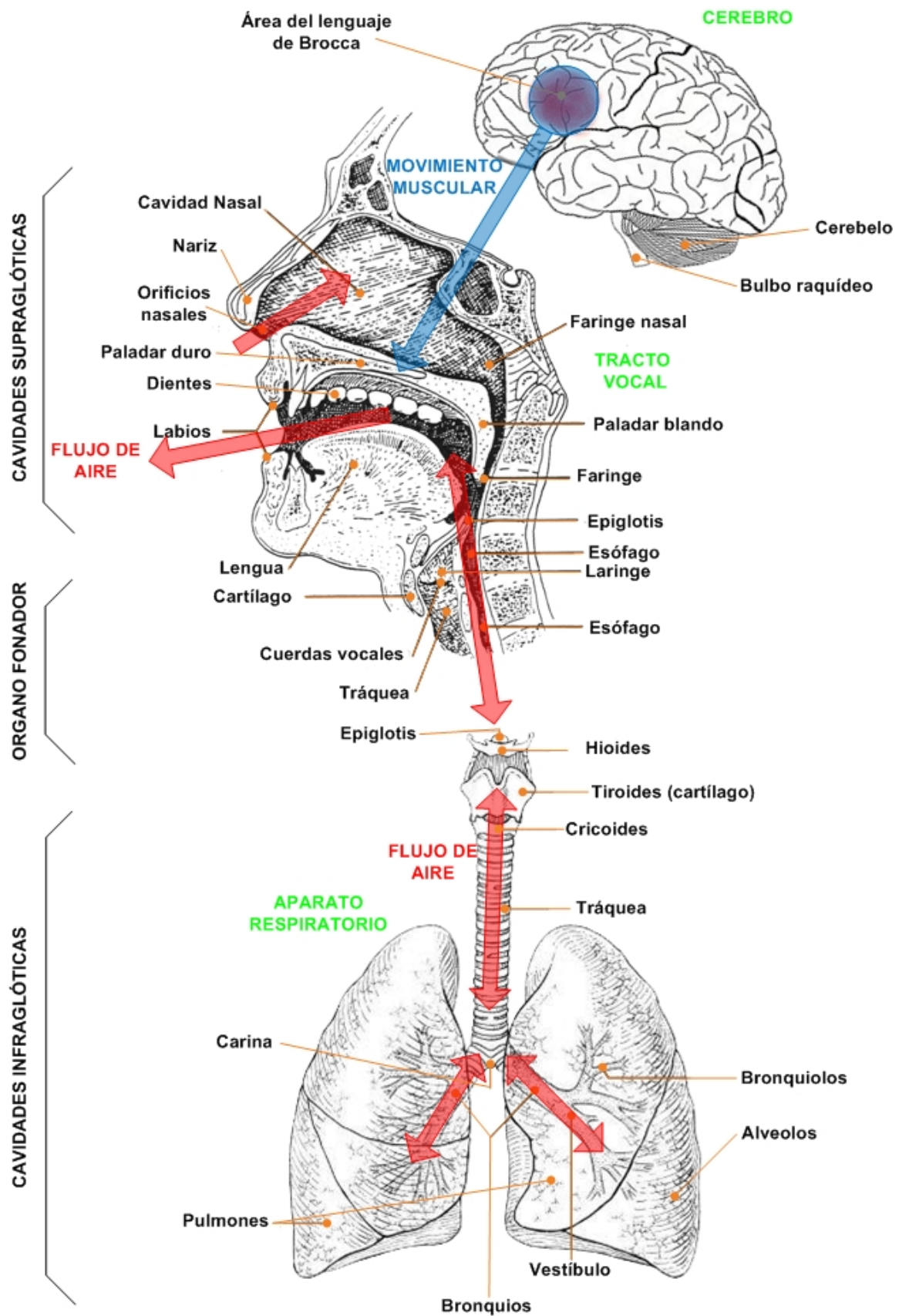


Ilustración 7. Esquema de funcionamiento del aparato fonador humano.

2.3. Percepción del sonido

La capacidad de comprender el lenguaje oral se deriva del funcionamiento de un conjunto muy complejo de procesos perceptivos, cognitivos y lingüísticos que permiten al oyente recuperar el significado de un enunciado cuando lo escucha.

La percepción puede verse como un proceso que une la onda acústica y su representación conceptual por medio de una serie de niveles:

- Estructura acústica.
- Habla.
- Estructura fonética.
- Fonología.
- Estructura superficial (información fonética).
- Sintaxis.
- Estructura profunda (información sintáctica).
- Semántica.
- Representación conceptual.

Además de las diferencias en la señal, también hay diferencias marcadas en el procesado de los sonidos de habla y los sonidos de no habla. Para los sonidos de habla: responde ante ellos como entidades lingüísticas más que como acontecimientos auditivos. El oyente aprovecha su background lingüístico para categorizar y etiquetar las señales de habla.

El problema fundamental es determinar como el estímulo acústico, que varía de manera continua, se convierte en una secuencia de unidades lingüísticas discretas de forma que sea posible recuperar el mensaje. Aunque la señal de habla sea de calidad pobre o distorsionada, el proceso de percepción se realiza perfectamente. Esto se debe a que el habla es una señal altamente estructurada y redundante de modo que las distorsiones no afectan a la inteligibilidad. La percepción también es posible porque el oyente tiene dos tipos de información disponibles, el contexto del habla (conocimiento pragmático) y el conocimiento de la lengua (sintaxis, semántica y fonología).

El mecanismo físico de la percepción del habla, al igual que la audición, se realiza por medio de dos órganos fundamentales, el Sistema auditivo periférico y el Sistema nervioso central auditivo.

El Sistema auditivo periférico es lo que vulgarmente se llama oído. En la figura siguiente pueden observarse las 4 partes en las que se divide el sistema auditivo: oído externo, oído medio, oído interno y el Sistema nervioso central auditivo.

Los modos de funcionamiento son los siguientes:

- **Oído externo:** funciona por vibración del aire. Canaliza la energía acústica y consiste de la parte externa visible y el canal auditivo externo, de aproximadamente 2.5 cm, a través del cual viaja el sonido.
- **Oído medio:** funciona por movimiento mecánico de los huesecillos. Transforma la energía acústica en energía mecánica, transmitiéndola hasta el oído interno.
- **Oído interno:** primero el funcionamiento mecánico, por el movimiento del estribo, luego hidrodinámico por el movimiento de los líquidos interiores a la cóclea y finalmente

electroquímico. Aquí se realiza la definitiva transformación de la energía mecánica en impulsos eléctricos.

- **Sistema nervioso central auditivo:** el funcionamiento es electroquímico, el movimiento de las células ciliadas provocan una reacción química que a su vez genera un impulso eléctrico.

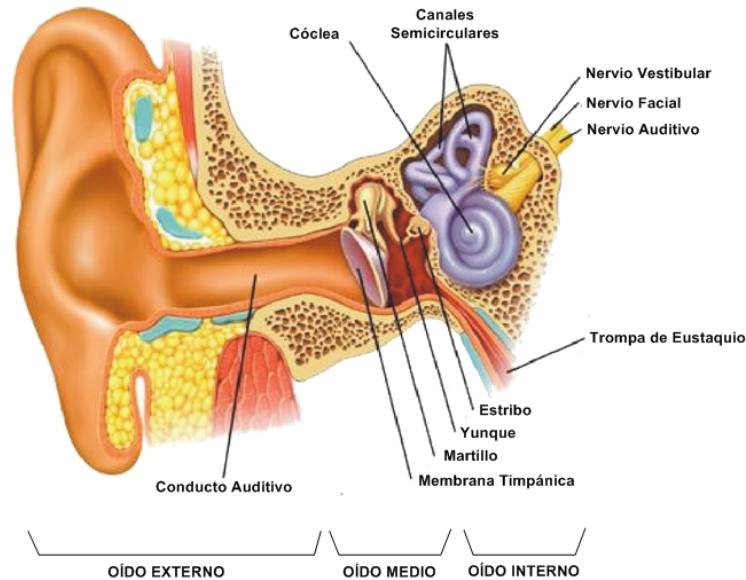


Ilustración 8. Estructura del sistema periférico auditivo humano.

Cuando el sonido llega al oído, las ondas sonoras son recogidas por el pabellón auricular (o aurícula). El pabellón auricular, por su forma helicoidal, funciona como una especie de "embudo" que ayuda a dirigir el sonido hacia el interior del oído. Sin la existencia del pabellón auricular, los frentes de onda llegarían de forma perpendicularmente y el proceso de audición resultaría ineficaz y gran parte del sonido se perdería: reflexión y difracción.

Hay que tener en cuenta que el pabellón auricular humano es mucho menos direccional que el de otros animales como los perros, que poseen un control voluntario de su orientación.

Una vez que ha sido recogido el sonido, las vibraciones provocadas por la variación de presión del aire cruzan el canal auditivo externo y llegan a la membrana del tímpano, ya en el oído medio.

El conducto auditivo actúa como una etapa de potencia natural que amplifica automáticamente los sonidos más bajos que proceden del exterior.

En el oído medio, se produce la transducción, es decir, la transformación la energía acústica en energía mecánica. En este sentido, el oído medio es un transductor mecánico-acústico.

La presión de las ondas sonoras hace que el tímpano vibre empujando a los ósculos que, a su vez, transmiten el movimiento del tímpano al oído interno. Cada ósculo empuja a su adyacente y finalmente a través de la ventana oval. Es un proceso mecánico, el pie del estribo empuja a la ventana oval, ya en el oído interno. Esta fuerza que empuja a la ventana oval es unas 20 veces mayor que la que empujaba a la membrana del tímpano, lo que se debe a la diferencia de tamaño entre ambas. Esta presión ejercida sobre la ventana oval, penetra en el interior de la cóclea, la cual se comunica directamente con el nervio auditivo, conduciendo una representación del sonido al cerebro. La cóclea es un tubo en forma de espiral (de 3.5 cm aproximadamente). La espiral está dividida longitudinalmente por la membrana basilar en dos cámaras que contienen líquido linfático.

La cóclea puede ser aproximada como un banco de filtros. Los filtros correspondientes al extremo más próximo a la ventana oval y al tímpano responden a las altas frecuencias, ya que la membrana es rígida y ligera. Por el contrario, en el extremo más distante, la membrana basilar es pesada y suave, por lo que los filtros correspondientes responden a las bajas frecuencias. Por ello los investigadores emprenden trabajos psicoacústicos experimentales para obtener las escalas de frecuencias que modelen la respuesta natural del sistema de percepción humano.

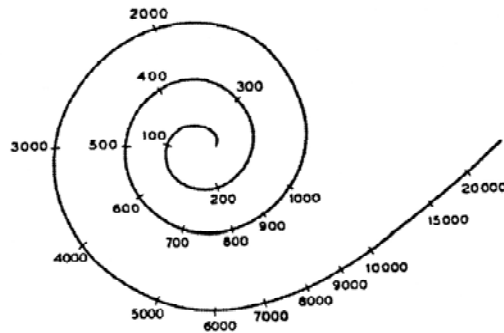


Ilustración 9. Esquema de funcionamiento del habla humana.

El comportamiento de la cóclea como analizador en frecuencia puede resumirse en dos características:

- La componente del espectro de la señal sonora es procesada por más de un receptor auditivo.
- Análogamente, cada receptor auditivo procesa diversas componentes del espectro de la señal.

La forma del patrón de excitación provocado por un tono puro ilustra bien estas dos características, e indica que la selectividad en frecuencia del sistema auditivo no es infinita. Una característica fundamental del sistema auditivo humano es su capacidad de resolución de frecuencia e intensidad. En este aspecto, es fundamental el concepto de banda crítica. Una forma de entender el funcionamiento del sistema auditivo es suponer que contienen una serie o banco de filtros paso banda solapados conocidos como filtros auditivos.

El sistema de audición lleva a cabo un análisis espectral de sonidos dentro de sus componentes de frecuencia. La cóclea actúa como si estuviese compuesta de filtros superpuestos con un ancho de banda igual al ancho de banda crítico. Una inquietud que surge de inmediato es preguntarse cuantas bandas críticas existen en el sistema auditivo y cuál es la frecuencia central de cada una.

Con objeto de aproximarse a la sensibilidad del oído humano, que no tiene una respuesta lineal, existen diferentes escalas. Existe una escala de medición de las bandas críticas llamada escala de Bark, que tiene un rango del 1 al 24 y corresponde a las primeras veinticuatro bandas críticas del sistema auditivo. Esta escala tiene relación con la escala de frecuencias Mel, que será explicada más adelante.

La escala Mel ha sido ampliamente utilizada en modernos sistemas de reconocimiento de habla y puede ser aproximada en función de la frecuencia lineal como:

$$\tilde{f} = 2595 \log_{10} \left(1 + \frac{f}{700} \right)$$

Donde f representa la frecuencia en escala lineal y \tilde{f} la frecuencia en escala Mel.

2.4. Reconocimiento fonético

Los sonidos del habla pueden ser estudiados desde diferentes puntos de vista, articulatorio, acústico, fonético y perceptual. En esta sección se describen desde el punto de vista fonético, acústico y articulatorio, es decir, se analizará las relaciones de las características lingüísticas de los sonidos en posiciones y movimientos de los órganos fonatorios, así como la relación entre los fonemas y sus realizaciones acústicas interpretando la señal de voz como la salida del proceso de producción.

2.4.1 Fono y fonema

Antes de abordar propiamente el estudio de la fonética y la fonología, es conveniente definir primero los términos lengua y habla:

- La **lengua** es un modelo general y constante que existe en la conciencia de todos los miembros de una comunidad lingüística determinada, constituyendo el sistema de comunicación verbal de la misma. Es abstracto y supraindividual (compuesto por reglas).
- El **habla** es la realización concreta de la lengua (en un momento y lugar determinados) por parte de cada uno de los miembros de esa comunidad lingüística (realizaciones).

El habla puede verse como una secuencia de unidades básicas de sonidos o fonemas. Los fonemas son unidades teóricas, postuladas para estudiar el nivel fonético - fonológico de una lengua humana. Son unidades lingüísticas abstractas y no pueden observarse directamente en la señal de voz. Un mismo fonema se aplica a muchos sonidos ligeramente diferentes llamados realizaciones del fonema o alófonos.

Desde un punto de vista estructural, el fonema pertenece a la lengua, mientras que el sonido pertenece al habla. La palabra <casa>, por ejemplo, consta de cuatro fonemas (/k/, /a/, /s/, /a/). A esta misma palabra también corresponden en el habla, acto concreto, cuatro sonidos, a los que la fonología denominara alófonos, y estos últimos pueden variar según el sujeto que lo pronuncie. La distinción fundamental de los conceptos fonema y alófono está en que el primero es una huella psíquica de la neutralización de los segundos que se efectúan en el habla.

Los fonemas no son sonidos con entidad física, sino abstracciones mentales o abstracciones formales de los sonidos del habla. Entre los criterios para decidir que constituye o no un fonema se requiere que exista una función distintiva: son sonidos del habla que permiten distinguir palabras en una lengua.

Así, los sonidos /p/ y /b/ son fonemas del español porque existen palabras como /pata/ y /bata/ que tienen significado distinto y su pronunciación solo difiere en relación con esos dos sonidos. De esta forma, se puede decir que un fonema es una unidad fonológica diferenciadora, indivisible y abstracta.

- **Diferenciadora:** cada fonema se delimita dentro del sistema por las cualidades que se distinguen de los demás y es portador de una intención significativa especial. Por ejemplo, /k-o-t-a/ y /b-o-t-a/ son dos palabras que se distinguen semánticamente debido a que /k/ se opone a /b/ por la sonoridad.
- **Indivisible:** no se puede descomponer en unidades menores. Por ejemplo, la sílaba o el grupo fónico sí pueden fraccionarse. Un análisis pormenorizado del fonema revela que está compuesto por un haz de diversos elementos fónicos llamados rasgos distintivos, cuya combinación forma el inventario de fonemas. El inventario de rasgos distintivos es asimismo limitado y viene a constituir una especie de tercera articulación del lenguaje.

- **Abstracta:** no son sonidos, sino modelos o tipos ideales de sonidos. La distinción entre sonido y fonema ha sido un gran logro en los últimos tiempos en la lingüística.

Se pueden clasificar los fonemas atendiendo a dos criterios: modo de articulación y punto de articulación. En el castellano se definen 24 fonemas que se clasifican en la siguiente tabla (donde SN significa sonido sonoro y SR, sonido sordo):

Clasificación de los fonemas del castellano		Punto de Articulación															
		Abierto		Labiales				Dentales		Alveolares		Palatales		Velares		Glatales	
				Bilabiales		Labiodentales											
		SN	SR	SN	SR	SN	SR	SN	SR	SN	SR	SN	SR	SN	SR	SN	SR
Modo de Articulación	Plosivas			b	p					d	t			g	k		
	Nasales			m						n			ɲ				
	Laterales									l			ll				
	Fricativas						f			s	y			x			
	Vibración simple									r							
	Vibración compuesta									r							
	Africadas												c				
	Vocales	a											e,i		o,u		
	Semivocales			w									j				

Tabla 1. Clasificación de los fonemas del castellano.

Las vocales en castellano y en la mayoría de idiomas no se suelen clasificar de la manera anterior, sino que responden a una clasificación más sencilla atendiendo a la posición de la lengua (anterior, media o posterior) y a la abertura de la boca (cerrada, medio cerrada o abierta), tal y como se ilustra en la siguiente tabla.

Clasificación de los fonemas vocálicos del castellano		Posición de la lengua		
		Anterior	Central	Posterior
Abertura de la boca	Cerrada	i		u
	Medio Cerrada	e		o
	Abierta		a	

Tabla 2. Clasificación de los fonemas vocálicos del castellano.

Un sonido o fono se caracteriza por una serie de rasgos fonéticos y articulatorios, el número de dichos rasgos y la identificación de los mismos es tarea de la fonética. Un fono es cualquiera de las posibles realizaciones acústicas de un fonema.

Dada la distinción entre fonema y fono, existe otra forma de concebir un fonema como una especificación incompleta de rasgos fonéticos. Esta relación es de hecho equivalente a la del fonema como conjunto de fonos: el fonema sería el conjunto de rasgos fonéticos comunes a todos los fonos que forman la clase de equivalencia del fonema.

Fijado un conjunto de rasgos fonéticos se pueden definir los sonidos de la lengua. En principio no hay límite a lo fina que pueda ser la distinción que establecen estos rasgos. Potencialmente la lista de sonidos puede hacerse tan grande como se quiera si se incluyen más y más rasgos, sin embargo, el número de fonemas es un asunto diferente, puesto que muchos de los anteriores sonidos serán equivalentes desde el punto de vista lingüístico.

2.4.2 Modelos fonéticos

Los fonemas, tal y como se ha citado en la sección anterior, representan un nivel superior de fragmentación de palabras. La modificación de un fonema puede cambiarle el sentido a la misma. Hay que distinguirlo de alófono, que es cada una de las pronunciaciones reales del modelo ideal que representa el fonema.

En función del grado de resolución que se quiera que presenten las unidades fonéticas se pueden obtener modelos más o menos específicos. La manera en que se hace la división en unidades fonéticas depende del contexto donde el alófono se localice:

- **Mono fonemas:** Son unidades totalmente libres de contexto. Un mono fonema tiene en cuenta todas las posibles realizaciones de un fonema independientemente de sus vecinos.
- **Bifonemas:** Son unidades que dependen solo de uno de sus contextos, ya sea este el derecho (bifonema derecho) o el izquierdo (bifonema izquierdo).
- **Trifonemas:** Son unidades que dependen de ambos contextos a la vez.
- **Trifonemas generalizados:** Dado que el número de unidades va aumentando al ir considerando más detenidamente la posición de los fonemas en su contexto, puede llegar a ser tan elevado que su entrenamiento no fuera posible. Surge el Trifonema generalizado como un primer nivel de compartición, en el que varios trifonemas cercanos se agrupan para reducir el número de modelos y que el entrenamiento de estos sea mejor.

Así, un reconocedor necesita disponer de un conjunto de unidades que permita construir cualquier palabra o frase a partir de su concatenación. Los fonemas representan este conjunto completo y reducido de unidades a partir de las cuales se puede generar cualquier palabra. Estas unidades pueden modelarse con mayor o menor resolución en función del contexto a considerar. Es necesario especificar para un determinado idioma, en este caso para el castellano e inglés, como son estas unidades. Además, las diferentes alternativas de entrenamiento en función del acceso a las bases de datos será un factor determinante a la hora de establecer el conjunto de unidades a entrenar.

La creación de modelos acústicos, para su posterior uso en reconocimiento de habla, se crean en dos fases: la primera es la extracción de características de la señal de voz, y la segunda es usar esas características para identificar los fonemas:

- **Extracción de las características del sonido del habla.** Por semejanza con el funcionamiento del sistema humano, la extracción de esas características, que se denominarán parámetros, se realiza en el dominio de la frecuencia. Asignación de los parámetros extraídos a las representaciones discretas de nuestro diseño (fonemas, trifonemas, palabras...) correspondientes, con el objetivo de crear un modelo para cada una de las representaciones discretas que las identifique. Tanto las técnicas para la extracción de parámetros como los distintos modelos, son definidos en las siguientes secciones.
- **Entrenamiento y reconocimiento de modelos para cada fonema a identificar.** A partir de la extracción de parámetros se construirán una serie de modelos estadísticos con los cuales se identificarán, con cierta probabilidad, fonemas en otras locuciones. Por ello, se mide la

distancia entre el modelo (conjunto de parámetros que constituye el modelo) y los parámetros de la pronunciación a reconocer. Hay varias técnicas para realizar el proceso:

- HMM (Hidden Markov Model): las aproximaciones estadísticas toman como referencia el modelo estocástico de los datos. Se basa en la creación de modelos de fonemas en estados. Hasta la fecha este método es el que mejores resultados proporciona y el más utilizado.
- DTW (Alineamiento Temporal Dinámico): consiste en alinear de forma temporal los parámetros del archivo de test y los parámetros de los modelos, obteniendo la función que alinea a ambos, eligiendo la función de menor coste posible para dicha adaptación.
- VQ (Cuantificación vectorial): consiste en representar las características de los fonemas como un espacio vectorial de dimensión el número de parámetros, de forma que al fonema a reconocer se le asigna el vector cuya distancia a él sea mínima. Por tanto, los fonemas quedaran representados por unos vectores determinados (centroides) de forma que todos los puntos que caigan en una zona determinada se asignaran a dicho vector.

2.5. Reconocimiento automático de audio

El objetivo del reconocimiento automático de una señal de audio, normalmente señales de voz humana, es imitar el proceso de reconocimiento que lleva a cabo el receptor en la comunicación oral. Hay varios niveles de reconocimiento en dicho proceso humano, y los diferentes sistemas de reconocimiento automático implementan todos, algunos o solo los más básicos dependiendo de cuáles sean su aplicación y complejidad. Se pueden distinguir ocho niveles de reconocimiento en orden de ascendente complejidad:

- **NIVEL ACÚSTICO**: la señal acústica analógica que ha enviado el emisor es recibida y traducida a un conjunto de rasgos relevantes no redundantes. En la comunicación oral, este reconocimiento se hace en el oído. Hay cuatro operaciones incluidas en este nivel:
 - *Parametrización*: la señal analógica se transforma en una señal numérica que pueda ser tratada por una máquina digital en la que se hace el reconocimiento. Hay varios métodos de parametrización, en el dominio del tiempo y de la frecuencia, que dan lugar a diferentes parámetros de caracterización.
 - *Segmentación*: determina como separar la señal analógica continua en una cadena de sonidos cuya sucesión es la señal en el tiempo. Se lleva a cabo con métodos basados en las curvas de variación de la energía o de variabilidad de la señal.
 - *Extracción de la información relevante*: se busca retener solo aquellos datos que proporcionan información útil para el reconocimiento como pueden ser los espectros de los instantes de mayor estabilidad o de los instantes de transición.
 - *Información relativa a la prosodia*: estudia la variación del armónico fundamental de la voz, variación de la intensidad, y el ritmo.
- **NIVEL FONÉTICO**: la secuencia de información relevante obtenida en el nivel acústico es traducida a una secuencia de fonemas.
- **NIVEL FONOLÓGICO**: los fonemas de la lengua que hacen que el contenido fonético de las palabras se modifique en una articulación rápida o por una sucesión de términos léxicos son analizados. Las variedades dialectales son también tratadas.
- **NIVEL LÉXICO**: se identifican las palabras de la lengua en la que se produce la comunicación.
- **NIVEL SINTÁCTICO**: se detectan las reglas gramaticales que permiten describir y analizar el lenguaje, y que relacionan las palabras reconocidas a nivel léxico.
- **NIVEL SEMÁNTICO**: analiza el sentido de las palabras, buscando la comprensión del mensaje y eliminando las interpretaciones que no tengan sentido. Es el nivel de conocimiento de las palabras que da un diccionario de la lengua.
- **NIVEL PRAGMÁTICO**: estudia el sentido del mensaje recibido teniendo en cuenta el contexto de su aplicación. Reconoce la información que viene determinada por la situación en la que se produce la comunicación.
- **NIVEL PROSÓDICO**: interviene de manera paralela al resto de niveles, sin formar parte de una estructura piramidal como los demás. Este nivel detecta la información que el mensaje comunica mediante los modos de pronunciación: palabras pronunciadas con cierto nivel de insistencia para ponerlas de relieve, fronteras entre grupos de palabras, naturaleza interrogativa o declarativa de una frase, etc.

Independientemente del nivel implementado en los diferentes sistemas de reconocimiento automático de habla, conceptualmente hablando son necesarios tres principales bloques de manipulación de datos. Dichos bloques son:

- **Bloque con datos de entrenamiento.** Es la información con la cual el sistema se entrena. Es un conjunto de datos que debe ser lo suficientemente significativo, equilibrado y amplio como para que el reconocedor aprenda a reconocer ese vocabulario.
- **Bloque de parametrización.** Las entradas de señal de voz pasaran por una etapa en la que se extraen sus características más representativas de ser clasificadas.
- **Bloque de reconocimiento.** La clasificación de los parámetros se realiza usando los datos de entrada y las referencias con las que cuenta el sistema: el aprendizaje de los datos de entrenamiento, y las referencias acústicas, léxicas y de lenguaje.

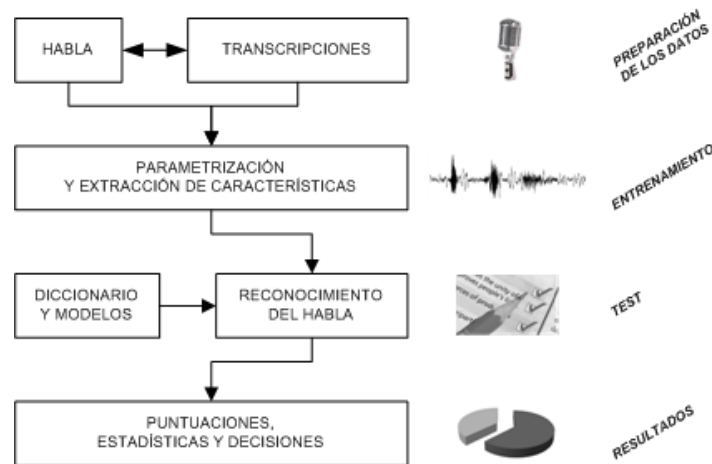


Ilustración 10. Esquema conceptual de un sistema de reconocimiento de voz.

2.5.1 Procesamiento de señales digitales de audio

El procesamiento digital de señales es una técnica que convierte señales procedentes de fuentes del mundo real (usualmente en forma analógica), en datos digitales que luego pueden ser analizados. Este análisis es realizado en forma digital, pues una vez que una señal ha sido reducida a valores numéricos discretos, sus componentes pueden ser aisladas, analizadas y reordenadas más fácilmente que en su primitiva forma analógica.

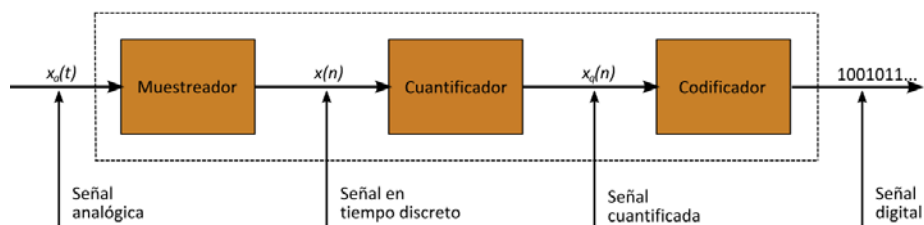


Ilustración 11. Esquema de digitalización de una señal.

Las señales digitales, no varían en forma continua, sino que cambian en pasos o en incrementos discretos. Las señales en tiempo discreto son aquellas que se representan matemáticamente como una secuencia de números. Además del carácter de estar definidas en tiempo discreto, la amplitud de la señal puede ser también discreta.

2.5.1.1. Muestreo

En la mayoría de los casos, las señales en tiempo discreto surgen de tomar muestras de una señal analógica. De esta forma, el valor numérico del n -ésimo número de la secuencia es igual al valor de la señal analógica $x_a(t)$, en el instante temporal nT_s , es decir:

$$\hat{x}(n) = x_a(nT_s), \quad -\infty < n < \infty$$

La cantidad T_s se denomina periodo de muestreo y su inversa es la frecuencia de muestreo f_s .

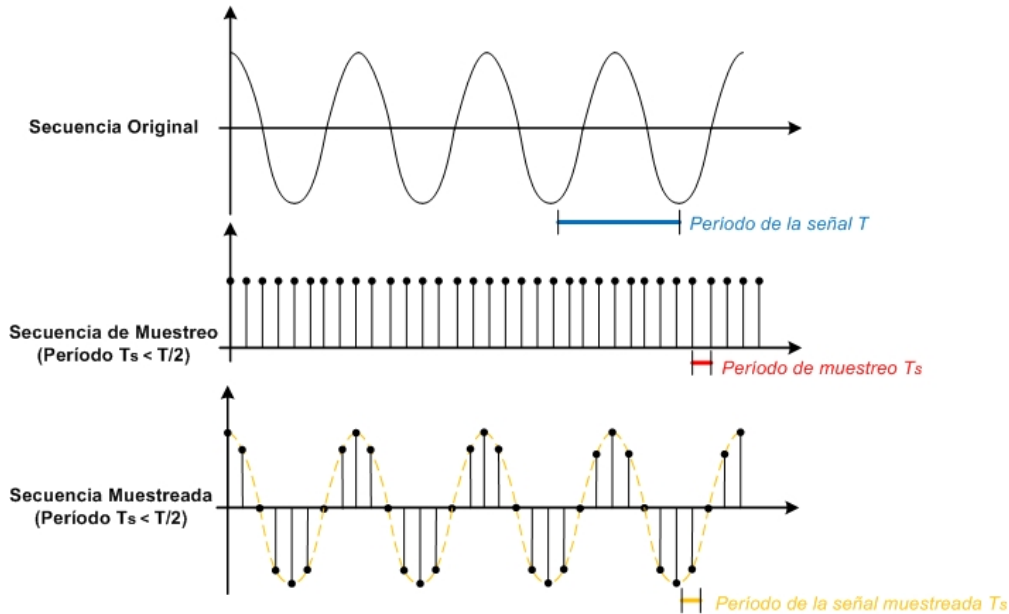


Ilustración 12. Proceso de muestreo de una señal analógica.

El teorema de Nyquist establece que, para poder reconstruir una señal a partir de sus muestras, se debe utilizar una frecuencia $N_s \geq 2 f_N$, al menos el doble de f_N . Siendo f_N la componente de más alta frecuencia de la señal.

El espectro de frecuencias del sonido audible por los humanos, es aproximadamente de 20 Hz a 20 kHz. Por esto, las señales de audio se muestrean generalmente a 44100 Hz, más del doble de la máxima frecuencia audible.

El contenido en frecuencia de las señales de voz puede abarcar hasta 15 kHz o más, pero la voz es altamente inteligible incluso con bandas de frecuencia limitadas a unos 4 kHz. Ese es el caso de los sistemas telefónicos comerciales donde la frecuencia de muestreo estándar utilizada para la voz es de 8 kHz. En la etapa de muestreo se obtiene una señal en tiempo discreto $\hat{x}(n)$ cuyas amplitudes son valores continuos. Para digitalizar la señal resta discretizar esos valores (cuantificarlos).

2.5.1.2. Cuantificación

El propósito del cuantificador es transformar la muestra de entrada $\hat{x}(n)$ en un valor $x(n)$ de un conjunto finito de valores preestablecidos. Esto se realiza redondeando los valores de las muestras hasta el nivel de cuantificación más próximo.

La precisión de los datos dependerá del número de bits con que se codifiquen los niveles de cuantización. Por tanto, se introduce un ruido de cuantización que se modela como ruido blanco.

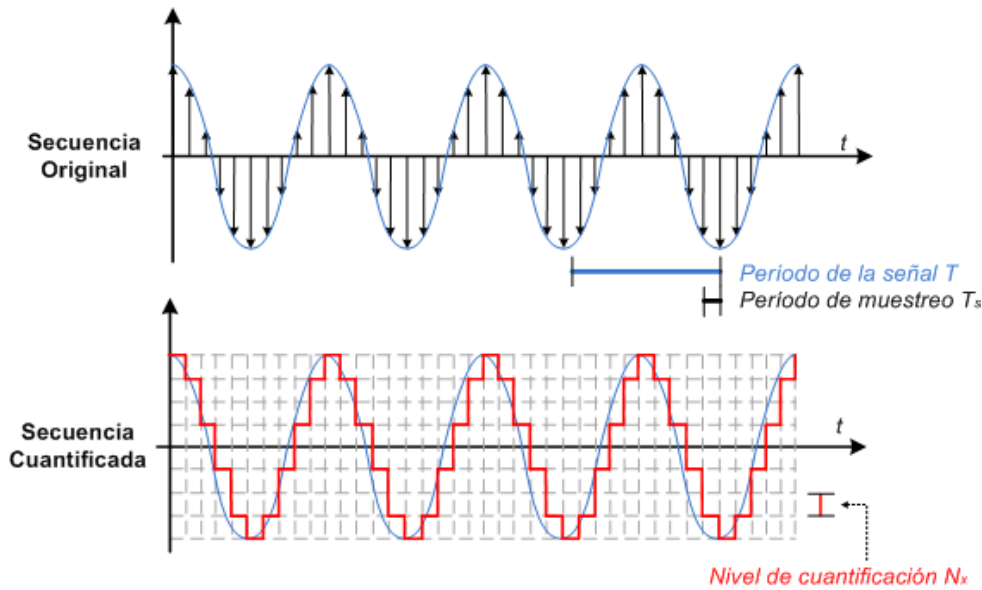


Ilustración 13. Proceso de cuantificación de una señal analógica.

2.5.2 Parametrización de la voz

La parametrización de la voz busca el objetivo de la extracción de información relevante de la señal acústica analógica, eliminando las redundancias y la información asociada a las fuentes de variabilidad que tiene la misma. La información relevante será aquella que permita:

- Diferenciar los fonemas existentes en las diferentes lenguas y que están caracterizados por:
 1. La *envolvente espectral* del fonema, determinada por los formantes que lo componen. Los formantes se definen como las frecuencias de resonancia del tracto vocal para cada fonema.
 2. El tipo de *excitación* que los produce. Las vocales y consonantes sonoras están generadas mediante una excitación periódica. La frecuencia fundamental de la excitación es también una característica definitoria del fonema, aunque es variable para los diferentes hablantes y las diferentes entonaciones.
 3. La *energía* de la señal. Las vocales y consonantes sonoras tienen mayor energía que las sordas, siendo la energía un buen parámetro de caracterización ya que presenta poca variabilidad para un mismo fonema una vez que ha sido convenientemente normalizada.
- Aportar datos sobre la prosodia de la frase tales como el acento, los tonos y la entonación. Esta información se obtiene analizando: las variaciones de la frecuencia fundamental, las variaciones de la duración de los fonemas y la variación en la intensidad de los fonemas diferenciados.

Teniendo en cuenta la información necesaria para caracterizar los fonemas y su prosodia, es razonable que la mayor parte de los sistemas de parametrización se basen en el análisis de la potencia espectral en tiempo corto. Al hacer este análisis, la señal se divide en tramas lo suficientemente cortas como para poder considerar la señal cuasi-estacionaria para someterla a un análisis espectral y quedando caracterizada por un vector de características que suele tener de 10 a 20 parámetros. La siguiente ilustración muestra de manera general el proceso de parametrización con las posibles variantes en cada una de las etapas.

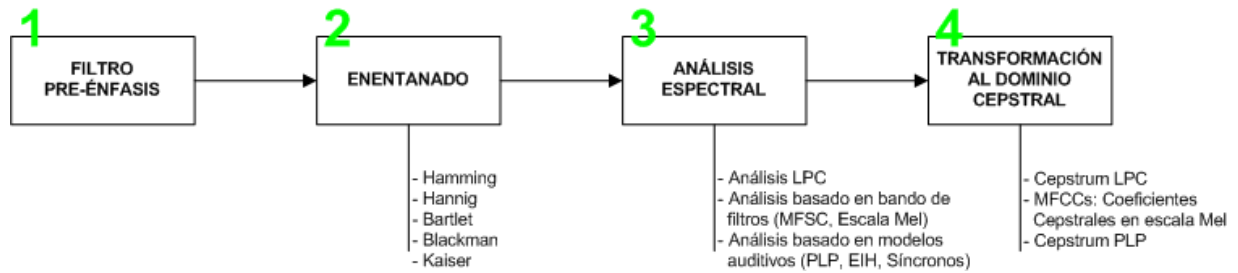


Ilustración 14. Proceso de parametrización de una señal de voz.

2.5.2.1. Filtrado de Preénfasis

La señal de voz muestreada pasa un filtro de preénfasis, típicamente un (Finite Impulse Response) de primer orden, que amplifica las altas frecuencias para compensar el efecto de los pulsos glotales y la impedancia de radiación. Generalmente, este filtro sigue la expresión:

$$H(z) = 1 - \mu \cdot z^{-1}, \quad \text{donde } 0.95 \leq \mu \leq 0.98$$

El filtro de preénfasis se destina para alzar el espectro de la señal aproximadamente 20 dB por década. Su utilización es necesaria porque los segmentos de voz sonoros tienen una pendiente espectral negativa (aproximadamente 20 dB por década) y con este filtro tiende a contrarrestar esta pendiente mejorándose la eficiencia de las etapas posteriores. Por otra parte, el sistema auditivo humano es más sensible por encima de 1 KHz en la región del espectro. Este filtro amplifica esta zona del espectro ayudando a las etapas posteriores de análisis.

2.5.2.2. Enventanado

La señal de voz es un proceso aleatorio y no estacionario, lo que supone un inconveniente a la hora de analizar la señal. No obstante, es posible salvar este problema si se tiene en cuenta que a corto plazo de tiempo (del orden de *ms*) la señal es cuasi-estacionaria. Este hecho da lugar a un tipo de análisis donde se obtienen segmentos o tramas de la señal de pocos *ms* denominado análisis localizado. A este proceso donde se obtienen tramas o segmentos consecutivos de señal se le denomina enventanado.

El enventanado requiere que cada una de las tramas sea multiplicada por una función limitada en el tiempo de tal manera que su valor fuera de ese intervalo sea nulo. De esta forma, el enventanado consiste en agrupar las muestras de la señal $x(n)$ en bloques de N elementos, y multiplicarlas por una ventana $w(n)$.

Para mantener la continuidad de la información de la señal, es muy común realizar el enventanado con bloques de muestras solapados entre sí, de esta forma no se pierden los eventos en la transición entre ventanas.

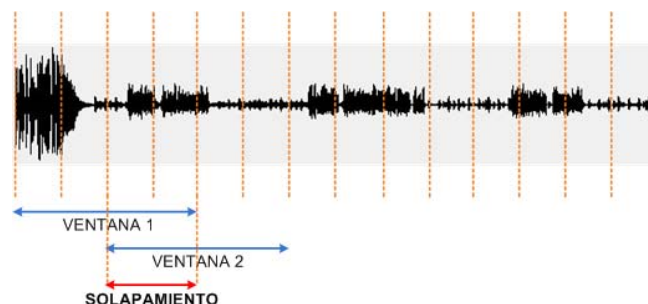


Ilustración 15. Enventanado y solapamiento de una señal de voz.

Las ventanas más utilizadas para la función de enventanado, así como su representación temporal se muestran en la siguiente tabla.

Ventana	Fórmula
Rectangular	$w(n) = 1 \quad 0 < n < N$
Hanning	$w(n) = \frac{1}{2} - \frac{1}{2} \cos\left(\frac{2\pi n}{N}\right) \quad 0 < n < N$
Hamming	$w(n) = \frac{27}{50} - \frac{23}{50} \cos\left(\frac{2\pi n}{N}\right) \quad 0 < n < N$
Bartlett	$w(n) = \begin{cases} \frac{2n}{N} & 0 < n < \frac{N}{2} \\ 2 - \frac{2n}{N} & \frac{N}{2} < n < N \end{cases}$
Blackman	$w(n) = \frac{21}{50} - \frac{1}{2} \cos\left(\frac{2\pi n}{N}\right) + \frac{2}{25} \cos\left(\frac{4\pi n}{N}\right)$
Kaiser	$w(n) = \frac{I_0\left(\pi\beta\sqrt{1 - \left(\frac{2n}{N}\right)^2}\right)}{I_0(\pi\beta)}$

Tabla 3. Ventanas más utilizadas para enventanado.

Cada ventana se caracteriza por la forma de sus lóbulos central y laterales en frecuencia. Se requiere de una ventana, que su lóbulo central sea lo más angosto posible y que los lóbulos laterales sean pequeños para tener una buena resolución en frecuencia.

La ventana rectangular posee el lóbulo central con menor ancho de banda de todos, pero sus lóbulos laterales decaen muy lentamente. Estos hacen aparecer el efecto de ‘ripple’ (fenómeno de Gibbs), no deseado por la distorsión armónica que genera. El resto de las ventanas tiene cada una distintas propiedades, que según la aplicación podrán ser de un modo u otro ventajosas.

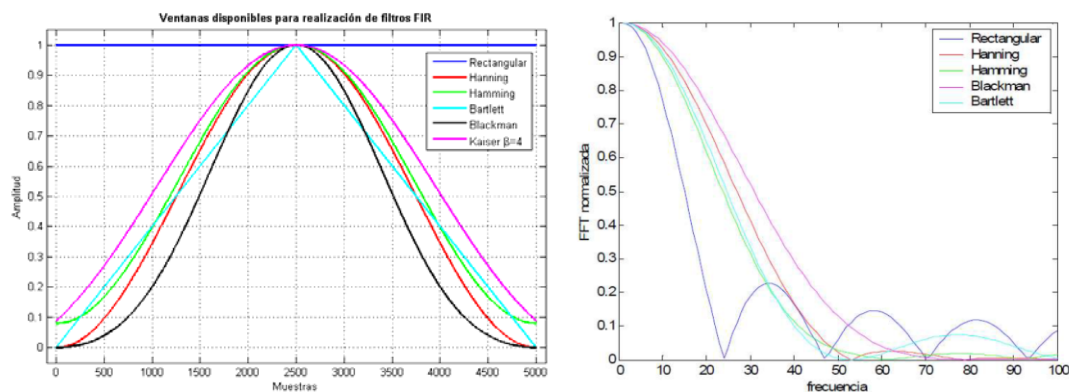


Ilustración 16. Representación temporal y frecuencial de las ventanas más utilizadas.

2.5.2.3. Transformada discreta de Fourier

Los sistemas lineales e invariantes en el tiempo, cumplen ciertas propiedades que permiten la representación de las señales en frecuencia. Una de las propiedades es que la respuesta a secuencias sinusoidales es también sinusoidal, de igual frecuencia y con amplitud y fase determinadas por el sistema.

Esta propiedad hace que las representaciones de las señales mediante sinusoides o exponenciales complejas (la transformada de Fourier) sean muy útiles. Para las secuencias de duración finita, se utiliza la Transformada Discreta de Fourier (DFT). Se llaman secuencias base a las exponenciales complejas que se utilizan para representar la señal. Dada una señal en tiempo discreto $x(n)$ con N muestras, su transformada $X(k)$ está dada por:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j\frac{\pi kn}{N}}$$

En la práctica, el costo computacional del cálculo de la DFT se reduce utilizando la transformada rápida de Fourier, FFT (Fast Fourier Transform), un algoritmo eficiente que permite calcular la transformada de Fourier discreta (DFT) y su inversa. La FFT es de gran importancia y uso en una amplia variedad de aplicaciones, desde el tratamiento digital de señales y filtrado digital en general, a la resolución de ecuaciones diferenciales parciales o los algoritmos de multiplicación rápida de grandes números enteros.

La evaluación directa de la fórmula anterior requiere $O(n^2)$ operaciones aritméticas. Mediante un algoritmo FFT se puede obtener el mismo resultado con solo $O(n \log n)$ operaciones. En general, dichos algoritmos dependen de la factorización de n pero, al contrario de lo que frecuentemente se cree, existen FFTs para cualquier n , incluso con n primo.

La idea que permite esta optimización es la descomposición de la transformada a tratar en otras más simples y estas a su vez hasta llegar a transformadas de dos elementos donde k puede tomar los valores cero y uno. Una vez resueltas las transformadas más simples hay que agruparlas en otras de nivel superior que deben resolverse de nuevo y así sucesivamente hasta llegar al nivel más alto. Al final de este proceso, los resultados obtenidos deben reordenarse.

Dado que la transformada discreta de Fourier inversa es análoga a la transformada discreta de Fourier, con distinto signo en el exponente y un factor $1/n$, cualquier algoritmo FFT puede ser fácilmente adaptado para el cálculo de la transformada inversa.

2.5.2.4. Transformada discreta del coseno

La DCT es una transformada muy similar a la DFT (Discrete Fourier Transform) en donde las secuencias base son cosenos y la representación de una señal real mediante esta transformada, es también real. La fórmula que expresa dicha señal es la siguiente:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cos \left[\frac{k\pi}{N} \left(n + \frac{1}{2} \right) \right]$$

Se utiliza en muchas aplicaciones de compresión de datos con preferencia sobre la DFT debido a una propiedad que se denomina generalmente “compactación de la energía”. La DCT tiende a concentrar la mayor parte de la información de la señal en los coeficientes de baja frecuencia. Gracias a

esto, se necesita un menor número de coeficientes para representarla. Esta propiedad se muestra en la siguiente figura:

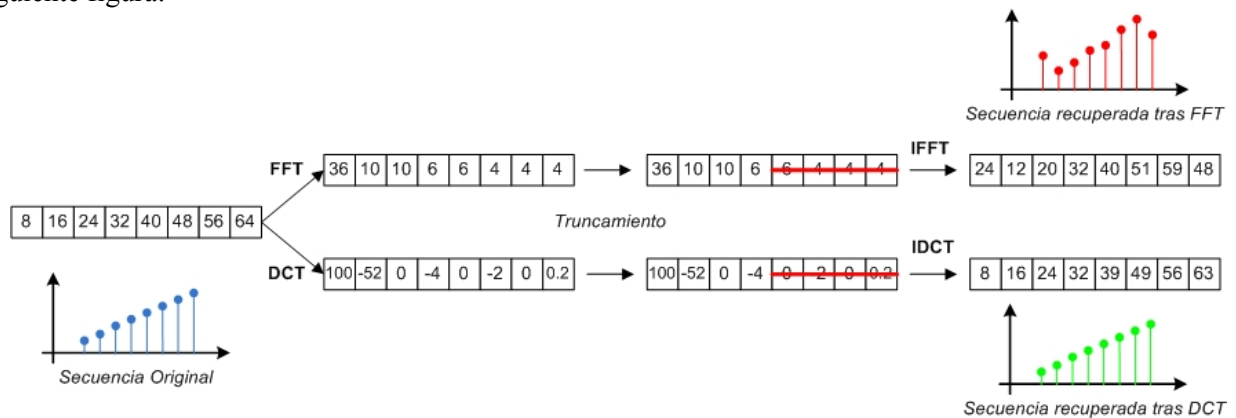


Ilustración 17. Propiedad de compactación de la DCT frente a la FFT.

2.5.2.5. Análisis Espectral

El análisis de los segmentos de voz obtenidos se puede hacer tanto en el dominio del tiempo como en el dominio de la frecuencia. En el dominio del *tiempo* las magnitudes que se analizan son la energía local, la tasa de cruces por cero de la señal y su autocorrelación. Este dominio aporta un análisis de la señal rápido, sencillo y con una interpretación inmediata. Sin embargo, el análisis *espectral* es el utilizado por su mayor potencia para caracterizar la información de la señal de voz.

La parametrización usada en el reconocimiento de voz se deriva del análisis de la potencia espectral de las tramas de voz. El análisis de la fase del espectro de frecuencias se omite debido a que los oídos son insensibles a las variaciones de la fase y en consecuencia los equipos de comunicaciones de voz y de grabación no preservan la fase original, que también se ve alterada por factores no deseados como la acústica del entorno. El análisis de la potencia espectral se hace además en escala logarítmica por motivos prácticos:

- La escala logarítmica hace que cuando la ganancia que tiene la señal cambia, *la forma del espectro de potencias se mantenga*.
- El filtrado lineal debido a la acústica del entorno o a variaciones en el canal, tiene un efecto convolucional en el dominio del tiempo, un *efecto multiplicativo* para el espectro de potencias lineal y un simple efecto de suma de una constante para los espectros logarítmicos de potencias.
- La forma de onda de la voz se puede modelar como la convolución en el dominio temporal de la excitación de una cuasi-periódica con un filtro variante en el dominio del tiempo, que está determinado por la configuración del tracto vocal para la producción de dicha señal de voz.

Esta configuración del tracto vocal como filtro variante en el tiempo va a ser de la que se obtenga información sobre los fonemas articulados. Es deseable poder separar estas dos componentes de la forma de onda (*excitación cuasi-periódica y filtro variante*), siendo el dominio de la potencia espectral logarítmica el óptimo para hacerlo ya que en dicho dominio ambos componentes son *aditivos*.

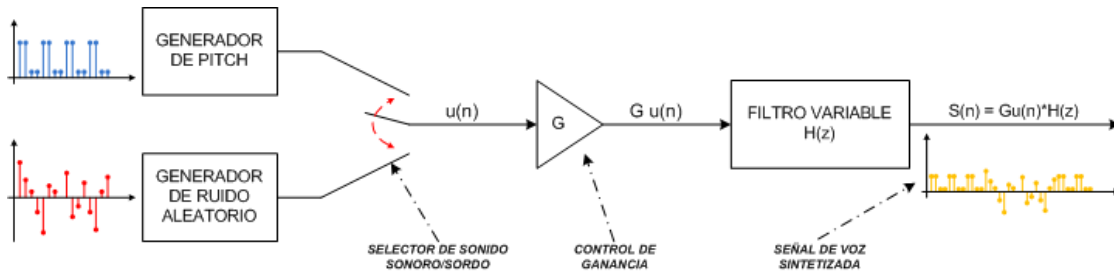


Ilustración 18. Modelo digital de producción de voz.

Existen tres técnicas de análisis espectral que serán descritas a continuación:

- **Representaciones basadas en el modelo LPC** (Linear Predictive Coding): El modelo digital de producción de voz de la ilustración anterior se usa para modelar el tracto vocal como un tubo acústico a través del cual el sonido se propaga como una onda plana. Los efectos del tracto vocal en la señal de excitación son la creación de un conjunto de sonidos resonantes, quedando así el tracto modelado como un filtro todo polos $H(z)$. El análisis espectral permite dar los coeficientes de dicho filtro, $H(z)$, sin calcular el espectro explícitamente.
- **Análisis basado en el banco de filtros mediante coeficientes (MFSC Mel Frequency Spectral Coefficients)**: El espectro de potencias de la señal se obtiene aplicando la transformada de Fourier a las tramas de voz de ventanas que se solapan y por ello aparecen armónicos a frecuencias múltiplo de la frecuencia fundamental de las tramas. Este efecto se puede subsanar agrupando los conjuntos de componentes cercanos en unas 20 bandas de frecuencias antes de hacerles el logaritmo de la potencia.

Cada filtro es un promedio ponderado de las componentes espectrales presentes en su banda, caracterizado el tracto de resolución perceptual del oído humano, haciendo que las bandas que abarcan los filtros sean más anchas para frecuencias superiores a 1 KHz. Esta escala recibe el nombre de "Escala Mel". El logaritmo de la energía a la salida de los filtros en escala Mel da lugar a los coeficientes MFSC.

- **Análisis basado en modelos auditivos**. Este análisis está basado en los aspectos fisiológicos y psicofísicos del proceso auditivo humano que incorpora a los criterios de parametrización. Las anteriores estrategias de parametrización estaban basadas en el modelo de producción de la voz.

Este grupo de técnicas imita el modelo de percepción de la voz humana para parametrizar el habla, intentando reproducir el comportamiento de la membrana basilar del oído humano en la percepción. Para ello se persigue implementar mecanismos que capturen la información fisiológica que caracteriza a la percepción humana:

- Análisis de frecuencias en canales paralelos.
- Conservación de la estructura temporal fina del sonido.
- Rango dinámico limitado en los canales individuales.
- Realce de los contrastes temporales.
- Realce de los contrastes espectrales en frecuencias adyacentes.

Los resultados obtenidos con estas técnicas alrededor de los años 90 son bastante buenos, ligeramente mejores que los de los parámetros MFCC del dominio Cepstral (que serán analizados más adelante y que son los más utilizados en la actualidad), ya que captan cierta información útil adicional:

- La estructura temporal detallada de la señal.
- La supresión lateral de los canales adyacentes.
- Los contrastes temporales.
- Otras características no lineales del proceso de audición.

Sin embargo, esta línea de investigación no siguió desarrollándose ya que, aunque los resultados eran buenos, llevaban asociado un coste computacional y de almacenamiento que no compensaba ni era factible para reconocimientos en tiempo real con coste computacional razonable. En la actualidad, existe un resurgimiento de esta línea de trabajo motivado por las capacidades de computación y almacenamiento superiores a las existentes en los años 80, por la necesidad de encontrar parametrización que mejoren MFCC y permitan enfrentarse a los actuales retos de reconocimiento.

2.5.2.6. Análisis Cepstral

Las técnicas de análisis espectral que operan en el dominio de la potencia espectral logarítmica tienen la limitación de que, debido a que los espectros de los filtros en bandas adyacentes están bastante correlados, originan coeficientes espectrales también bastante correlados. Es deseable eliminar esa correlación manteniendo solo la información que sea útil para el reconocimiento. Para ello, se utiliza un filtro de decorrelación homomórfica o Cepstrum que, mediante la transformada inversa de Fourier del logaritmo del espectro de potencias, lleva los coeficientes cepstrales al dominio de la *cuefrecia* convirtiéndolos en coeficientes cepstrales.

Los coeficientes cepstrales representan la señal temporal que corresponde al espectro logarítmico de potencia. El dominio de la *cuefrecia* es un dominio homomórfico del dominio temporal. Esto implica que las convoluciones en el dominio temporal se convierten en sumas en su dominio homomórfico de la *cuefrecia*.

Esto será enormemente útil ya que permitirá separar las señales de voz de los ruidos convolucionales con los que estén mezcladas. Las componentes de excitación y envolvente espectral del tracto vocal aparecerán en zonas separadas del dominio transformado de la *cuefrecia*, que se podrán separar mediante ventanas. Haciendo un juego de paralelismo, los inventores de este operador homomórfico llamado Cepstrum (cuyo nombre crearon intercambiando la posición de las cuatro primeras letras del término spectrum), llamaron a ese enventanado en el dominio de la *cuefrecia* "liftering" (cambiando la posición de las primeras letras del término correspondiente filtering del dominio spectrum).

El análisis en el dominio espectral tiene su correspondiente homólogo en el dominio de la *cuefrecia* que reciben los nombres de coeficientes cepstrales LPC (Linear Predictive Coding), MFCC (Mel Frequency Cepstral Coefficients), Cepstrum PLP y HFCC (Human Factor Cepstral Coefficients). Los coeficientes MFCC parecen ser los que mejores resultados dan como técnica de parametrización teniendo en cuenta el compromiso entre coste computacional y resultados obtenidos.

2.6. Reconocimiento de voz usando HMMs

Un modelo oculto de Markov ó Hidden Markov Model (HMM) es la representación de un proceso estocástico que consta de dos mecanismos interrelacionados: una cadena de Markov de primer orden subyacente, con un número finito de estados, y un conjunto de funciones aleatorias, cada una de las cuales asociada a un estado.

En un instante discreto de tiempo se supone que el proceso está en un estado determinado y que genera una observación mediante la función aleatoria asociada. Al instante siguiente, la cadena subyacente de Markov cambia de estado siguiendo su matriz de probabilidades de transición entre estados, produciendo una nueva observación mediante la función aleatoria correspondiente. El observador externo solo es sensible a la salida de las funciones aleatorias asociadas a cada estado, siendo incapaz de observar directamente la secuencia de estados de la cadena de Markov.

La ilustración siguiente muestra un ejemplo de una cadena de Markov típica. El ejemplo se ha extraído del manual de uso del conjunto de herramientas software HTK [16], el cual se utiliza para desarrollo de este proyecto.

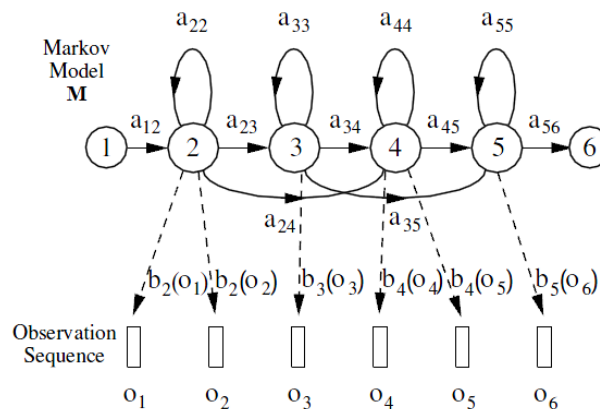


Ilustración 19. Ejemplo de cadena de Markov con seis estados.

El ejemplo anterior de un modelo oculto de Markov es llamado de tipo "Bakis" o de izquierda a derecha, con dos estados no emisores y cuatro estados emisores, es decir, que emiten función de probabilidad. Entonces, un modelo oculto de Markov es la composición de dos procesos estocásticos (X,Y) definidos como:

- Una cadena oculta de Markov X que tiene en cuenta la variabilidad temporal, y que no es directamente observable.
- Un proceso observable Y que tiene en cuenta la variabilidad espectral y va tomando valores en el espacio de las características acústicas u observaciones.

La combinación de ambos procesos modela las fuentes de variabilidad de la señal de voz y permite reflejar una secuencia de parámetros acústicos como concatenación de los procesos elementales del modelo con la flexibilidad suficiente para hacer sistemas de reconocimiento. Los modelos ocultos de Markov usados en el reconocimiento de voz tienen dos asunciones formales características:

- a. La historia de la cadena no influye en la evolución futura de la misma si existe información actual (hipótesis de Markov de primer orden).
- b. Ni la evolución de la cadena ni las observaciones pasadas determinan la observación actual si se ha especificado la última transición de la cadena (hipótesis de independencia de las salidas).

Una vez hechas esas asunciones, si se llama "y" $\in Y$ a una variable que representa las observaciones y si se denomina $i, j \in X$ a las variables que representan los estados del modelo, el modelo $\lambda = (A, B, P)$ queda representado por las siguientes matrices de parámetros según puede verse en la ilustración anterior:

$$\begin{aligned} A &\equiv a_{i,j} \mid i, j \in X, && \text{Probabilidades de transición} \\ B &\equiv b_{i,j} \mid i, j \in X, && \text{Probabilidades de salida} \\ \Pi &\equiv \pi_i \mid i \in X, && \text{Probabilidades iniciales} \end{aligned}$$

Donde los términos de las matrices se definen como:

$$\begin{aligned} a_{i,j} &\equiv p(X_t = j \mid X_{t-1} = i) \\ b_{i,j} &\equiv p(X_{t-1} = j \mid X_t = i) \\ \pi_i &\equiv p(X_0 = i) \end{aligned}$$

La técnica de HMM se usa en la actualidad en aquellos sistemas en los que el modelado tiene una dependencia del tiempo como pueden ser los sistemas reconocimiento fonético y del habla. Es una práctica universal usar los modelos ocultos de Markov para calcular las probabilidades acústicas debido a su capacidad de modelar estadísticamente de manera adecuada la generación de voz. Los HMM en reconocimiento de voz se utilizan teniendo en cuenta dos hipótesis:

1. La voz se puede dividir en segmentos o estados, en los que la señal de voz se puede considerar estacionaria. Es decir, en la ventana de análisis la señal mantiene la estructura de principio a fin. Se asume que las transiciones entre segmentos contiguos son instantáneas.
2. La probabilidad de observación de que un vector de características se genere depende solo del estado actual y no de símbolos anteriores. Esta es una suposición de Markov de primer orden denominada hipótesis de independencia.

Otra razón por la que los HMMs son populares, es porque pueden ser entrenados automáticamente, siendo factible realizar los cálculos en un tiempo razonable. El reconocimiento fonético es la disposición más simple posible, para los que los modelos de Markov tendrán en cada estado una distribución estadística llamada mezcla de Gaussianas de matriz de covarianza diagonal, que de una probabilidad para cada vector observado.

Un modelo oculto de Markov para una secuencia de fonemas se construye concatenando los modelos ocultos entrenados para los fonemas separados. Estas facilidades han servido para el desarrollo del presente proyecto.

El uso de los HMM permite eludir las limitaciones de algunos otros sistemas en el reconocimiento de fonemas como son los siguientes:

- **DTW** (Alineamiento temporal Dinámico) no hay posibilidad de realizar un entrenamiento estadístico, ya que se realiza comparaciones entre secuencias de vectores de parámetros.
- **VQ** (Cuantificación temporal) asignación dura entre los vectores y la clase que modela. Además tiene que respetar el compromiso entre el tamaño del codebook y el error de cuantificación.

Los modelos ocultos de Markov se caracterizan por tres problemas que es necesario resolver para que resulten modelos útiles en aplicaciones de reconocimiento de voz reales:

1. **Evaluación.** Es el problema sencillo, pues dada una observación acústica y un modelo oculto de Markov, determinar la probabilidad de que el modelo genere esa observación, es decir, la probabilidad acústica $P(O|\lambda)$. Esta probabilidad se determina con el algoritmo forward-backward.
2. **Decodificación.** Determinación de la secuencia óptima (de mayor probabilidad) de estados $X = x_1, x_2, \dots, x_T$ dada la observación acústica y el modelo oculto de Markov. Es decir, se busca la alineación de la observación con el modelo, asignando cada vector a un estado del modelo. Se lleva a cabo mediante el algoritmo de Viterbi, siendo una de las utilidades implementadas en el software de herramientas HTK.

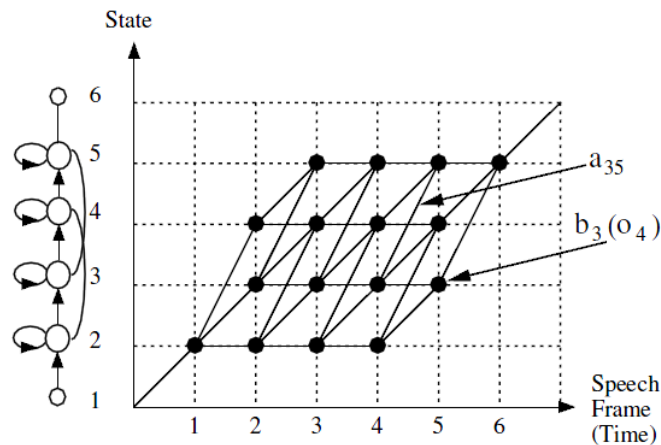


Ilustración 20. Camino de mayor probabilidad mediante el algoritmo de Viterbi.

3. **Estimación o entrenamiento de los HMMs.** Consiste en el cálculo de los parámetros que caracterizan el modelo. Dada un conjunto de datos y una colección de secuencias observables, se determina el HMM que con mayor probabilidad ha generado la secuencia. Este problema se resuelve comúnmente con el algoritmo Baum-Welch.

2.7. Algoritmos de extracción de características

En el reconocimiento del habla, la señal de voz, una vez digitalizada, se procesa para producir una nueva representación de la voz en forma de secuencia de vectores o agrupaciones de unos valores que se denominan parámetros y que deben representar la información contenida en la envolvente del espectro.

El número de parámetros debe ser reducido, puesto que la base de datos de entrenamiento siempre es limitada, por lo que cuantos más parámetros tenga la representación, menos fiables son los valores entrenados y, por otro lado, más costoso es el proceso de reconocimiento.

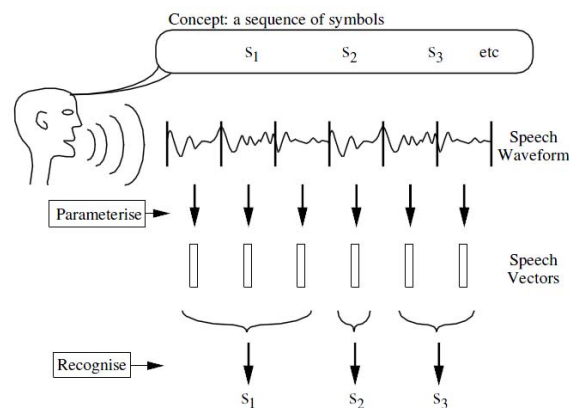


Ilustración 21. Esquema de extracción de características en el proceso de habla humana.

Existen varios algoritmos que permiten extraer, de una forma u otra, las características frecuenciales de la voz humana como pueden ser MFCC (coeficientes cepstrales en la escala Mel de frecuencias), HFCC (coeficientes cepstrales basados en factores humanos), PLP (análisis porcentual lineal predictivo), LPC (codificación lineal predictiva), LPC-Cepstrum. A continuación se detalla de forma precisa el algoritmo MFCC, usado para la realización de este proyecto.

2.7.1 Mel Frequency Cepstral Coefficients

A imitación de lo que sucede en el sistema auditivo humano, la identificación de los sonidos se hace en el dominio de la frecuencia. Entre las muchas técnicas de parametrización del habla, la más empleada es la denominada MFCC (Mel Frequency Cepstral Coefficients). Es la técnica de parametrización del habla más utilizada en los sistemas automáticos de reconocimiento de voz, principalmente porque se adapta bien a las hipótesis utilizadas para estimar las distribuciones de estado de los HMMs, y también, debido a la robustez de ruido superior que ofrece sobre otras técnicas alternativas de extracción de características, como LPCC.

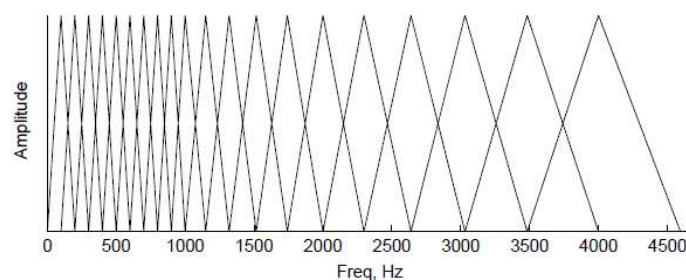


Ilustración 22. Banco de filtros logarítmicos para la obtención de coeficientes MFCC.

El ancho de banda de los filtros triangulares en MFCC viene determinado por la distribución de la frecuencia de centro de cada filtro, siendo esta función de la frecuencia de muestreo y el número de filtros. Es decir, si el número de filtros en el banco de filtros aumenta, el ancho de banda de cada filtro decrece.

Si bien las características del banco de filtros en MFCC se derivan del sistema auditivo humano, la elección del número de filtros o la forma de estos, así como el factor de solapamiento entre ellos queda a elección del programador. La forma en triángulo de cada filtro utilizado en MFCC, aproxima a modelos de la banda de paso natural de las bandas críticas del sistema auditivo humano, aunque la conocida relación entre la frecuencia central y ancho de banda crítico no se utiliza para establecer el ancho de banda del filtro.

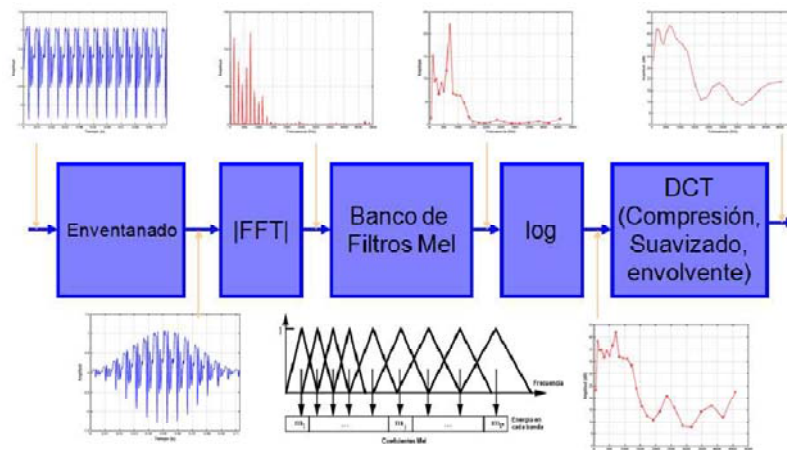


Ilustración 23. Proceso de extracción de coeficientes MFCC.

Los coeficientes MFCC representan la envolvente espectral de la señal de voz, obteniendo importantes características identificadoras del habla. En concreto, el primer coeficiente C_0 , indica la energía de la señal y se usa o no dependiendo de la aplicación. El segundo coeficiente C_1 , tiene una razonable interpretación como indicador del balance global de energía entre bajas y altas frecuencias.

Para obtener más información, como por ejemplo la de coarticulación de fonemas, es necesario introducir datos de la velocidad y aceleración de los parámetros. Así surgen los MFCC-Delta (o Δ MFCC) y los MFCC Delta-Delta (o $\Delta\Delta$ MFCC), que representan la evolución temporal de los fonemas en su transición a otros fonemas. Los Δ MFCC se calculan como la variación de los coeficientes MFCC con respecto a un instante de tiempo. Por ello, son denominados coeficientes de velocidad (ya que dan los cambios por tiempo) o de primera derivada. Los coeficientes $\Delta\Delta$ MFCC representan la variación de los coeficientes de velocidad, por lo que son llamados coeficientes de aceleración.

Sin embargo, los coeficientes MFCC son difíciles de relacionar con cualquier aspecto cerrado de la producción o percepción del habla. Los detalles espectrales que contienen permiten la discriminación entre sonidos similares, pero su carencia de interpretación los hace altamente vulnerables a condiciones no lineales tales como el ruido o acentos. En particular, los MFCCs dan igual peso a las altas y bajas amplitudes en el espectro logarítmico, cuando es bien conocido que la alta energía domina en la percepción.

2.8. Segmentación de Audio

Se conoce como Segmentación de Audio la división temporal en segmentos de idéntica longitud de una secuencia de audio. El fin de dicha división es facilitar la extracción de parámetros de la secuencia completa, analizando segmentos más pequeños.

Desde hace varios años, la comunidad mundial dedicada al reconocimiento del locutor y del idioma, se ha dedicado con especial atención a resolver situaciones de aplicación real. Las grabaciones de audio existentes disponen de una señal no muy limpia debido a interferencias, ruidos de canal o música de fondo. Este hecho y ayudados por la investigación de nuevos algoritmos más eficientes y potentes, así como por la mejora computacional de los ordenadores ha servido de motivación para que el grupo de reconocimiento de voz ATVS participara en la evaluación de Albayzin 2010 [1].

Otros métodos de segmentación con buenos resultados son el Análisis Factorial a corto plazo propuesto por Castaldo [2] o la mejora de este sistema mediante el modelado del habla y la variabilidad del canal propuesta por Najim y Kenny [3]. También se han obtenido mejoras con un entrenamiento discriminativo de clases o la estimación de probabilidades basadas en la distancia del coseno según lo explicado por Najim [7].

Para el desarrollo de este proyecto se ha utilizado el Sistema de Segmentación de Audio ATVS-UAM. Está basado en la parametrización de secuencias de audio usando un HMM de 5 estados, uno por clase acústica: 'voz', 'voz con ruido de fondo', 'voz con música de fondo', 'música' y 'otros'.

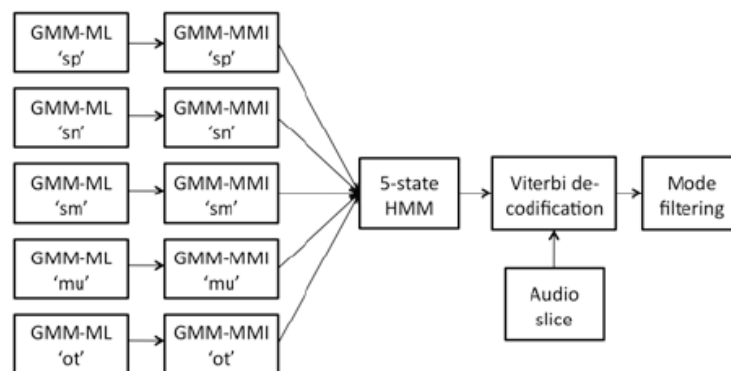


Ilustración 24. Esquema de funcionamiento del sistema de Segmentación de Audio ATVS-UAM.

Cada estado del HMM consiste en un modelo de mezclas de gaussianas (GMM) de 1024 mezclas entrenadas mediante el método de máxima verosimilitud, y mejorado con 18 iteraciones del criterio de máxima verosimilitud mutua con el kit de herramientas STK de la Universidad de Brno [9].

Las características han sido obtenidas tras la división del audio en segmentos de 60 segundos con solapamiento entre ellos de 2 segundos usando las herramientas disponibles en el Toolkit de Matlab [19]. Después se emplea la mencionada decodificación por Viterbi, se ha utilizado un filtrado con una ventana de 700 ms y se han eliminado los segmentos cuya longitud era inferior de 3 segundos.

Finalmente, para cada secuencia completa de audio se obtienen muchos segmentos caracterizados por tres parámetros: el tiempo de inicio, el tiempo de fin y la clase de audio que mejor lo identifica.

2.9. Diarización de Locutores

Se entiende por Diarización de Locutores la identificación de segmentos de voz con la persona que lo ha pronunciado, basándose en características de propias del habla y en bases de datos de dichas personas, siempre salvando cualquier tipo de identidad [5].

Para este proyecto se han utilizado dos sistemas que se describen a continuación:

2.9.1 Sistema de Diarización de Locutores ATVS-UAM

El sistema de Diarización de Locutor UAM-ATVS [1], extrae primero las características MFCC tras la división del audio en segmentos de 90 segundos con un 33% de solapamiento.

Los vectores de información (iVectors) son procesados usando la medida de similitud conocida como la distancia del coseno [7]. Con ello se podrá averiguar la frecuencia de aparición de un locutor en los segmentos de audio. Una segunda agrupación permite reestimar los centroides que mejor representan el modelo específico del locutor, asignando una probabilidad de aparición de cada locutores en la secuencia de audio.

Por último, se obtienen las puntuaciones mediante el algoritmo de Viterbi, y se procede a etiquetar los segmentos.

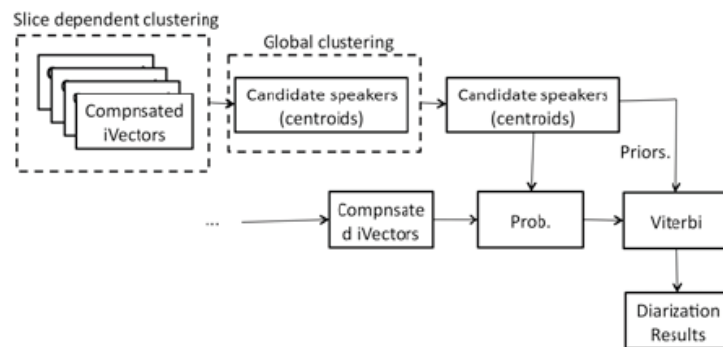


Ilustración 25. Esquema de funcionamiento del sistema de Diarización de Locutores ATVS-UAM.

2.9.2 Sistema de Diarización de Locutores empleando Reconocimiento Fonético

El Sistema de Reconocimiento Fonético tiene como fin identificar los fonemas de una determinada lengua que existen en una secuencia de audio.

El Reconocedor Fonético disponible hace uso de las herramientas implementadas en el Toolkit HTK [17]. Con él se puede obtener una lista detallada de segmentos que corresponden a fonemas o segmentos que corresponden a silencios.

Dichos segmentos se encuentran caracterizados por tiempos de inicio y final y puntuación o *score* en referencia a una base de datos con parámetros obtenidos tras un entrenamiento inicial. Esta base de datos está compuesta por: un diccionario fonético de la lengua (en este caso castellano), una lista de trifenemas más frecuentes de dicha lengua y HMMs obtenidos en grabaciones espontáneas de locuciones telefónicas.

Tras un análisis de la velocidad de pronunciación de fonemas en los segmentos en los que haya fonemas, se podrá determinar de forma más o menos acertada, los segmentos en los que hay voz y los segmentos en los que no hay voz y por lo tanto habrá silencios, música, ruidos, etc...

2.10. Identificación de Idioma

2.10.1. Introducción

La señal de voz contiene gran cantidad de información. Esta información se clasifica en dos grandes grupos, información de bajo nivel e información de alto nivel. En un análisis más fino se puede dividir cada uno de estos grandes grupos en dos, dando como resultado una clasificación de las particularidades de la señal de voz en cuatro niveles: acústico, fonético, prosódico y léxico.

Al igual que sucede con los aspectos relacionados con la identidad del locutor, las particularidades específicas de cada idioma se encuentran esparcidas por todos los niveles. Si bien es verdad que los niveles superiores parecen ser los más relevantes en la tarea de reconocimiento de idioma.

Las particularidades articulatorias y la configuración fisiológica independiente de cada idioma hacen que existan diferencias a nivel acústico. También se muestran diferencias a nivel prosódico ya que cada idioma presentará unos patrones prosódicos representativos, es decir, patrones de energía, duración y tono característicos. Ya en el nivel más alto se podrán observar las diferentes palabras y estructuras gramaticales características de cada idioma.

2.10.2. Niveles de distinción idiomática

Estos parámetros característicos permiten establecer una diferenciación idiomática por niveles:

- *Nivel acústico*: Los diferentes idiomas pueden tener patrones acústicos distintos, por ejemplo siendo más nasales, dentales, palatales, guturales, etc.
- *Nivel fonético*: Los idiomas difieren también en el conjunto de fonemas que utilizan, así como en la frecuencia de utilización de los distintos sonidos y en la frecuencia de aparición de secuencias de sonidos. A este grupo pertenecen las características que se utilizarán para el reconocimiento de idioma en la tarea de reconocimiento de idioma del proyecto. Con ello se pretende modelar la aparición de una secuencia de sonido para caracterizar un idioma.
- *Nivel prosódico*: También se diferencian por tener distintos patrones prosódicos (duraciones, energía y tono de los fonemas), es decir, cada idioma tiene una entonación característica.
- *Niveles léxico, gramatical y superiores*: Finalmente, y posiblemente lo más importante desde un punto de vista conceptual, los idiomas tienen distintos vocabularios y distintas formas de combinar las palabras. El conjunto de palabras es posiblemente lo más característico de un idioma, de modo que un idioma puede reconocerse como tal si se emplea el vocabulario correcto pero no se emplean los fonemas o prosodia correcta. Pese a ser el nivel que más información sobre el idioma puede aportar, en la actualidad son pocas las técnicas que emplean este nivel para realizar la clasificación.

Para tratar de acercarse a un detector de idioma preciso, idealmente se debería de hacer uso de las particularidades en los cuatro niveles del idioma descritos anteriormente. Sin embargo, el cuarto nivel (léxico, gramatical y superiores) resulta muy difícil de manejar porque requiere ser capaz de determinar la secuencia de palabras pronunciada a partir de exclusivamente la voz. En definitiva, para sacar partido de las particularidades del idioma en niveles léxicos y superiores es necesario disponer de un reconocedor automático de voz con una precisión suficiente y capaz de manejar todos los idiomas que se desee detectar.

Considerando que el reconocimiento de voz no está resuelto de modo satisfactorio ni siquiera en un único idioma sino que sigue siendo un tema de investigación muy activo (como lo atestiguan los resultados de D. T. Toledano y E. Campos, [10] y D. T. Toledano y A. Moreno [11]), lo habitual es que la detección de idioma se centre exclusivamente en los tres primeros niveles: acústico, fonético y prosódico. La ventaja de centrarse en estos niveles es que permite técnicas de modelado que pueden llegar a ser razonablemente independientes de los idiomas que se desea detectar, lo que proporciona una versatilidad que no se conseguiría con los niveles léxico y superiores.

2.10.3. Aplicaciones

Las aplicaciones del reconocimiento del habla se centran en las tres áreas siguientes:

- *Indexado y recuperación de información en contenidos de audio y audiovisuales.* La creciente proliferación de contenidos multimedia en diversos ámbitos, como las emisiones de radiodifusión o Internet hace necesario la clasificación por idioma de los contenidos de los mismos. Hoy en día se necesita que los buscadores clasifiquen los contenidos multimedia de Internet y se indexen por idioma ya que es un entorno con una gran variedad lingüística.

El presente proyecto intenta iniciar un camino en este tipo de aplicaciones, sobre todo dirigidas a la recuperación de información multimedia en Internet.

- *En entornos telefónicos multilingües tanto automáticos como con operador.* En un entorno automático es necesaria la clasificación para que el usuario sea atendido en la lengua concreta desde el principio hasta el final. Por el contrario, para el caso de un operador se hace que la consulta sea más rápida y se elimina la necesidad de personal que haga la distinción de idioma. Por tanto, en este medio es importante la detección de idioma para poder hacer un enrutado automático de la consulta.
- *Sistemas de traducción simultánea voz a voz.* En estos procesos es necesario conocer el idioma de los interlocutores. Con la utilización de un sistema de reconocimiento de idioma no será necesaria la configuración de estos sistemas.

Todas estas aplicaciones tienen un interés evidente para grandes empresas relacionadas con servicios telefónicos, con radio y televisión, con buscadores de Internet, y empresas e instituciones dedicadas a la vigilancia y seguridad.

2.10.4. Técnicas empleadas

Al igual que sucede con el reconocimiento biométrico de locutor, la señal de voz es la portadora de la información relativa al idioma. Por este motivo las técnicas aplicadas al reconocimiento de locutor se pueden extrapolar al reconocimiento de idioma.

De entre las técnicas basadas en los niveles superiores de información es preciso destacar los sistemas de reconocimiento fonético PRLM, PPRLM utilizados en el presente proyecto y que se describen con más detalle en las siguientes secciones.

2.10.4.1. Sistemas basados en GMMs (Gaussian Mixture Models)

Los sistemas de reconocimiento de idioma de GMM se basan en el principio de que los idiomas tienen diferentes sonidos y que la frecuencia de aparición de los sonidos es diferente de un idioma a otros.

La realización de esta técnica consiste principalmente en seguir los siguientes pasos:

1. *Extraer características de la voz*, en particular se suele usar la parametrización de MFCC (es la misma que la que se usaba para la creación de los modelos fonéticos) o la SDC.
2. *Modelar los parámetros de entrada para cada idioma* como una mezcla de gaussianas multidimensionales por cada idioma. Cada vector de información O_t para $t=\{1...T\}$ y dado un modelo λ , la probabilidad de observación de dicha secuencia de parámetros vendrá dada por una mezcla de múltiples gaussianas:

$$p(O_t | \lambda) = \sum N_m(O_t; \mu_m, \Sigma_m)$$

Donde μ y Σ_m son respectivamente la media y la matriz de covarianza de la gaussiana m ; λ es modelo de parámetros $\lambda = \{ \omega_m, \mu_m, \Sigma_m \}$. Con los parámetros del modelado se realizan varias repeticiones del algoritmo EM (algoritmo de esperanza-maximización para encontrar estimadores de máxima verosimilitud). Dada una probabilidad de observación de la secuencia O_t para una componente gaussiana m , los parámetros se obtendrían con las fórmulas:

$$\bar{\omega}_m = \frac{1}{T} \sum_{t=1}^T p(O_t | \lambda)$$

$$\bar{\mu}_m = \frac{\sum_{t=1}^T p(O_t | \lambda) \cdot O_t}{\sum_{t=1}^T p(O_t | \lambda)}$$

$$\bar{\Sigma}_m = \frac{\sum_{t=1}^T p(O_t | \lambda) \cdot (O_t - \bar{\mu}_m) \cdot (O_t - \bar{\mu}_m)^T}{\sum_{t=1}^T p(O_t | \lambda)}$$

3. **Reconocer**, se determina la probabilidad de que los vectores acústicos de la voz a clasificar hayan sido generados por el GMM de cada uno de los idiomas, seleccionando aquel que muestre un valor más alto. También se puede detectar la presencia de un idioma comparando con el UBM.

Las ventajas de esta técnica están en su relativa sencillez, así como en que no requiere que las locuciones estén etiquetadas fonéticamente, ya que se puede considerar como un HMM con sólo un estado. Su principal limitación es que modela únicamente los vectores de parámetros considerándolos de forma independiente e independientemente del fonema del que provengan: se modela únicamente información puramente acústica y a muy corto plazo (cada vector de parámetros se obtiene típicamente a partir de sólo 25 ms. de voz).

Debido a estas limitaciones el sistema no era competitivo con respecto a los PPRLM y por ello en la evaluación de 1996 estos sistemas estaban claramente por debajo. Pero la inclusión de un nuevo tipo de parametrización como era la de SDC de P.A. Torres-Carrasquillo [4], que introducía una mayor cantidad de información al expandir el tiempo de la ventana de cálculo de los parámetros, hizo que fuesen competitivos e incluso superasen a los PPRLM en las evaluaciones posteriores.

2.10.4.2. Sistemas SVMs (Support Vector Machines)

Las máquinas de vectores soporte (SVMs) son herramientas discriminativas genéricas de clasificación de patrones, que en los últimos años se ha demostrado que pueden ser muy potentes, por ejemplo, en reconocimiento de locutores. Extrapolando la experiencia en reconocimiento de locutores, se aplican a la problemática de detección de idioma obteniendo resultados bastante competitivos tanto respecto a las técnicas PPRLM como a las GMM.

La técnica consiste en partir de una serie de puntos que representan los vectores de parámetros del idioma a reconocer y de los idiomas impostores. Lo primero que se realiza es trabajar en un espacio de dimensión mayor, una vez en dicho espacio, se calcula el hiperplano que separa mejor los dos grupos impostores y legítimos.

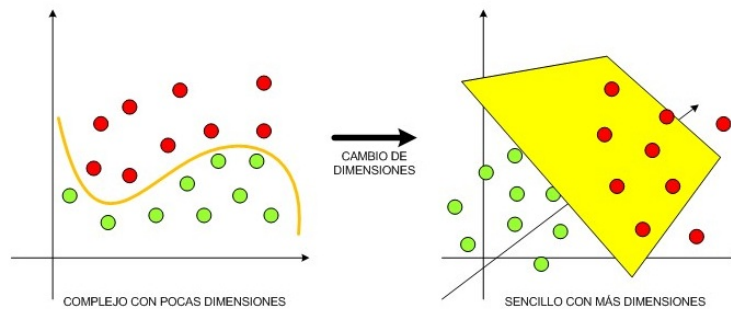


Ilustración 26. Discriminación de características usando la clasificación SVM.

2.10.4.3. Sistemas de reconocimiento fonético: PRLM, PPRLM y PPR

Estas técnicas se asientan en la combinación del reconocimiento fonético, basado en modelos ocultos de Markov, y el modelado estadístico del lenguaje, conjunto de fonemas y frecuencias de aparición.

- *PRLM* es una técnica en la que se usa un modelo estadístico de lenguaje, habitualmente un n-grama, de las secuencias de fonemas reconocidas con mayor probabilidad por un único reconocedor fonético, pudiendo ser este del mismo idioma o distinto, para reconocer al idioma. La identificación del idioma consiste en determinar el modelo de lenguaje que habría generado la secuencia de fonemas reconocida con mayor probabilidad, para ello se aplica el reconocedor de fonemas a la locución.

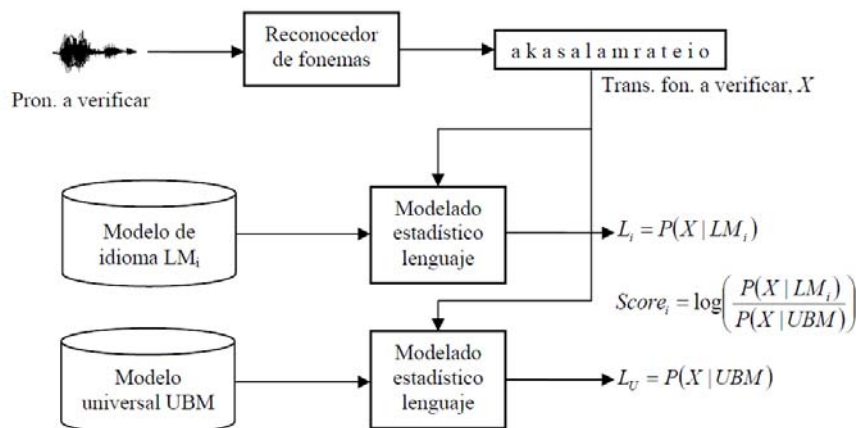


Ilustración 27. Esquema PRLM de verificación de idioma.

- *PPRLM* es una técnica extendida de la anterior. El sistema dispondrá de varios reconocedores fonéticos, correspondientes a distintos idiomas, de cada uno de ellos se obtiene una probabilidad o puntuación. La decisión final resulta de combinar distintas las puntuaciones obtenidas.
- *PPR* consiste en aplicar la locución a un reconocedor que combina HMMs fonéticos y modelos de lenguaje por cada locutor a reconocer. De esta forma se combina el reconocimiento fonético y el modelo de lenguaje, a diferencia de lo que se hacía en las técnicas anteriores donde se aplicaban secuencialmente existiendo un desacoplamiento total.

PPRLM supera ampliamente a cualquier otra técnica. En las evaluaciones de NIST, 2003 y 2005, los sistemas basados en GMMs y SVMs se acercaron bastante a los resultados obtenidos por PPRLM. Esta mejora protagonizada por los sistemas acústicos se debió en gran medida a la cantidad de avances llevados a cabo en diversos campos, entre ellos destacaron la implementación de un nuevo tipo de parametrización conocida como SDC (Shifted Delta Cepstral) [4]. Esta parametrización amplía la ventana de tiempo donde se calculan los parámetros, trabajando con vectores de características cuya información temporal es mucho mayor.

Los progresos en sistemas como GMM [8], SVM o SuperVector han obligado a la comunidad científica a desarrollar sensibles mejoras para mantener a PPRLM en el estado del arte. Algunas de estas mejoras son la extracción de más información a nivel fonético (lattices) [12], o el empleo de SVM como criterio de decisión en lugar de comparar probabilidades [13].

Todos los sistemas y algoritmos descritos anteriormente (salvo los que expresamente se han desarrollado para este proyecto), han presentado sus resultados en las diversas evaluaciones anuales organizadas por el instituto NIST (National Institute of Standards and Technology) , tanto en reconocimiento de locutor [14], como en identificación de idioma [15].

3

Diseño y Desarrollo

3.1. Introducción

El objeto de este apartado es servir de explicación del software utilizado en este proyecto. También se explicará de forma concisa los medios disponibles para su realización.

La finalidad requerida de la identificación de información multimedia en vídeos con carácter turístico ha requerido la implementación de tres sistemas. Dos sistemas han perseguido la meta de la segmentación de audio e identificación de locutores. El tercero se ha diseñado para la identificación de idioma en los segmentos de voz detectados anteriormente. Más en concreto, se han desarrollado:

- ***Sistema de Segmentación de Audio y Diarización de Locutores.***

El objetivo es conseguir separar el audio disponible en la base de datos de vídeos MA2VICMR [18], en segmentos hablados por distintos locutores o segmentos de silencio o música.

Para la tarea de segmentación se ha decidido usar dos herramientas de segmentación, con el fin de combinarlos para obtener mejores resultados.

- La primera herramienta que se ha utilizado es el Sistema de "Segmentación de Audio ATVS-UAM" [1], con la que se separa el audio disponible en segmentos de voz y no voz.
- La segunda herramienta es el paquete HTK [17] con la que se ha realizado un reconocimiento fonético para detectar segmentos de voz frente a los de no voz.

▪ **Sistema de Reconocimiento de Idioma.**

El sistema de reconocimiento de idioma permite identificar el idioma, basándose en un reconocimiento fonético adaptado a los idiomas que existen en la base de datos. Se decidirá entre español o inglés en base a puntuaciones o scores de acierto.

Para esta tarea se han empleado dos diccionarios fonéticos (uno en español y otro en inglés) adaptados a la voz telefónica hablada.

La siguiente imagen se puede visualizar en una primera aproximación y de una forma clara y sencilla, un posible resultado tras la aplicación en cadena de los sistemas mencionados y que se describen en detalle en el apartado 3.3:



Ilustración 28. Ejemplo de resultados de la aplicación de los sistemas del presente proyecto.

3.2. Medios disponibles

3.2.1. Bases de datos

Es este proyecto se han utilizado tres bases de datos con fines diferentes:

- *Base de datos de vídeos MA2VICMR [18].*

Esta base de datos pertenece al proyecto MA2VICMR (Mejorando el Acceso, el Análisis y la Visibilidad de la Información y los Contenidos Multilingüe y Multimedia en Red para la Comunidad de Madrid), segunda edición del proyecto inicial MAVIR.

El Consorcio MAVIR es una red de investigación co-financiada por la Comunidad de Madrid y el Fondo Social Europeo bajo los programas de I+D en TIC MA2VICMR (2010-2013) y MAVIR (2006-2009) formada por un equipo multidisciplinar de científicos, técnicos, lingüistas y documentalistas para desarrollar un esfuerzo integrador en las áreas de investigación, formación y transferencia de tecnología.

El núcleo del consorcio está formado por siete grupos de investigación de universidades y centros de la Comunidad de Madrid (entre los que se encuentra en grupo de la UAM "*Human Language Technologies & Information Retrieval*") que, desde un perspectiva pluridisciplinar, se complementan en varias dimensiones: mundo académico vs. mundo profesional, investigación vs. oferta de servicios, generación de recursos vs. aplicaciones.

Dicha base de datos está formada por cuarenta y seis (46) vídeos de información turística pertenecientes al proyecto MA2VICMR [18] y que se puede consultar en el Anexo A, cuyo audio está grabado en estéreo (2 canales). El período de muestreo del audio original es de 44,100 KHz con 16 bits/muestra. Existen treinta (30) vídeos en español y en dieciséis (16) vídeos en inglés, identificables en el propio nombre mediante las etiquetas '*_esp*' para español y '*_eng*' para inglés, resumidos en la tabla siguiente:

Alcala_esp.mp4	Caceres_esp.mp4	GaudiGenio_eng.mp4	Salamanca_esp.mp4
Andalucia_eng.mp4	CaminoSantiago_esp.mp4	GaudiGenio_esp.mp4	SantiagoCompostela_esp.mp4
Andalucia_esp.mp4	CastillaLeon_esp.mp4	Ibiza_esp.mp4	Segovia_eng.mp4
Aranjuez_eng.mp4	Cid_esp.mp4	Laguna_esp.mp4	Segovia_esp.mp4
Aranjuez_esp.mp4	Cuenca_esp.mp4	LaMancha_eng.mp4	Toledo_eng.mp4
Avila_esp.mp4	Dalisurre_eng.mp4	LaMancha_esp.mp4	Toledo_esp.mp4
Baeza_eng.mp4	Dalisurre_esp.mp4	Madrid_eng.mp4	Ubeda_eng.mp4
Baeza_esp.mp4	Escorial_esp.mp4	Madrid_esp.mp4	Ubeda_esp.mp4
Bcngotico_eng.mp4	Extremadura_eng.mp4	Montjuic_eng.mp4	Valencia_eng.mp4
Bcngotico_esp.mp4	Extremadura_esp.mp4	Montjuic_esp.mp4	Valencia_esp.mp4
Bcnmodernista_eng.mp4	Gaudi_eng.mp4	PueblosEdadMedia_esp.mp4	
Bcnmodernista_esp.mp4	Gaudi_esp.mp4	RomanicoAragon_esp.mp4	

Tabla 4. Vídeos de la base de datos corpus MA2VICMR.

Cada video está acompañado con su respectivo archivo etiquetado con información de la transcripción fonética, transcripción prosódica, calidad de grabación, duración, número y sexo de los locutores que intervienen, palabras clave, número de palabras, volumen de la música. A continuación se detalla un extracto de estos archivos.

```

Andalucia_esp.txt
xmlns:uam-ATVS="http://www.mavir.net/uam/atvs"
xmlns:uam-ir="http://www.mavir.net/uam/ir"
<Video id="R801_858c" >
  <uam-lli:Transcription mode="manual" <!-- mode puede ser "manual", "automatic" o "any" -->
    flow="output"> <!-- flow puede ser "input" o "output" -->
  </uam-lli:Transcription >
  <HEADER>
    @Title: Andalucía, un recorrido por la historia y el arte
    @Topic: Andalucía history and art
    @Keywords: Andalucía, Muslims, Alhambra
    @Language: Spanish
    @Length: 2'30''
    @Words: 213
    @Participants: 1
    @Sex: male
    @Acoustic quality: high
    @Music: high
  </HEADER>
  <TEXT>
    <Prosodic>
      <UNIT speaker="LOC" startTime="10.982" endTime="30.218"> Andalucía / está poblada desde tiempos prehistóricos / y todas
      las civilizaciones / han dejado aquí una parte de su historia // hay vestigios fenicios y romanos / pero es sin duda la huella de la época musulmana / la
      que no tiene parangón en ningún otro país de Europa //</UNIT>
      <UNIT audio_element="MUS" startTime="30.218" endTime="40.365">
      </UNIT>
      <UNIT speaker="LOC" startTime="40.365" endTime="60.192"> ocho siglos de presencia árabe en Andalucía / han dejado
      monumentos como la mezquita de Córdoba / que fue la capital del Califato / y ciudad de convivencia entre musulmanes / cristianos y judíos // en Sevilla /
      la giralda / enseña el poder almonade //</UNIT>
      <UNIT audio_element="MUS" startTime="60.192" endTime="64.974">
      </UNIT>
      <UNIT speaker="LOC" startTime="64.974" endTime="74.173"> Granada / fue el último reino musulmán de la península / y la
      Alhambra y sus jardines / la mejor muestra de su esplendor //</UNIT>
      <UNIT audio_element="MUS" startTime="74.173" endTime="80.022">
      </UNIT>
    </Prosodic>
  </TEXT>
</Video>

```

Ilustración 29. Ejemplo de formato de las transcripciones originales.

Esta base de datos será la principal fuente de procesamiento de datos a la que se aplicarán los sistemas de segmentación de audio y diarización de locutores e identificación de idioma desarrollados en el presente proyecto y que se explican más adelante.

- Base de datos para la utilización del reconocedor fonético en la segmentación de audio y reconocimiento de locutores.

Esta base de datos ha sido generada para la evaluación de Albayzin 2010 [1]. Está diseñada en base a una cadena de estados correspondiente a modelos ocultos de Markov. En concreto está formada por un GMM-MMI de 5 estados caracterizado por:

- ✓ Un HMM generado durante el entrenamiento, caracterizado por la matriz de probabilidades de transición entre estados en base a un entrenamiento con ficheros de audio de dicha evaluación. La cantidad de ficheros de audio utilizada para el entrenamiento ha sido alta, y se puede considerar lo suficientemente representativa del habla.

```

hmmdefsPhones
~0
<STREAMINFO> 1 39
<VECSIZE> 39-NULLD<<MFCC_D_A_0><DIAGC>
-t "T A"
<TRANS> 5
0.000000e+00 1.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 5.708905e-01 4.291095e-01 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 6.582302e-01 3.417698e-01 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 5.745838e-01 4.254162e-01
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
-t "T B"
<TRANS> 5
0.000000e+00 1.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 3.868027e-01 6.131973e-01 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 4.491720e-01 5.508280e-01 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 4.492647e-01 5.507353e-01
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
-t "T D"
<TRANS> 5
0.000000e+00 1.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 3.936498e-01 6.063502e-01 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 4.572273e-01 5.427727e-01 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 4.963998e-01 5.036002e-01
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
-t "T E"
<TRANS> 5
0.000000e+00 1.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 5.803970e-01 4.196031e-01 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 5.887566e-01 4.112435e-01 0.000000e+00

```

Ilustración 30. Probabilidades de transición del HMM de cinco estados generado para la evaluación de Albayzin 2010.

- ✓ Un diccionario fonético de habla en español. Lista ordenada de los fonemas de uso en la lengua de grabación de los vídeos, aunque se usará en un principio para detectar tan sólo segmentos de voz sin interesar el idioma propio.
- ✓ Una lista de trifenemas generada durante el entrenamiento del HMM. En ella se encuentran las mezclas de fonemas que mejor representan los sonidos trifenéticos de la lengua.

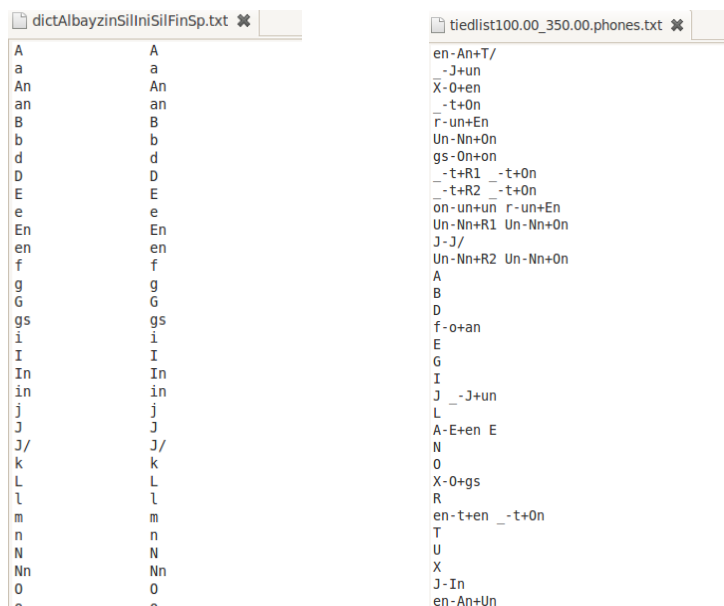


Ilustración 31. Diccionario fonético de habla en español y listado de trifenemas generados para la evaluación de Albayzin 2010.

Esta base de datos tiene como fin servir de herramienta para la identificación de segmentos de habla / no habla y no la del reconocimiento de idioma, para la que se usará un reconocedor fonético y diccionarios adaptados a los dos idiomas.

▪ Base de datos para la identificación de idioma:

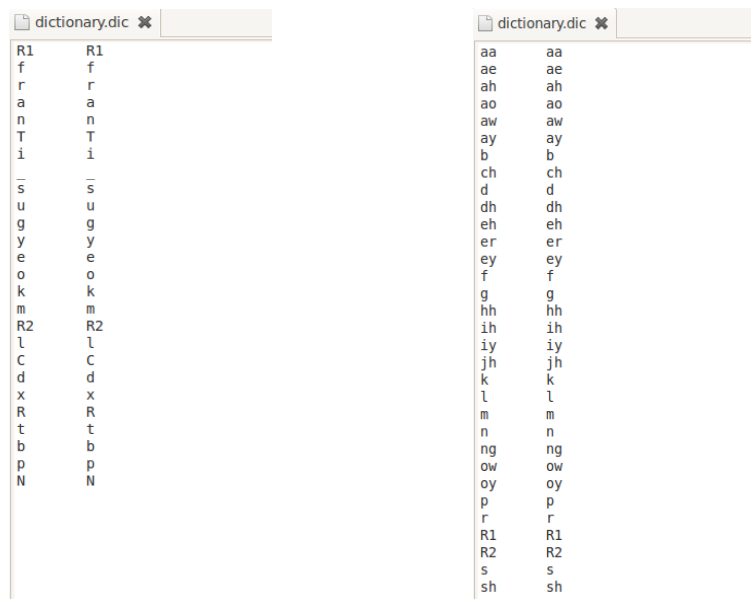
Esta base de datos desarrollada bajo el software de parametrización AURORA, está compuesta por 4 HMMs (GMM de 1 a 4 gaussianas) entrenados con voz telefónica a 8 KHz, tanto para español como para inglés. Por lo tanto se pueden diferenciar:

- ✓ Un HMM de una (1) mezcla de Gaussiana a 8 KHz para idioma español.
- ✓ Un HMM de dos (2) mezclas de Gaussiana a 8 KHz para idioma español.
- ✓ Un HMM de tres (3) mezclas de Gaussiana a 8 KHz para idioma español.
- ✓ Un HMM de cuatro (4) mezclas de Gaussiana a 8 KHz para idioma español.

- ✓ Un HMM de una (1) mezcla de Gaussiana a 8 KHz para idioma inglés.
- ✓ Un HMM de dos (2) mezclas de Gaussiana a 8 KHz para idioma inglés.
- ✓ Un HMM de tres (3) mezclas de Gaussiana a 8 KHz para idioma inglés.
- ✓ Un HMM de cuatro (4) mezclas de Gaussiana a 8 KHz para idioma inglés.

- ✓ Un diccionario fonético de habla en español, con la lista de fonemas existentes.
- ✓ Un diccionario fonético de habla en inglés, con la lista de fonemas existentes.

- ✓ Una lista de fonemas en español, con la lista de fonemas más frecuentes.
- ✓ Una lista de fonemas en inglés, con la lista de fonemas más frecuentes.



The image shows two side-by-side screenshots of a text editor window titled 'dictionary.dic'. The left window displays a list of Spanish phonemes, and the right window displays a list of English phonemes. Both lists are organized into two columns, with the first column representing the phoneme and the second column representing its frequency or classification.

Spanish Phonemes	English Phonemes
R1	aa
f	ae
r	ah
a	ao
n	aw
T	ay
i	b
—	ch
s	d
u	dh
g	eh
y	er
e	ey
o	f
k	g
m	hh
R2	ih
l	iy
C	jh
d	k
x	l
R	m
t	n
b	ng
p	ow
N	oy
	p
	r
	R1
	R2
	s
	sh

Ilustración 32. Diccionario fonético de habla en español e inglés respectivamente.

Esta base de datos tiene como fin servir de representación estadística de los fonemas más frecuentes de cada idioma, español o inglés. Con ella se identificará si un audio es más probable que esté pronunciado en uno u otro idioma debido a las puntuaciones o scores obtenidos en la evaluación de reconocimiento fonético con cada idioma.

3.2.2. Software

El sistema operativo empleado en la realización del presente proyecto ha sido Ubuntu 10.04, distribución basada en Debian GNU/Linux.

Como herramientas de desarrollo de software que se han utilizado en este proyecto han sido los lenguajes de programación Perl, Bash y C para generar un conjunto de scripts, permitiendo una ejecución ordenada de los diferentes sistemas.

El principal software de trabajo ha sido el conjunto de herramientas HTK (Hidden Markov Model Toolkit) [17], usado para la construcción y manipulación de los modelos ocultos de Markov.

HTK fue usado en principio en aplicaciones de reconocimiento de voz, aunque se ha encontrado otras muchas aplicaciones como la síntesis de voz o secuencias de ADN. HTK consiste en un conjunto de librerías y herramientas desarrolladas en C, cuya sofisticación y optimización facilitan el análisis de la voz, el entrenamiento de los HMM, la evaluación y extracción de resultados. El software soporta la creación de HMM con distribuciones continuas de mezclas de Gaussianas o por medio de distribuciones discretas pudiendo crear de esta forma complejos sistemas de HMM.

HTK fue desarrollado originalmente en el laboratorio de la inteligencia de máquinas (conocido antes como el grupo “the Speech Vision and Robotics”) del departamento de ingeniería de la Universidad de Cambridge (CUED) donde se ha utilizado para construir grandes sistemas de reconocimiento de habla.

En 1993 Entropic Research Laboratory Inc. adquirió los derechos de vender HTK y el desarrollo de HTK fue transferido completamente a Entropic en 1995 en que el laboratorio de investigación de Entropic Cambridge Ltd fue establecido. HTK fue vendido por Entropic hasta que en 1999 Microsoft compró Entropic. Microsoft ahora ha licenciado HTK de nuevo a CUED y está proporcionando la ayuda de modo que CUED pueda redistribuir HTK.

Por último se ha utilizado el software multiplataforma "*wavesurfer*" [20], un conjunto de herramientas en código abierto para el análisis visual de habla y sonidos. Entre otras muchas cosas, con él se pueden representar ficheros de audio y analizar transcripciones en formato HTK.

3.2.3. Hardware

El hardware empleado en el desarrollo del presente proyecto ha sido un ordenador de sobremesa, cuya CPU se compone de un procesador Intel Pentium IV a 1.8 GHz y 1024 MB de RAM, un monitor TFT de 17", teclado, ratón y altavoces.

Además, ha estado disponible la red interna del laboratorio formada por ordenadores personales así como por varios servidores de almacenamiento y ejecución.

Todos estos medios han sido suministrados por el grupo de trabajo ATVS de la Universidad Autónoma de Madrid (UAM).

3.3. Sistemas desarrollados

3.3.1 Preparación de los datos

La filosofía de trabajo ha sido la de abordar el desarrollo del proyecto mediante la ejecución ordenada de pequeñas tareas específicas, dividiendo en partes pequeñas y realizables de forma sencilla y eficiente.

Como se ha mostrado anteriormente, la base de datos disponible contiene vídeos en formato mp4, y audio grabado en estéreo a una frecuencia de 44.100 Hz. Los sistemas de segmentación de audio y reconocimiento de locutor son bastante sensibles a este tipo de características.

Si bien, de cuanto mayor información en frecuencia se dispone, de mayor calidad del audio se disfruta, a la hora del tratamiento de datos y computación, este hecho influye notablemente en los tiempos de ejecución, siendo mayor cuanto más información hay.

Se ha comprobado que es suficiente disponer de grabaciones de audio a una frecuencia de 8 KHz (un poco por debajo de la mitad de la frecuencia máxima audible por el ser humano) o mejor a 16 KHz, con 16 bits por muestra y con tan sólo un canal de grabación (mono) para poder desarrollar sistemas fiables y con los que se obtengan buenos resultados a un tiempo de computación razonable.

Por ello, lo primero que se ha realizado es la extracción del audio a 16 KHz, con 16 bits por muestra con codificación PCM *little-endian* y un solo canal en formato *wav* (forma de onda sin compresión ni cabeceras), pues los sistemas utilizados se encuentran diseñados con estas características.

Tras la preparación inicial de los datos, ha sido necesario entender y adaptar dichos audios a los dos sistemas de segmentación de audio.

Por ejemplo, según lo expresado en el artículo de Javier Franco Pedroso, Ignacio López-Moreno, Doroteo T. Toledano [1], el sistema de Segmentación de audio ATVS-UAM es capaz de segmentar y etiquetar eficaz y eficientemente los ficheros de audio, si éstos se dan como entrada al sistema en *slices* o ventanas de sesenta (60) segundos de duración, con un solapamiento de dos (2) segundos entre ellos.

De la misma forma, el Reconocedor fonético es algo más estricto en este sentido y se requieren *slices* de cuatro (4) segundos con un solapamiento de dos (2) segundos como archivos de entrada al sistema, pues trabaja a nivel de fonema pudiéndose reconocer varios fonemas por segundo de audio.

Analizando de forma visual la base de datos del proyecto MA2VICMR [18], se observó que el audio de cada video tan sólo contenía información turística hablada por un único locutor cuyo idioma no cambiaba y estaba acompañado de música instrumental de fondo durante toda la grabación. Estas premisas han sido claves para tomar decisiones a lo largo de la realización del proyecto.

3.3.2 Sistemas de Segmentación de Audio y Diarización de Locutores

A partir de estas observaciones, la tarea siguiente se limita a separar los segmentos de habla en cualquiera de los idiomas posibles, del resto de segmentos que puedan existir (identificables como música, ruido, silencios, etc.).

La siguiente imagen muestra de una forma visual, la idea de reconocimiento implementada mediante los dos sistemas anteriormente mencionados y que se detallan a continuación:

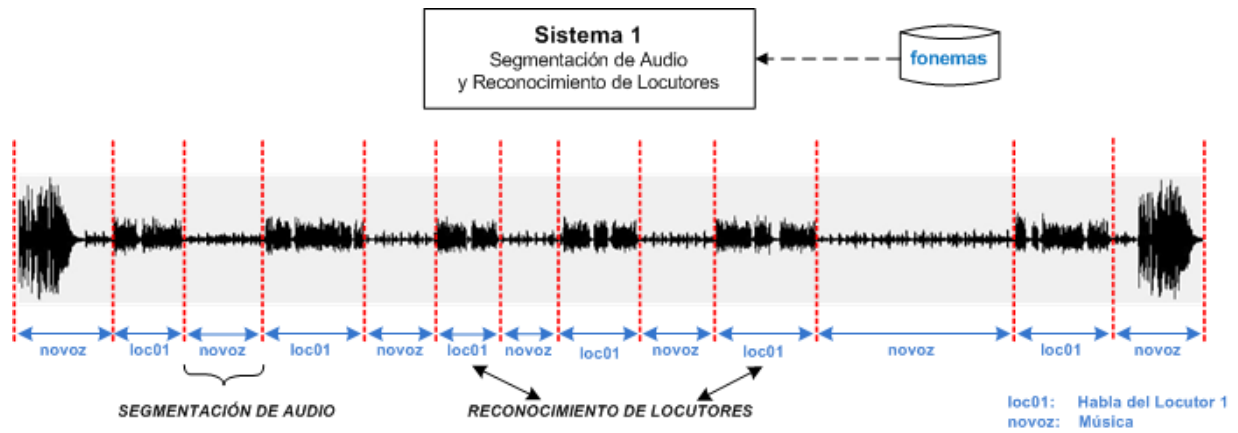


Ilustración 33. Imagen del Sistema de Segmentación de Audio y Diarización de Locutores.

3.3.2.1. Segmentación de Audio ATVS-UAM

El sistema de Segmentación de Audio desarrollado por el grupo de investigación ATVS-UAM bajo las herramientas proporcionadas por el entorno de trabajo de MATLAB, detecta los diferentes segmentos de una grabación de audio y los clasifica en diferentes estados.

Como paso previo a la segmentación es necesario realizar una extracción de características de los ficheros de audio. Estas características se consiguen usando el "parametrizador disponible en el grupo de investigación ATVS", y que a partir de un fichero de audio *wav*, se obtiene archivos de parámetros MFCC con formato FPG.

Una vez se han obtenido dichos parámetros, en el sistema de segmentación se usa un HMM formado por 1024 mezclas de gaussianas (GMM) el cual se ha optimizado con hasta dieciocho (18) iteraciones MMI (Maximun Mutual Information). Dicho HMM está compuesto por cinco (5) GMMs cada uno de ellos correspondiente a un estado posible de los segmentos.

Inicialmente, los estados en los que se puede clasificar un segmento son:

Iniciales del estado	Significado del estado	
	Inglés	Español
SP	SPeech	Voz
SM	Spech & Music	Voz y música
MU	MUsic	Música
SN	Speech & Noise	Ruido
OT	Other	Otro

Tabla 5. Estados del HMM del sistema de Segmentación de Audio ATVS-UAM.

Este sistema se ha readaptado de cara a obtener una mejora computacional, tanto en ejecución como en cantidad de datos iniciales. Las probabilidades de estancia inicial, matrices de probabilidades de salto de estado, matrices de covarianza, pesos, transiciones disminuyen notablemente en cuanto a volumen de datos, ya que existen estados que se pueden clasificar bajo un mismo estado debido a los fines que se han propuesto. Por ello, se ha adaptado el sistema de segmentación ATVS-UAM inicial formado por un HMM con cinco (5) estados, en un HMM de dos (2) estados, siendo estos dos estados *voz y no voz*, según se muestra en la imagen siguiente.

Debido a las características de las grabaciones, es imposible aislar los segmentos de voz de la música, pues ésta se encuentra presente de forma ininterrumpida durante todo los audios (siendo posible una mejora realizando un filtrado en alta frecuencia). Se puede observar una bajada del volumen cuando el locutor habla, pero ésta se encuentra presente como música de fondo. Durante el resto del proyecto se hablará de segmentos de voz cuando se identifiquen segmentos "voz y música" por el sistema de segmentación original.

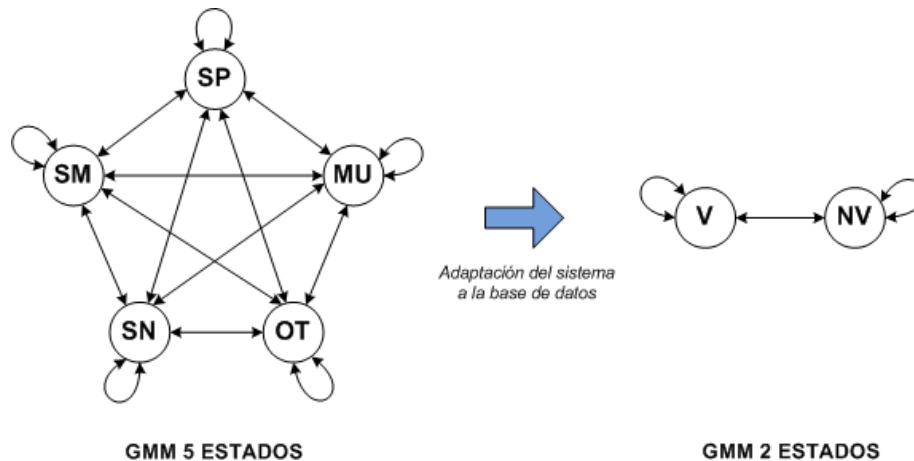


Ilustración 34. Adaptación del HMM de 5 estados a un HMM de 2 estados.

Estas sencillas y necesarias decisiones permiten sacar más conclusiones y facilitar el trabajo de reconocimiento de locutores. Observando los audios extraídos, (mas tarde se comprobará con su correspondiente etiquetado o transcripción disponible), se puede detectar que en todos los ficheros tan sólo interviene un único locutor y en un único idioma. Como el fin no es identificar la identidad del locutor, sino identificar qué segmentos de voz pertenecen a un locutor y cuales pertenecen a otro, esta decisión es sencilla. A estos segmentos de voz se les ha identificado con la etiqueta "loc01", en referencia al primer locutor que aparece.

El método de almacenamiento utilizado para los resultados intermedios ha sido mediante archivos de texto plano, manteniendo una estructura lógica sencilla durante todo el proyecto. Se han diferenciado cuatro columnas, título del video, inicio del segmento, final del segmento y clasificación del segmento. Ello permitirá obtener gráficas y estadísticas de una forma rápida y sencilla más adelante. Un ejemplo de archivo de almacenamiento de datos se muestra a continuación:

Andalucia_esp	Inicio	Final	Clasificación
Andalucia_esp	0.00	3.00	novoz
Andalucia_esp	3.00	5.00	novoz
Andalucia_esp	5.00	7.00	novoz
Andalucia_esp	7.00	9.00	novoz
Andalucia_esp	9.00	11.00	novoz
Andalucia_esp	11.00	13.00	loc01
Andalucia_esp	13.00	15.00	loc01
Andalucia_esp	15.00	17.00	loc01
Andalucia_esp	17.00	19.00	loc01
Andalucia_esp	19.00	21.00	loc01
Andalucia_esp	21.00	23.00	loc01
Andalucia_esp	23.00	25.00	loc01
Andalucia_esp	25.00	27.00	loc01
Andalucia_esp	27.00	29.00	loc01
Andalucia_esp	29.00	31.00	novoz
Andalucia_esp	31.00	33.00	novoz

Ilustración 35. Ejemplo de almacenamiento de datos en archivo de texto plano.

3.3.2.2. Reconocimiento fonético

Un sistema basado en reconocimiento fonético se centra en el análisis de la información de la señal de voz, atendiendo a la unidad mínima de formación del habla: el fonema. Un fonema es la parte más esencial de una lengua, permitiendo establecer diferencias y dotando de significado a las palabras.

En este proyecto se ha utilizado un reconocedor fonético con una finalidad derivada del uso normal del sistema. Con él se pretende reconocer los fonemas que existen en una grabación de audio determinada para realizar un análisis de velocidad de pronunciación de fonemas.

Los audios disponibles pertenecen a una serie de capítulos en los que se muestran los lugares de especial interés turístico de una determinada zona, las grabaciones están realizadas a una velocidad que permite la normal asimilación de la información. En concreto, la velocidad normal del habla está en torno a los quince (15) fonemas por segundo con un margen por encima y por debajo de cinco (5) fonemas por segundo. Éstos serán nuestro umbrales a la hora de realizar la identificación de tramos de frente a tramos de no voz, pues tanto el ruido, como la música instrumental son identificados por el sistema con velocidades muy inferiores ó superiores de estos umbrales.

Como se ha comentado anteriormente, el propio reconocimiento se realiza con las herramientas y facilidades disponibles en el software de reconocimiento de audio HTK [17]. En la siguiente imagen se muestra el esquema de funcionamiento del software mencionado. Los distintos programas para la preparación de datos, entrenamiento, testeo y el análisis final comienzan por la 'letra H' y se encuentran recuadrados. El resto son ficheros de entrada necesarios para el funcionamiento, o datos de salida con resultados valorables.

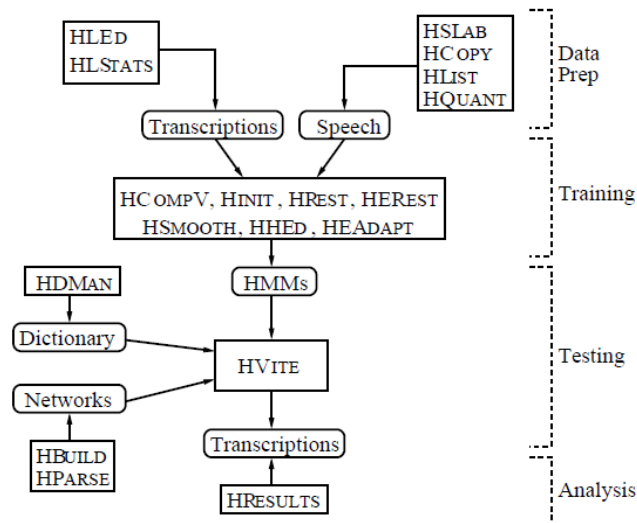


Ilustración 36. Esquema de procesamiento de audio y herramientas HTK.

En concreto para este caso, se ha utilizado la herramienta *HVite*. La principal funcionalidad es la decodificación mediante el algoritmo de Viterbi, cuando se proporciona un HMM previamente entrenado con audio lo suficientemente representativo en la lengua adecuada, un diccionario fonético y unas reglas gramaticales propias de la de la lengua a reconocer.

La herramienta *HVite* se ha configurado para el reconocimiento fonético para el uso de coeficientes MFCC, Δ MFCC y $\Delta\Delta$ MFCC incluido el coeficiente inicial C_0 . Además los archivos de entrada están en formato *wav* sin cabeceras, con un período de $625 \cdot 10^{-7}$ seg (o bien 16KHz). También se ha utilizado una ventana de *Hamming* de 25 ms de duración y un filtro de pre-énfasis de coeficiente 0.97. Por último, un banco de filtros de Mel compuesto por veintidós (22) filtros. En concreto, estas variables han sido:

```

FORCECTEXP = T
ALLOWXRDEXP = T
SOURCEKIND = WAV
TARGETKIND = MFCC_D_A_0
SOURCEFORMAT = WAV
NATURALREADORDER = T
NATURALWRITEORDER = T
SOURCERATE = 625
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
ENORMALISE = T
    
```

Tabla 6. Valores de configuración de *HVite*.

Se ha podido aplicar directamente el reconocimiento a este nivel porque se disponía de modelos de HMM previamente entrenados con gran cantidad de audio telefónico (aunque no con la base de datos utilizada pues la cantidad de datos existente podría ser escasa), un diccionario en castellano disponible y un esquema de redes de funcionamiento del idioma (gramática) de la evaluación de Albayzin 2010.

Como se puede comprobar, este reconocimiento es en exclusiva en castellano y se ha aplicado a audios tanto en castellano como el inglés. Se ha decidido aplicar a audios de ambos idiomas debido a la proximidad de fonemas existente entre ellos y de cara a un reconocimiento a nivel fonético, la utilización de un sistema adaptado a los dos idiomas no proporcionaría grandes cambios.

La ejecución de esta herramienta deporta unos ficheros de reconocimiento con la extensión *'mfl'*, en los que se detallan: el fonema reconocido, acompañado de los tiempos de inicio y final de segmento en formato HTK (10^7 seg), y un valor que corresponde al *score* o puntuación conforme a probabilidades y cálculos propios del software para cada *slice* de audio definido en la fase de preparación de datos.

```

Andalucia_esp_recout_adapt.train.mfl
#IMLF#
"../../../../Lattices_Adapt/Andalucia_esp/Andalucia_esp_slice-0001_start-0_end-4.rec"
200000 3800000 a -2755.827148
3900000 39600000 R -25516.458984
.
"../../../../Lattices_Adapt/Andalucia_esp/Andalucia_esp_slice-0002_start-2_end-6.rec"
200000 1600000 a -1096.774902
1700000 32200000 R -20866.611328
32200000 37200000 R -3224.639648
37300000 39600000 n -1461.567993
.
"../../../../Lattices_Adapt/Andalucia_esp/Andalucia_esp_slice-0003_start-4_end-8.rec"
200000 12200000 R -8065.996582
12200000 17200000 R -3224.639648
17300000 20600000 n -2166.416260
20700000 30300000 o -4480.891602
30400000 32400000 t -1276.584351
32500000 33400000 a -582.222168
33500000 35600000 s -1456.663696
35700000 39600000 s -2608.079834
.
"../../../../Lattices_Adapt/Andalucia_esp/Andalucia_esp_slice-0004_start-6_end-10.rec"
200000 10300000 o -4893.840820
10400000 12400000 t -1276.584351
12500000 13400000 a -582.222168
13500000 15600000 s -1456.663696
15700000 21400000 s -3862.159668
21500000 39600000 n -12064.046875
    
```

Ilustración 37. Ejemplo de fichero tras reconocimiento *HVite* con extensión *'mfl'*.

Tras obtener este tipo de resultados, se han tratado estos ficheros para adaptarlos al formato común establecido en la fase de diseño.

3.3.2.3. Reconocimiento fonético con adaptación al locutor

Como mejora al sistema de reconocimiento fonético aplicado a la segmentación de audio y reconocimiento de locutores se ha desarrollado una readaptación de los modelos fonéticos a los propios locutores que intervienen en la base de datos. Este hecho permite obtener un nuevo modelo de HMM adaptado a las características de la voz del propio locutor. Con ello se podría obtener una mejora en el reconocimiento fonético, siempre y cuando los datos de entrenamiento sean lo suficientemente grandes y representativos como para identificar a un locutores entre varios.

Esta adaptación se ha realizado usando la herramienta *HMMAdapt* del Toolkit HTK. El método de adaptación ha sido MLLR manteniendo el resto de características utilizadas anteriormente. Tras esta adaptación, se ha vuelto a realizar el reconocimiento fonético usando de nuevo la herramienta *HVite*, pero esta vez con los HMMs reentrenados para tal efecto.

3.3.2.4. Combinación de resultados de Segmentación de Audio y Reconocimiento fonético

La idea principal es la de combinar los resultados que se han obtenido con ambos sistemas de segmentación y reconocimiento, mediante las sencillas funciones lógicas AND y OR. Esta combinación se puede llevar a cabo gracias a que los resultados se reflejan en segmentos temporales, los cuales pueden diferir en cuanto a duración, inicio o final con respecto a los segmentos reales del audio.

Con la combinación de estas funciones lógicas, se puede ampliar o reducir la duración de los segmentos reconocidos para adaptarse a los reales, cubriendo de esta forma los excesos o defectos de los sistemas de reconocimiento. A continuación se detalla la tabla de conversión de ambas funciones. En ellas se pueden ver los dos tipos de segmentos que el reconocedor ofrece: segmentos de voz del locutor 1 (identificados con la etiqueta *loc01*) y segmentos en los que no hay voz, aunque serán segmentos con música de fondo, identificados con la etiqueta *novoz*:

AND			OR		
Sistema A	Sistema B	Resultado	Sistema A	Sistema B	Resultado
novoz	novoz	novoz	novoz	novoz	novoz
novoz	loc01	novoz	novoz	loc01	loc01
loc01	novoz	novoz	loc01	novoz	loc01
loc01	loc01	loc01	loc01	loc01	loc01

Tabla 7. Tablas de aplicación de las funciones lógicas AND y OR.

La elección de funciones no es aleatoria, y su aplicación se hace sobre los segmentos de voz. Con la función AND se intenta seleccionar tan sólo los segmentos en los que ambos sistemas hayan detectado lo mismo. Es una forma de confirmar y asegurar que el segmento así identificado el correcto, restringiendo la duración del segmento.

Por otra parte, con la función OR se consigue el efecto contrario. Esta función selecciona los segmentos en los que ambos sistemas hayan reconocido los mismos datos, siempre limitados por el máximo o mínimo de los segmentos coincidentes.

En la siguiente imagen se muestra una explicación visual de este efecto. Se puede observar cómo el reconocimiento realizado tanto con el sistema de Segmentación de Audio ATVS-UAM, tanto como con el Reconocedor fonético ofrece un etiquetado de segmentos que se aproximan al real, pero con pequeños desplazamientos con respecto a él.

Ello es debido principalmente a dos efectos:

- Las transiciones articulatorias propias de la vocalización humana, en las que durante una conversación normal, no se pronuncia palabras de idéntica duración, sino que se unen los finales de las palabras con los inicios de las siguientes, manteniendo entonación y duración de las vocales o consonantes finales.

- El cambio estados que se han denominado de *novoz*, en los que tan sólo existe música de fondo sin palabras, a estados denominados de *voz*, en los que predominan las palabras del locutor, pero existe la música de fondo que dificulta el reconocimiento.

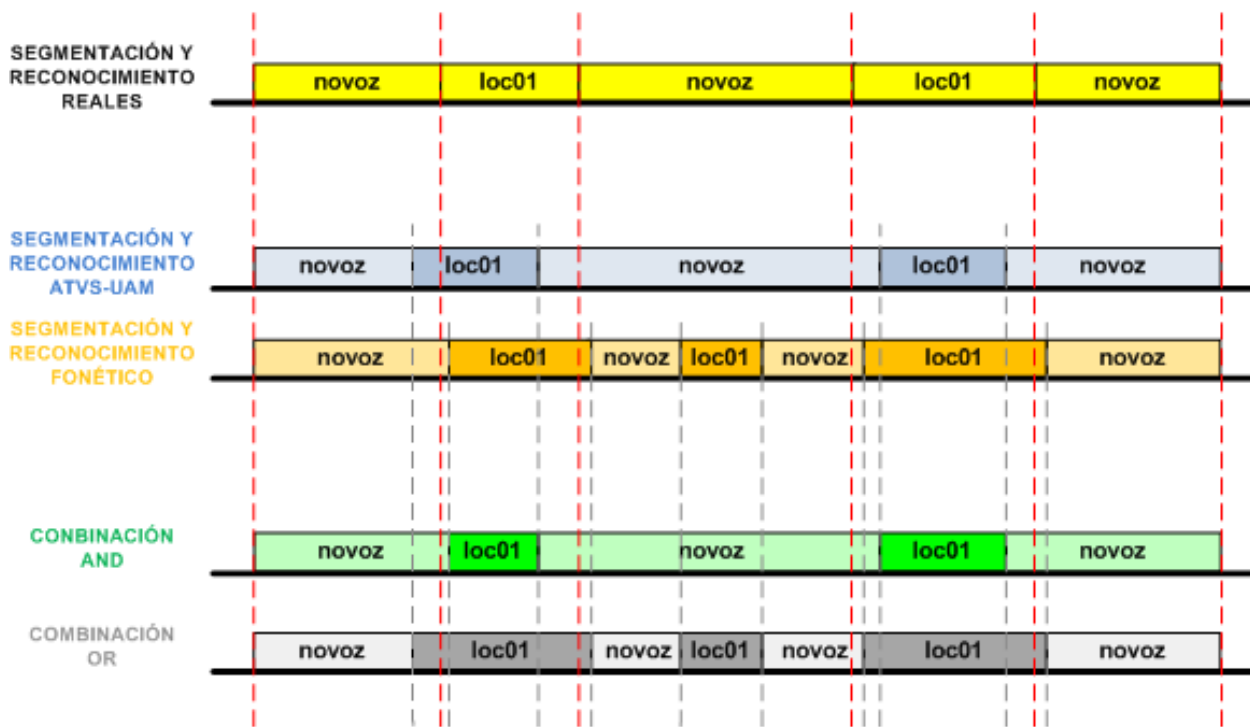


Ilustración 38. Esquema visual de reconocimiento, combinando resultados.

3.3.3 Sistema de Identificación de Idioma

Este sistema se ha desarrollado como mejora con respecto a lo inicialmente propuesto. Con él se pretende añadir a la idea inicial basada en la recuperación de información multimedia de videos, la identificación del idioma de los audios. La siguiente imagen muestra el esquema de funcionamiento del sistema desarrollado:

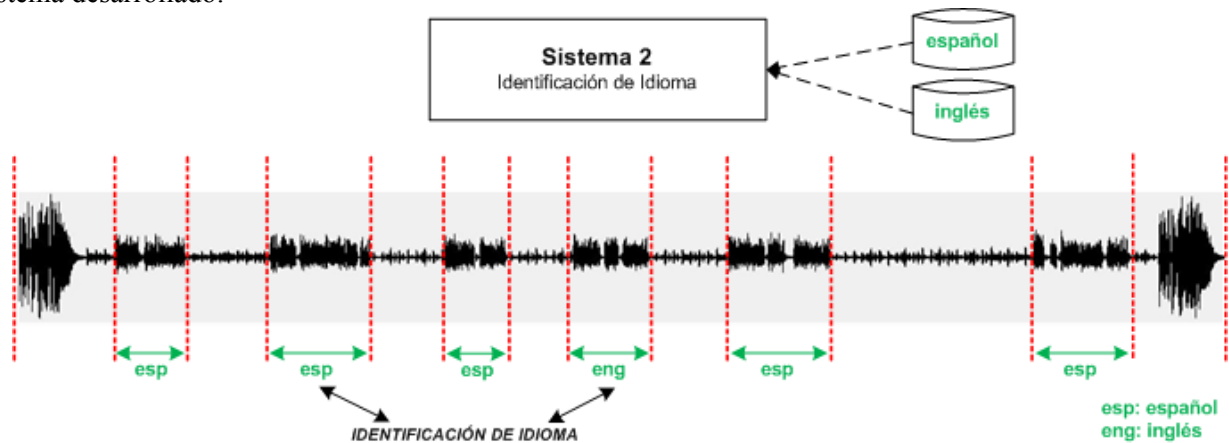


Ilustración 39. Sistema de Identificación de Idioma.

Para ello se ha utilizado un reconocedor fonético de dos idiomas, español e inglés. Para la identificación se disponía de cuatro (4) modelos ocultos de Markov (MHH) por idioma, entrenados con audio generalmente obtenido de habla telefónica. Estos cuatro HMMs corresponden a cuatro modelos diferentes de la base de datos empleando desde una (1) gaussiana hasta cuatro (4) gaussianas para audio a 8 KHz y 16 KHz.

Además se facilitaron otros datos para la identificación como han sido: una lista de fonemas para cada idioma, en los que se encuentra especificados los fonemas básicos de ambos idiomas, inglés y español. Un diccionario fonético en inglés y otro en español, compuesto por trifenemas de uso más frecuente en ambos idiomas. También se facilitó un archivo para cada idioma de la red de funcionamiento de los fonemas, conocido como gramática de la lengua.

Para la implementación de este sistema se ha recurrido al conjunto de herramientas HTK, cuyas facilidades de manejo con HMMs y reconocimiento de voz se han aprovechado.

La identificación del idioma se ha realizado comparando los *scores* o puntuaciones obtenidas al aplicar los dos reconocedores (el de inglés y el de español) a todos los videos. De esta forma, el reconocedor fonético de inglés proporcionará mejores puntuaciones sobre los audios en este idioma, que el reconocedor en español, y viceversa.

A continuación se muestra un diagrama completo del flujo de la información de este proyecto. En él se pueden ver los sistemas de extracción y preparación de los datos, segmentación y reconocimiento de locutores, y por último el sistema de identificación de idioma.

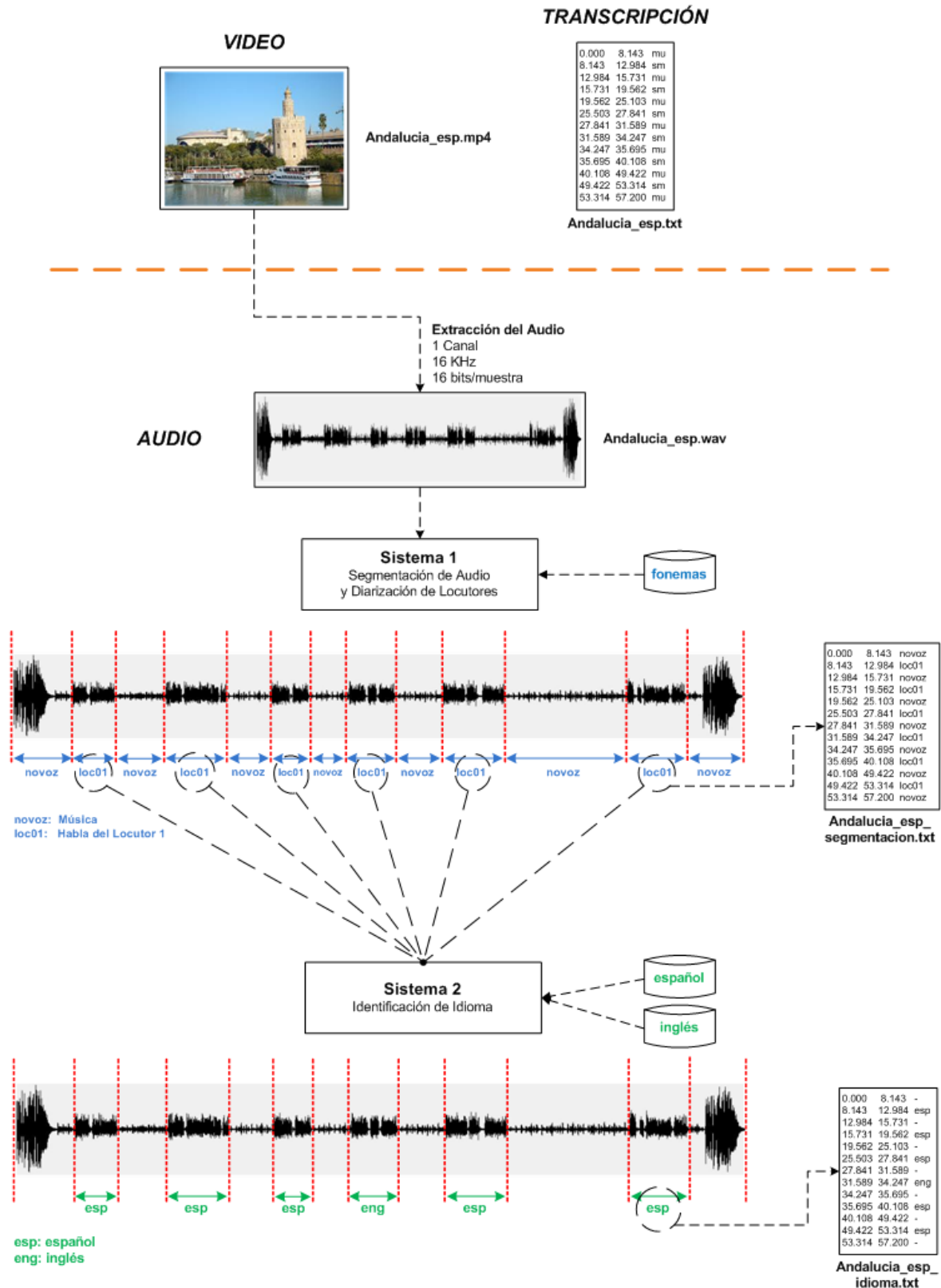


Ilustración 40. Secuencia de ejecución en cadena de los sistemas implementados.

4

Pruebas y Resultados

4.1. Pruebas realizadas

En este capítulo se describen las pruebas experimentales más importantes realizadas para este proyecto, con el fin de comparar y evaluar las diferentes bondades o fallos de los sistemas implementados.

Para ello ha sido necesario disponer de unos archivos de referencia que se denominan *transcripciones*, con los que se pudiera comparar si el reconocimiento era correcto total o parcialmente, y en tal caso ver tasas de acierto y fallo.

Dichas transcripciones componen la base de datos principal para realización de pruebas y verificación de resultados. Contienen los datos más característicos e interesantes para este proyecto del video que se pretenda analizar: tiempos de inicio y fin de todos los segmentos de un audio e identificación real de la información que contiene (*voz o no voz* que se relacionan biunívocamente con las etiquetas de los sistemas de reconocimiento *loc01* y *novoz*).

Las pruebas realizadas son muy sencillas. Tanto para probar la segmentación y reconocimiento de locutores, como para la identificación de idioma, se dispone de las duraciones y etiquetado de los segmentos de todos los audios. Con ello se obtienen los tiempos máximos de duración por categorías, y porcentajes de acierto o error al reconocer los diferentes tipos de segmento.

De esta forma, se pueden realizar pequeñas y visuales representaciones de las tasas de acierto o error sobre cada video y valorar de forma rápida el comportamiento del sistema evaluado.

Principalmente se han realizado las siguientes pruebas:

- **Pruebas de los sistemas de Segmentación y reconocimiento de locutores.**
 - Segmentación y reconocimiento de locutores con el sistema de Segmentación de Audio ATVS-UAM.
 - Segmentación y reconocimiento de locutores con el sistema de Reconocimiento Fonético y herramientas HTK.

- Segmentación y reconocimiento de locutores con el sistema de Reconocimiento Fonético y herramientas HTK con adaptación al locutor del audio del propio video.
 - Segmentación y reconocimiento de locutores mediante combinación con función AND de los sistemas ATVS-UAM y reconocimiento fonético.
 - Segmentación y reconocimiento de locutores mediante combinación con función OR de los sistemas ATVS-UAM y reconocimiento fonético.
-
- **Pruebas del sistema de Identificación de Idioma.**
 - Identificación de idioma con el sistema de reconocimiento fonético de idioma aplicando el modelo HMM-1-GAUSS.
 - Identificación de idioma con el sistema de reconocimiento fonético de idioma aplicando el modelo HMM-2-GAUSS.
 - Identificación de idioma con el sistema de reconocimiento fonético de idioma aplicando el modelo HMM-3-GAUSS.
 - Identificación de idioma con el sistema de reconocimiento fonético de idioma aplicando el modelo HMM-4-GAUSS.

4.2. Resultados experimentales

En vista a la gran cantidad de información obtenida, se analizarán los resultados de dos videos conscientemente seleccionados cuyas pruebas hayan ofrecido resultados dispares y que permitan sacar conclusiones claras y determinantes. Los videos cuyos audios han sido elegidos para este análisis han sido: '*Andalucia_esp.mp4*' y '*PueblosEdadMedia_esp.mp4*', pudiendo consultar el resultado del resto de los videos en las tablas creadas para este fin e incluidas en el Anexo B. En dicho Anexo B también se pueden consultar los diagramas de barras que expresan las tasas de acierto o fallo de los sistemas implementados, así como el alineamiento de todos los reconocimientos para estos dos audios.

4.2.1. Forma de presentación de resultados

4.2.1.1. Evaluación de los sistemas de Segmentación y Reconocimiento de locutores.

La evaluación de las tasas de acierto o error se realiza sobre los tiempos originales, mediante comparación directa con la información disponible en las transcripciones reales. El diseño del sistema permite realizar tan sólo comparaciones de segmentación real frente a segmentaciones obtenidas mediante reconocimiento de los dos sistemas implementados. Para ello se han diseñado unas tablas individuales de análisis de tiempos y porcentajes en las que se representan datos reales frente a datos reconocidos para cada audio, como la de la siguiente imagen.

AUDIO	REAL	TIEMPOS (s)		PORCENTAJES (%)	
NOMBRE	RECONOC	NOVOZ	LOC01	NOVOZ	LOC01
	NOVOZ	A	C	A'	C'
	LOC01	B	D	B'	D'

Tabla 8. Ejemplo de presentación de los resultados obtenidos en segmentación y reconocimiento de locutores.

Para realizar una correcta lectura de dichos resultados se han de mirar resultados haciendo dos hipótesis reflejadas en la tabla anterior. En la parte de las columnas se han indicado los tiempos y porcentajes de los segmentos referidos a las condiciones reales del audio analizado. Mientras que en la parte de las filas se representan los tiempos y porcentajes referidos a los segmentos reconocidos por los sistemas implementados. De esta forma, cada celda se puede leer de la siguiente manera:

- **Celda A:** Representa el tiempo etiquetado como *novoz* por los diferentes sistemas de segmentación y reconocimiento frente al tiempo real del audio original etiquetado como *novoz*.

Esta celda aportará información del correcto funcionamiento de los sistemas en la identificación de segmentos de *novoz* como tales, y por lo tanto es una medida del acierto en este etiquetado. Esta casilla lleva **asociada la celda A'**, cuyo valor es más visual al tratarse de porcentajes, y representa el porcentaje del tiempo etiquetado como *novoz* por los diferentes sistemas de segmentación y reconocimiento, frente al total de tiempo.

A la hora de tomar decisiones se realizarán en base a esta celda, para la que valores altos de porcentajes indicarán que los segmentos se han reconocido correctamente y por el contrario, valores bajos de porcentaje indicarán que el reconocimiento ha fallado en este segmento.

- Celda B:** Representa el tiempo etiquetado como *loc01 (voz)* por los diferentes sistemas de segmentación y reconocimiento frente al tiempo real del audio original etiquetado como *novoz*.
De la misma manera, **lleva asociada la celda B'**, que será un indicador del porcentaje de tiempo real de *novoz* que se ha reconocido como *voz*. Por lo tanto, valores altos en esta celda, al contrario que en la anterior, supone que ha existido una alta tasa de error en la identificación del segmento. Análogamente, un porcentaje de tiempo bajo, indica que el error cometido al reconocer ese segmento ha sido bajo.
- Celda C:** Representa el tiempo etiquetado como *novoz* por los diferentes sistemas de segmentación y reconocimiento frente al tiempo real del audio original etiquetado como *loc01 (voz)*.
La celda C' está asociada a ella, siendo un claro indicador del porcentaje de tiempo real de *voz* que se ha etiquetado como *novoz*. Ello supone que en valores de porcentaje alto se refleje un fallo de reconocimiento del sistema en este segmento, y por el contrario, valores bajos suponga un acierto del sistema en el reconocimiento.
- Celda D:** Representa el tiempo etiquetado como *loc01 (voz)* por los diferentes sistemas de segmentación y reconocimiento frente al tiempo real del audio original etiquetado como *loc01 (voz)*.
La celda D' está directamente asociada a ella, siendo un claro indicador del porcentaje de tiempo real de *voz* que se ha etiquetado como *loc01 (voz)*. Ello supone que en valores de porcentaje alto se refleje un acierto de reconocimiento del sistema en este segmento, y por el contrario, valores bajos suponga un fallo del sistema en el reconocimiento.

Por lo explicado anteriormente, se ha de observar que las celdas se corresponden biunívocamente dos a dos en cuanto a tiempos y porcentajes, pero también se corresponden directamente casillas de tiempos y porcentajes reales. Esto es, que la suma de los porcentajes A' y B' componen la totalidad del porcentaje de audio de los segmentos reales de *novoz*, de la misma forma que la suma de los porcentajes C' y D' corresponden a la totalidad de los segmentos reales de *loc01 (voz)*.

A continuación se muestra a modo de ejemplo, un extracto de la transcripción original del audio 'Andalucia_esp', y a su lado la transcripción segmentada por ATVS-UAM. Las pruebas se realizan en base a este tipo de datos.

TRANSCRIPCIÓN ORIGINAL				TRANSCRIPCIÓN SEGMENTACION ATVS-UAM			
Andalucia_esp	0.000	10.982	novoz	Andalucia_esp	0.00	11.28	novoz
Andalucia_esp	10.982	30.218	loc01	Andalucia_esp	11.28	29.40	loc01
Andalucia_esp	30.218	40.365	novoz	Andalucia_esp	29.40	39.43	novoz
Andalucia_esp	40.365	60.192	loc01	Andalucia_esp	39.43	59.00	loc01
Andalucia_esp	60.192	64.974	novoz	Andalucia_esp	59.00	64.97	novoz
Andalucia_esp	64.974	74.173	loc01	Andalucia_esp	64.97	73.12	loc01
Andalucia_esp	74.173	80.022	novoz	Andalucia_esp	73.12	73.16	novoz
Andalucia_esp	80.022	92.216	loc01	Andalucia_esp	73.16	73.21	loc01
Andalucia_esp	92.216	98.484	novoz	Andalucia_esp	73.21	79.28	novoz
Andalucia_esp	98.484	106.536	loc01	Andalucia_esp	79.28	92.52	loc01
Andalucia_esp	106.536	107.642	novoz	Andalucia_esp	92.52	98.51	novoz
Andalucia_esp	107.642	132.356	loc01	Andalucia_esp	98.51	130.58	loc01
Andalucia_esp	132.356	150.00	novoz	Andalucia_esp	130.58	150.00	novoz

Tabla 9. Extracto de transcripciones original y segmentada del audio 'Andalucia_esp.wav'.

Con el fin de obtener mejores indicadores visuales del reconocimiento realizado, se presenta otra forma de análisis de los datos mediante el software '*wavesurfer*'. Dicho software es un conjunto de herramientas que permiten realizar análisis de la señal de voz en todos los niveles: temporal, espectral, cepstral, etc..

En este proyecto se ha utilizado para representar la señal de audio junto a las transcripciones de cada etapa de reconocimiento, y así poder obtener los segmentos alineados. A continuación se presenta una captura de pantalla de este software, que en concreto se trata de una sección de audio '*Andalucia_esp.wav*'.

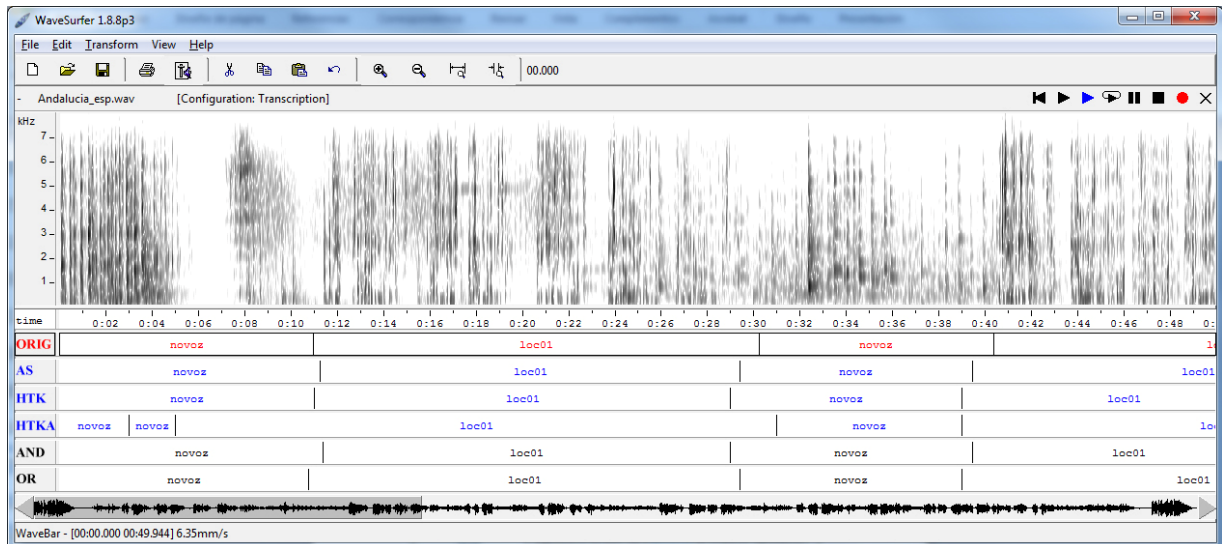


Ilustración 41. Ejemplo de presentación de resultados con '*wavesurfer*'.

En la parte inferior se pueden observar seis filas con los títulos ORIG, AS, HTK, HTKA, AND y OR. Cada una de ellas se corresponde con cada una de las transcripciones obtenidas en la aplicación en cadena del sistema de segmentación de audio y reconocimiento de locutores:

ETIQUETA	SIGNIFICADO
ORIG	Segmentación del audio extraído de la transcripción fonética original
AS	Segmentación del audio tras la aplicación del sistema de Segmentación y Reconocimiento de locutores ATVS-UAM
HTK	Segmentación del audio tras la aplicación del sistema de Segmentación y Reconocimiento de locutores por reconocimiento fonético
HTKA	Segmentación del audio tras la aplicación del sistema de Segmentación y Reconocimiento de locutores por reconocimiento fonético con adaptación al locutor
AND	Segmentación del audio tras la combinación de los resultados obtenidos con AS y HTK, mediante la función AND
OR	Segmentación del audio tras la combinación de los resultados obtenidos con AS y HTK, mediante la función OR

Tabla 10. Significado de las etiquetas de '*wavesurfer*'.

4.2.1.2. Evaluación del sistema de Identificación de Idioma.

La evaluación de las tasas de acierto y fallo que se producen tras la aplicación del sistema de identificación de idioma mediante reconocimiento fonético se recogen en tablas como la que se detalla a continuación:

VIDEO	IDIOMA ORIGINAL	SCORE SPANISH	SCORE ENGLISH	IDIOMA RECONOCIDO	PORCENTAJE
NOMBRE	SPANISH / ENGLISH	X	Y	SPANISH / ENGLISH	100%

Tabla 11. Ejemplo de presentación de los resultados obtenidos en identificación de idioma.

En ella se ha especificarán los *scores* o puntuaciones obtenidas tras la aplicación de los reconocedores fonéticos de cada idioma, a cada audio en particular. Estas puntuaciones, que en la tabla se identifican con las letras X e Y, son datos ofrecidos por el software HTK y cuyo valor indica mayor o menor penalización. Por lo tanto, valores altos (en módulo) indicarán mayor penalización y por el contrario, valores más bajos indicarán mejor puntuación.

También se detallará el idioma del audio original y el idioma reconocido en base a decisiones sobre las puntuaciones anteriores, obteniendo un porcentaje de acierto o no acierto sobre la correcta identificación de idioma.

4.2.2. Tablas de Resultados

Como se ha comentado anteriormente, ante la gran cantidad de datos obtenidos durante la realización de las pruebas, tan sólo se mostrarán y analizarán los resultados de las pruebas realizadas sobre dos audios en particular, tanto para segmentación como para idioma. Se pueden consultar el resto de resultados de las pruebas de segmentación de audio, reconocimiento de locutores e identificación de idioma en el Anexo B.

En pruebas de segmentación de audio, dichos audios corresponden a los vídeos '*Andalucia_esp.mp4*' y '*PueblosEdadMedia_esp.mp4*', siendo los datos reales en cuanto a duración temporal de los segmentos de los audios los siguientes:

ID	VIDEO	DUR (M)	DUR (s)	TIEMPOS REALES (s)		
				NOVOZ (s)	LOC01 (s)	DUR (s)
3	Andalucia_esp.wav	00:02:30	150	56,778	93,222	150,000
35	PueblosEdadMedia_esp.wav	00:01:46	106	36,826	69,174	106,000

Tabla 12. Duración de tiempos real de los audios mostrados en las pruebas.

En pruebas de identificación de idioma, dichos audios corresponden a los vídeos '*Andalucia_esp.mp4*' y '*Montjuic_eng.mp4*', siendo los datos reales los siguientes:

ID	VIDEO	IDIOMA
3	Andalucia_esp.mp4	SPANISH
33	Montjuic_eng.mp4	ENGLISH

Tabla 13. Idiomas reales de los audios mostrados en las pruebas.

4.2.2.1. Sistema de Segmentación de Audio ATVS-UAM.

Los resultados para los audios de '*Andalucia_esp*' y '*PueblosEdadMedia_esp*' tras la aplicación del sistema de segmentación de audio ATVS-UAM, basado en HMMs han sido:

AUDIO	AUDIO SEGMENTATION ATVS-UAM				
	REAL	TIEMPOS (s)		PORCENTAJES	
	RECON	NOVOZ	LOC01	NOVOZ	LOC01
Andalucia_esp.wav	NOVOZ	55,177	1,016	97,18%	1,09%
	LOC01	1,601	92,206	2,82%	98,91%
PueblosEdadMedia_esp.wav	NOVOZ	0,000	0,297	0,00%	0,43%
	LOC01	36,826	68,877	100,00%	99,57%

Tabla 14. Resultados del sistema de segmentación de audio ATVS-UAM.

Si se analizan los porcentajes del audio segmentado y reconocido, se pueden sacar algunas interpretaciones inmediatas:

- El sistema tiene una tasa de acierto cercana al 98% a la hora de detectar los segmentos correctos en el audio '*Andalucia_esp.wav*'. Por el contrario, la segmentación del audio '*PueblosEdadMedia_esp*' se realiza de forma errónea en los segmentos en los que no hay voz, pues el 100% de ellos son reconocidos como segmentos de voz.
- Lo primero que se puede pensar es que el entrenamiento de los HMMs proporcionados se realizara con grabaciones de audio que no eran lo suficientemente representativas del habla, en este caso. En particular cabe destacar que el entrenamiento de los HMMs se ha realizado con habla telefónica mientras que el habla en este caso es de características muy distintas (microfónica y con calidades diversas).
- También se puede pensar directamente en las características externas que más afecta a este tipo de sistemas. El diseño y entrenamiento de los sistemas de reconocimiento se realiza con grabaciones de audio controladas, es decir, bajo niveles de ruidos, ecos o interferencias casi nulos y suelen realizarse en salas insonorizadas. En este caso, se dispone de audios con música de fondo (que parece no ser el mayor problema pues funciona bien las tasas de acierto son altas en la mayoría de los reconocimientos), y en este caso en especial, el audio '*PueblosEdadMedia_esp.wav*' se encuentra altamente influenciado por ecos y reverberaciones cuya naturaleza es desconocida.

En el Anexo B se encuentran el resto de tablas que se pueden consultar para obtener información del conjunto completo de pruebas realizadas a todos los audios de la base de datos. En concreto pueden resultar de especial interés las tablas:

- **Anexo B:** B1. Tiempos y porcentajes tras la aplicación del Sistema de Segmentación de Audio ATVS-UAM.
- **Anexo B:** B10. Diagramas de barras de las tasas de acierto y fallo, para segmentos de VOZ en el reconocimiento con el sistema ATVS-UAM.
- **Anexo B:** B11. Diagramas de barras de las tasas de acierto y fallo, para segmentos de NOVOZ en el reconocimiento con el sistema ATVS-UAM.
- **Anexo B:** B22. Representación espectral, temporal y etiquetado de segmentos del audio '*Andalucia_esp.wav*' con *software wavesurfer*.
- **Anexo B:** B23. Representación espectral, temporal y etiquetado de segmentos del audio '*PueblosEdadMedia_esp.wav*' con *software wavesurfer*.

Debido a la gran cantidad de información, tan sólo se han incluido en el Anexo B, las comparativas con 'wavesurfer' de los dos audios mencionados, estando disponibles el resto en caso de ser necesario.

4.2.2.2. Sistema de Segmentación de Audio mediante rec. fonético.

Los resultados tras la aplicación del sistema de segmentación de audio mediante reconocimiento fonético, basado en HMMs entrenados con conversación telefónica en español han sido:

AUDIO	REAL RECON	RECONOCIMIENTO FONETICO			
		TIEMPOS (s)		PORCENTAJES	
		NOVOZ	LOC01	NOVOZ	LOC01
Andalucia_esp.wav	NOVOZ	52,265	11,735	92,05%	12,59%
	LOC01	4,513	81,487	7,95%	87,41%
PueblosEdadMedia_esp.wav	NOVOZ	36,762	43,238	99,83%	62,51%
	LOC01	0,064	25,936	0,17%	37,49%

Tabla 15. Resultados del sistema de segmentación de audio por rec. fonético.

Si se analizan los porcentajes del audio segmentado y reconocido, se pueden sacar algunas interpretaciones claras:

- Con el reconocimiento fonético se obtienen tasas de acierto altas, cercanas al 95% en el caso de reconocimiento de segmentos en los que no existe voz, pero esta mejoría no se ve reflejada en los segmentos que se debería identificar tan sólo como voz.
- De nuevo se puede comprobar que existe una alta influencia de factores externos al locutor en el reconocimiento de voz. Además en este caso, y por el diseño del sistema fonético, la alta tasa de fallo en el reconocimiento de segmentos reales de no voz como segmentos de voz, cerca de un 62%, se puede deber a la coexistencia de la música y habla, y la consecuente distorsión de los fonemas reconocidos con respecto a los fonemas de la base de datos.

En el Anexo B se encuentran el resto de tablas y en concreto se estiman de especial interés:

- **Anexo B:** B2. Tiempos y porcentajes tras la aplicación del Reconocedor fonético para segmentación de audio.
- **Anexo B:** B12. Diagramas de barras de las tasas de acierto y fallo, para segmentos de VOZ en reconocimiento con Reconocedor fonético.
- **Anexo B:** B13. Diagramas de barras de las tasas de acierto y fallo, para segmentos de NOVOZ en reconocimiento con Reconocedor fonético.
- **Anexo B:** B22. Representación espectral, temporal y etiquetado de segmentos del audio 'Andalucia_esp.wav' con *software wavesurfer*.
- **Anexo B:** B23. Representación espectral, temporal y etiquetado de segmentos del audio 'PueblosEdadMedia_esp.wav' con *software wavesurfer*.

4.2.2.3. Sistema de Segmentación de Audio mediante rec. fonético y adaptación al locutor.

Los resultados tras la aplicación del sistema de segmentación de audio mediante reconocimiento fonético, basado en HMMs entrenados con conversación telefónica en español, y adaptado al propio locutor del audio mediante algoritmo MLLR han sido:

AUDIO	RECONOCIMIENTO FONÉTICO CON ADAPTACIÓN AL LOCUTOR				
	REAL	TIEMPOS (s)		PORCENTAJES	
	RECON	NOVOZ	LOC01	NOVOZ	LOC01
Andalucia_esp.wav	NOVOZ	38,000	0,000	66,93%	0,00%
	LOC01	18,778	93,222	33,07%	100,00%
PueblosEdadMedia_esp.wav	NOVOZ	31,970	9,030	86,81%	13,05%
	LOC01	4,856	60,144	13,19%	86,95%

Tabla 16. Resultados del sistema de segmentación de audio por rec. fonético y adaptación al locutor.

En este caso, si se realiza un análisis idéntico a los anteriores, las conclusiones al aplicar la adaptación al locutor al sistema de segmentación por reconocimiento fonético pueden ser erróneas. De éstos porcentajes se pueden sacar las observaciones:

- Las tasas de acierto en el audio '*Andalucia_esp.wav*' han empeorado al aplicar la adaptación al locutor MLLR. Ello es debido a la cantidad de datos con la que se han reestimado los HMMs disponibles. Es necesario disponer de una cantidad de datos de audio del locutor lo suficientemente representativa como para que los resultados mejoren, en lugar de empeorar. En este caso, los audios son de apenas dos o tres minutos, de los cuales el locutor pronuncia palabras en la mitad del tiempo, algo insuficiente.
- Por el contrario, para el audio '*PueblosEdadMedia_esp*', las tasas de acierto de ambos segmentos han mejorado, llegando a establecerse cerca de un 87%, cantidad que puede considerarse aceptable debido a las condiciones internas de diseño del sistema y externas de la grabación de los audios.

En el Anexo B se encuentran el resto de tablas siendo de especial interés:

- **Anexo B:** B3. Tiempos y porcentajes tras la aplicación del Reconocedor fonético con adaptación al locutor para segmentación de audio.
- **Anexo B:** B14. Diagramas de barras de las tasas de acierto y fallo, para segmentos de VOZ en reconocimiento con Reconocedor fonético.
- **Anexo B:** B15. Diagramas de barras de las tasas de acierto y fallo, para segmentos de NOVOZ en reconocimiento con Rec. fonético con adaptación al locutor.
- **Anexo B:** B22. Representación espectral, temporal y etiquetado de segmentos del audio '*Andalucia_esp.wav*' con *software wavesurfer*.
- **Anexo B:** B23. Representación espectral, temporal y etiquetado de segmentos del audio '*PueblosEdadMedia_esp.wav*' con *software wavesurfer*.

4.2.2.4. Combinación de sistemas mediante función AND.

Los resultados tras la aplicación combinada mediante la función lógica AND, de los sistemas de segmentación de audio ATVS-UAM y reconocimiento fonético, han sido:

AUDIO	COMBINACIÓN AND				
	REAL	TIEMPOS (s)		PORCENTAJES	
	RECON	NOVOZ	LOC01	NOVOZ	LOC01
Andalucia_esp.wav	NOVOZ	50,686	9,509	89,27%	10,20%
	LOC01	6,092	83,713	10,73%	89,80%
PueblosEdadMedia_esp.wav	NOVOZ	31,424	34,721	85,33%	50,19%
	LOC01	5,402	34,453	14,67%	49,81%

Tabla 17. Resultados de la combinación mediante función AND de los sistemas de segmentación de audio y reconocimiento de locutores.

Si se analizan los porcentajes del audio segmentado y reconocido en este caso, se pueden sacar algunas interpretaciones:

- La combinación mediante la función AND no ofrece resultados que mejoren el reconocimiento global en segmentos de voz. Ello es debido que uno de los sistemas ha proporcionado altas tasas de fallo en estos segmentos, que unido a la capacidad restrictiva de la propia función AND, hace que no se obtengan resultados favorables.
- Esta función se podría utilizar para determinar segmentos que ambos sistemas tengan en común siempre de inferior longitud con respecto a los segmentos identificados por los sistemas. Así se podría afirmar con total seguridad que el etiquetado es correcto en dichos segmentos a cambio de obtener una tasa más baja de acierto.

En el Anexo B se encuentran el resto de tablas de aplicación de la función AND, resultando de especial interés:

- **Anexo B:** B4. Tiempos y porcentajes tras la combinación de resultados de los sistemas de Segmentación de Audio y Rec. fonético con la función lógica AND.
- **Anexo B:** B16. Diagramas de barras de las tasas de acierto y fallo, para segmentos de VOZ en combinación de sistemas con función lógica AND.
- **Anexo B:** B17. Diagramas de barras de las tasas de acierto y fallo, para segmentos de NOVOZ en combinación de sistemas con función lógica AND.
- **Anexo B:** B22. Representación espectral, temporal y etiquetado de segmentos del audio 'Andalucia_esp.wav' con *software wavesurfer*.
- **Anexo B:** B23. Representación espectral, temporal y etiquetado de segmentos del audio 'PueblosEdadMedia_esp.wav' con *software wavesurfer*.

4.2.2.5. Combinación de sistemas mediante función OR.

Los resultados tras la aplicación combinada mediante la función lógica OR, de los sistemas de segmentación de audio ATVS-UAM y reconocimiento fonético, han sido:

AUDIO	COMBINACIÓN OR				
	REAL	TIEMPOS (s)		PORCENTAJES	
	RECON	NOVOZ	LOC01	NOVOZ	LOC01
Andalucia_esp.wav	NOVOZ	51,254	4,981	90,27%	5,34%
	LOC01	5,524	88,241	9,73%	94,66%
PueblosEdadMedia_esp.wav	NOVOZ	0,037	17,397	0,10%	25,15%
	LOC01	36,789	51,777	99,90%	74,85%

Tabla 18. Resultados de la combinación mediante función OR de los sistemas de segmentación de audio y reconocimiento de locutores.

En este caso, si se realiza un análisis idéntico al anterior, se puede observar que:

- La mejoría en la segmentación y el reconocimiento del video '*Andalucia_esp.wav*' es evidente. Ello es debido a la propia definición de la función OR. Esta función permite seleccionar la duración completa de los segmentos de ambos reconocimientos en los que se haya determinado la existencia de voz.
- Por el contrario, los resultados ofrecidos para el audio '*PueblosEdadMedia_esp*' son bastante bajos. Como es de esperar cuando uno de los sistemas tiene altas tasas de error, la aplicación de una combinación de los mismos mediante la función OR, hace que los errores también se combinen, ampliando el rango de segmentos mal reconocidos.

En el Anexo B se encuentran el resto de tablas de aplicación de la función OR, resultando de especial interés:

- **Anexo B:** B5. Tiempos y porcentajes tras la combinación de resultados de los sistemas de Segmentación de Audio y Rec. fonético con la función lógica OR.
- **Anexo B:** B18. Diagramas de barras de las tasas de acierto y fallo, para segmentos de VOZ en combinación de sistemas con función lógica OR.
- **Anexo B:** B19. Diagramas de barras de las tasas de acierto y fallo, para segmentos de NOVOZ en combinación de sistemas con función lógica OR.
- **Anexo B:** B22. Representación espectral, temporal y etiquetado de segmentos del audio '*Andalucia_esp.wav*' con *software wavesurfer*.
- **Anexo B:** B23. Representación espectral, temporal y etiquetado de segmentos del audio '*PueblosEdadMedia_esp.wav*' con *software wavesurfer*.

4.2.2.6. Resultados Globales de Segmentación de audio y Reconocimiento de locutor.

En este apartado se resumen los resultados globales, promedio de reconocimiento de las diferentes estrategias utilizadas anteriormente. Las tablas que se presentan a continuación contienen los porcentajes de audio original (disponible en transcripciones) frente al detectado con los sistemas de reconocimiento.

SISTEMA DE SEGMENTACIÓN ATVS-UAM		DETECTADO	
		NOVOZ	LOC01
TRANSCRIPCIÓN	NOVOZ	84,98%	6,40%
	LOC01	15,02%	93,60%

SISTEMA DE SEGMENTACIÓN RECONOCEDOR FONÉTICO		DETECTADO	
		NOVOZ	LOC01
TRANSCRIPCIÓN	NOVOZ	92,42%	34,48%
	LOC01	7,58%	65,52%

SISTEMA DE SEGMENTACIÓN REC. FONÉTICO Y ADAPTACIÓN		DETECTADO	
		NOVOZ	LOC01
TRANSCRIPCIÓN	NOVOZ	79,56%	9,88%
	LOC01	20,44%	90,12%

SISTEMA DE SEGMENTACIÓN COMBINACIÓN OR		DETECTADO	
		NOVOZ	LOC01
TRANSCRIPCIÓN	NOVOZ	89,54%	33,07%
	LOC01	10,46%	66,93%

SISTEMA DE SEGMENTACIÓN COMBINACIÓN AND		DETECTADO	
		NOVOZ	LOC01
TRANSCRIPCIÓN	NOVOZ	80,14%	7,74%
	LOC01	19,86%	92,26%

En el Anexo B se encuentran el resto de resultados con los que se acompaña y se completa a este apartado.

4.2.2.7. Sistema de Identificación de Idioma mediante HMMs de 1 Gaussiana.

Los resultados tras la aplicación del sistema de identificación de idioma mediante reconocimiento fonético con HMM de 1 Gaussiana tanto en español como en inglés, han sido:

		MODELO 1 GAUSS			
AUDIO	IDIOMA ORIGINAL	SCORE SPANISH	SCORE ENGLISH	IDIOMA RECONOCIDO	PORCENTAJE
Andalucia_esp.wav	SPANISH	-127,4496	-128,1953	SPANISH	100%
Montjuic_eng.wav	ENGLISH	-130,7432	-130,7099	ENGLISH	100%

Tabla 19. Resultados de la identificación de Idioma mediante HMMs de 1 Gaussiana.

En este caso se puede ver como:

- El sistema de identificación de idioma se ha comportado como se esperaba ante la correcta identificación del idioma original de cada audio. El *score* o puntuación ofrecida por los reconocedores fonéticos de inglés y español aplicados a cada audio ofrecen puntuaciones más bajas en el caso correcto.
- Por otro lado, los scores apenas difieren en décimas e incluso en centésimas. Este hecho indica un funcionamiento parecido de los dos sistemas reconocedores de idioma y se debe de tener en cuenta al sacar conclusiones.

En el Anexo B se encuentran el resto de resultados que acompañan a este apartado, resultando de especial interés:

- **Anexo B:** B6. Puntuaciones y resultados obtenidos tras el reconocimiento de Idioma y decisiones tomadas con el modelo de HMM de 1 Gaussiana.
- **Anexo B:** B20. Diagramas de barras de las tasas de acierto y fallo, para identificaciones de idioma de audios en ESPAÑOL.
- **Anexo B:** B21. Diagramas de barras de las tasas de acierto y fallo, para identificaciones de idioma de audios en INGLÉS.

4.2.2.8. Sistema de Identificación de Idioma mediante HMMs de 2 Gaussianas.

Los resultados tras la aplicación del sistema de identificación de idioma mediante reconocimiento fonético con HMM de 2 Gaussianas tanto en español como en inglés, han sido:

		MODELO 2 GAUSS			
AUDIO	IDIOMA ORIGINAL	SCORE SPANISH	SCORE ENGLISH	IDIOMA RECONOCIDO	PORCENTAJE
Andalucia_esp.wav	SPANISH	-127,0865	-127,9860	SPANISH	100%
Montjuic_eng.wav	ENGLISH	-130,5982	-130,5991	SPANISH	0%

Tabla 20. Resultados de la identificación de Idioma mediante HMMs de 2 Gaussianas.

Se puede comprobar de forma inmediata:

- Cómo la identificación del idioma en el archivo '*Andalucia_esp.wav*' se realiza de forma correcta, pero para el archivo '*PueblosEdadMedia_esp*', la decisión es errónea. Se pueden observar las puntuaciones de ambos reconocedores, que en este caso son muy próximas y hay que recurrir al tercer decimal para tomar la decisión.
- Se podría pensar en la conveniencia o no del uso de modelos de varias gaussianas para la identificación de idioma. En este caso el reconocimiento no ha sido correcto a pesar de haberse realizado correctamente con modelos HMM de una gaussiana.

En el Anexo B se encuentran el resto de resultados que acompañan a este apartado, resultando de especial interés:

- **Anexo B:** B7. Puntuaciones y resultados obtenidos tras el reconocimiento de Idioma y decisiones tomadas con el modelo de HMM de 2 Gaussianas.
- **Anexo B:** B20. Diagramas de barras de las tasas de acierto y fallo, para identificaciones de idioma de audios en ESPAÑOL.
- **Anexo B:** B21. Diagramas de barras de las tasas de acierto y fallo, para identificaciones de idioma de audios en INGLÉS.

4.2.2.9. Sistema de Identificación de Idioma mediante HMMs de 3 Gaussianas.

Los resultados tras la aplicación del sistema de identificación de idioma mediante reconocimiento fonético con HMM de 3 Gaussianas tanto en español como en inglés, han sido:

		MODELO 3 GAUSS			
AUDIO	IDIOMA ORIGINAL	SCORE SPANISH	SCORE ENGLISH	IDIOMA RECONOCIDO	PORCENTAJE
Andalucia_esp.wav	SPANISH	-126,8953	-126,8992	SPANISH	100%
Montjuic_eng.wav	ENGLISH	-130,4568	-130,7234	SPANISH	0%

Tabla 21. Resultados de la identificación de Idioma mediante HMMs de 3 Gaussianas.

En este caso sucede algo parecido a lo que se observa en la identificación con modelo de 2 gaussianas:

- La identificación del idioma en el archivo '*Andalucia_esp.wav*' se realiza de forma correcta, pero para el archivo '*PueblosEdadMedia_esp*', la decisión es errónea esta vez se observa un distanciamiento a tener en cuenta en las puntuaciones del reconocedor de inglés.
- Ambos reconocedores de idioma español o inglés son idénticos, salvo por la gramática propia de cada idioma, el diccionario adaptado a cada idioma y los HMMs. entrenados con audio propio de cada idioma. En estos tres factores se halla el motivo de tal efecto.

En el Anexo B se encuentran el resto de resultados que acompañan a este apartado, siendo de especial interés:

- **Anexo B:** B8. Puntuaciones y resultados obtenidos tras el reconocimiento de Idioma y decisiones tomadas con el modelo de HMM de 3 Gaussianas.
- **Anexo B:** B20. Diagramas de barras de las tasas de acierto y fallo, para identificaciones de idioma de audios en ESPAÑOL.
- **Anexo B:** B21. Diagramas de barras de las tasas de acierto y fallo, para identificaciones de idioma de audios en INGLÉS.

4.2.2.10. Sistema de Identificación de Idioma mediante HMMs de 4 Gaussianas.

Los resultados tras la aplicación del sistema de identificación de idioma mediante reconocimiento fonético con HMM de 4 Gaussianas tanto en español como en inglés, han sido:

		MODELO 4 GAUSS			
AUDIO	IDIOMA ORIGINAL	SCORE SPANISH	SCORE ENGLISH	IDIOMA RECONOCIDO	PORCENTAJE
Andalucia_esp.wav	SPANISH	-126,9100	-128,1523	SPANISH	100%
Montjuic_eng.wav	ENGLISH	-130,4098	-130,8037	SPANISH	0%

Tabla 22. Resultados de la identificación de Idioma mediante HMMs de 4 Gaussianas.

En este caso se confirma el distanciamiento de puntuaciones y por lo tanto el error cometido para el audio '*PueblosEdadMedia_esp*', mientras se identifica de forma correcta el idioma el audio '*Andalucia_esp.wav*'.

En el Anexo B se encuentran el resto de resultados que acompañan a este apartado, siendo de especial interés:

- **Anexo B:** B9. Puntuaciones y resultados obtenidos tras el reconocimiento de Idioma y decisiones tomadas con el modelo de HMM de 4 Gaussianas.
- **Anexo B:** B20. Diagramas de barras de las tasas de acierto y fallo, para identificaciones de idioma de audios en ESPAÑOL.
- **Anexo B:** B21. Diagramas de barras de las tasas de acierto y fallo, para identificaciones de idioma de audios en INGLÉS.

4.2.2.11. Resultados Globales en Identificación de Idioma.

En este apartado se resumen los resultados promedio obtenidos en el reconocimiento de idioma para los cuatro HMMs utilizados.

HMM 1 GAUSSIANA	ACIERTOS	FALLOS
NUMERO	29	17
PORCENTAJE	63,04%	36,96%

HMM 2 GAUSSIANAS	ACIERTOS	FALLOS
NUMERO	31	15
PORCENTAJE	67,39%	32,61%

HMM 3 GAUSSIANAS	ACIERTOS	FALLOS
NUMERO	32	14
PORCENTAJE	69,57%	30,43%

HMM 4 GAUSSIANAS	ACIERTOS	FALLOS
NUMERO	31	15
PORCENTAJE	67,39%	32,61%

En el Anexo B se encuentran el resto de resultados con los que se acompaña y se completa a este apartado.

5

Conclusiones y Trabajo Futuro

5.1. Conclusiones

Durante el desarrollo del proyecto y a la vista de los resultados obtenidos, se pueden extraer las siguientes conclusiones:

- **Conclusiones obtenidas en Segmentación de audio y Reconocimiento de locutores.**
 - ✓ Debido a las limitaciones propias de la base de datos, no se ha podido evaluar el objetivo de la segmentación de la voz en locutores. Dicha base de datos se compone por videos en cuya narración participaba tan sólo un único locutor por cada video. De esta manera, la identificación de locutores se ha realizado de manera inmediata y no se ha requerido el uso de ningún algoritmo.
 - ✓ Otro factor a evaluar era la agrupación de locutores, siendo imposible por las mismas características. Por lo tanto, el sistema se emplea como un sistema de segmentación de audio.
 - ✓ La calidad de las grabaciones de voz es fundamental para obtener buenos resultados. Son necesarios medios de grabación bien diseñados, aislados frente a perturbaciones, ruidos e interferencias que no produzcan ecos o reverberaciones.
 - ✓ La robustez del sistema pasa por obtener grabaciones con mejor calidad. Si ello no es posible existen métodos como la adaptación al locutor, que permiten dotar de mayor fiabilidad al sistema.

- ✓ La adaptación al locutor no siempre mejora resultados. Para ello es necesario reestimar los HMMs con una cantidad de audio lo suficientemente representativa del locutor que interviene en la grabación.
- ✓ La tendencia del sistema desarrollado es a clasificar el audio más como voz que como música (o 'novoz'). Para optimizar esta inclinación habría que trabajar en mejorar el umbral de decisión.
- **Conclusiones obtenidas en Identificación de Idioma.**
 - ✓ La cantidad de idiomas de la base de datos es limitada, disponiendo de tan sólo dos.
 - ✓ Los ficheros evaluados contiene habla en un solo idioma, siendo imposible evaluar la segmentación y reconocimiento de idioma simultáneamente.
 - ✓ La calidad del audio afecta directamente a los sistemas de identificación de idioma basados en reconocimiento fonético. Los resultados no son todo lo buenos que se esperaban debido principalmente a este efecto.
 - ✓ Los HMM del reconocimiento fonético de los que se disponían, se entrenaron con habla telefónica, distinta al habla de los audios de la base de datos. Este hecho afecta directamente al correcto reconocimiento de idioma al tratarse de datos de diferente naturaleza.
 - ✓ La cantidad de datos disponible no ha sido suficiente como para diferenciar entre datos de entrenamiento y datos de test, usando todos los audios como test. Sería necesaria una cantidad de datos mayor (en tiempo y cantidad) para poder estimar correctamente los HMMs y luego proceder a su reconocimiento.

5.2. Trabajo futuro

Dados los experimentos realizados y las conclusiones extraídas, sería interesante tener en cuenta de cara a trabajos futuros:

- ✓ Sería interesante extender la base de datos a videos multilocutor para poder evaluar la eficacia de esta tecnología y así se podría evaluar el sistema de diarización de locutores propuesto.
- ✓ El ajuste del umbral de velocidad de habla humana es siempre un problema a mejorar, debido a las diferentes velocidades y características físicas y psicológicas de cada ser humano.
- ✓ La separación del audio en segmentos de inferior longitud permite abordar con mayores expectativas, labores como la detección de palabras clave. Esto es especialmente interesante para dotar de individualidad y personalización.
- ✓ También sería interesante realizar análisis de audios en los que aparecieran múltiples idiomas.
- ✓ Una posible mejora podría ser la adaptación del sistema de reconocimiento a voz de banda ancha, pues los modelos utilizados han sido adaptados a la voz telefónica.
- ✓ Los tiempos de ejecución son relativamente aceptables una vez se poseen todos los parámetros de los audios. Con algunas optimizaciones tanto en software como en hardware, podría comenzar a plantearse un análisis próximo al análisis en tiempo real.

Bibliografía

- [1] Javier Franco Pedroso, Ignacio López-Moreno, Doroteo T. Toledano, and Joaquín González-Rodríguez, "ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation", in FALA 2010, Vigo, España. pp. 415-417
- [2] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based Speaker Segmentation Using Speaker Factors and Eigenvoices," in Proc. ICASSP, Las Vegas, Nevada, Mar. 2008, pp. 4133 – 4136.
- [3] Najim Dehak, Reda Dehak, James Glass, Douglas Reynolds, and Patrick Kenny, "Cosine similarity scoring without score normalization techniques," in Odyssey, 2010
- [4] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller, Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features." Proc. ICSLP 2002, Sept. 2002, pp. 89-92.
- [5] Patrick Kenny, Douglas Reynolds and Fabio Castaldo. "Diarization of Telephone Conversations using Factor Analysis". IEEE Journal on Selected Topics In Signal Processing. 2010.
- [6] Najim Dehak, Patrick Kenny, Rda Dehak, Pierre Ouellet, and Pierre Dumouchel, "Front end Factor Analysis for Speaker Verification," IEEE Transactions on Audio, Speech and Language Processing, 2010.
- [7] Najim Dehak, Reda Dehak, James Glass, Douglas Reynolds, and Patrick Kenny, "Cosine similarity scoring without score normalization techniques," in Odyssey, 2010
- [8] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller, Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features." Proc. ICSLP 2002, Sept. 2002, pp. 89-92.
- [9] P. Matejka, L. Burget, P. Sckwarz, J Cernocky. "Brno University of Technology System for NIST 2005 Language Recognition Evaluation", in Proceedings of Odyssey 2006. Puerto Rico.
- [10] D. T. Toledano, E. Campos, A. Moreno, J. Colás, J. Garrido, "Resultados preliminares de decodificación fonética sobre distintos tipos de habla espontánea", in Proceedings III Jornadas de Tecnología del Habla, 17-19, Noviembre 2004, Valencia, Spain, pp. 227-232.
- [11] D. T. Toledano, A. Moreno, J. Colás, J. Garrido, "Acoustic-phonetic decoding of different types of spontaneous speech in Spanish", in Proceedings of the ISCA Workshop on Disfluency in Spontaneous Speech 2005, 10-12, September 2005, Aix-en-Provence, France.

- [12] A. O. Hatch, B. Peskin y A. Stolcke. "Improved phonetic speaker recognition using lattice decoding". Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 1, pp. 169-172, Marzo 2005.
- [13] W. M. Campbell, J. R. Campbell, D. A. Reynolds, D. A. Jones y T. R. Leek. "High-level speaker verification with support vector machines". Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 1, N. 17- 21, pp. 73-76, Mayo 2004a.
- [14] NIST SRE. Descripciones de las distintas evaluaciones NIST de locutor.
<http://www.nist.gov/speech/tests/spk>
- [15] NIST LRE. Descripciones de las distintas evaluaciones NIST de idioma.
<http://www.nist.gov/speech/tests/lang>
- [16] HTKBook. <http://htk.eng.cam.ac.uk/ftp/software/htkbook.pdf.zip>.
- [17] <http://htk.eng.cam.ac.uk/>
- [18] <http://www.mavir.net/>
- [19] <http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.htm>
- [20] <http://www.speech.kth.se/wavesurfer/>

Glosario de Términos

- **BASH** (Bourne Again SHell)
Programa informático para interpretar órdenes de diversos lenguajes de programación.
- **DTW** (Dynamic Time Warping)
Alineamiento Temporal Dinámico.
- **Fonema**
Unidad básica de una lengua humana. Sonido del habla que permite distinguir palabras por su pronunciación.
- **GMM** (Gaussian Mixture Models)
Modelo de mezclas de Gaussianas.
- **HMM** (Hidden Markov Models)
Modelos ocultos de Markov.
- **HTK** (Hidden markov models ToolKit)
Kit de herramientas para creación y tratamiento de HMM.
- **Lattice**
Referido a reconocimiento de locutores, entramado de características o parámetros en reconocimiento de patrones de varias dimensiones que identifica una secuencia de audio.
- **MAP** (Maximun A Posteriori)
Método de adaptación de modelos independientes de locutor a los distintos locutores.
- **MFCC** (Mel-Frequency Cepstral Coefficient)
Coeficientes cepstrales en escala de frecuencias MEL.
- **MLF** (Master Label Format)
Formato de salida del HTK para los *lattices*.
- **MLLR** (Maximum Likelihood Linear Regression)
Método de adaptación de modelos independientes de locutor a los distintos locutores mediante transformaciones lineales.
- **NIST** (National Institute of Standards and Technology)
Organismo federal no regulador, perteneciente a la Cámara de Comercio de los Estados Unidos que desarrolla y promueve medidas, estándares y tecnología para aumentar la productividad, facilitar el comercio y mejorar la calidad de vida.

- **PERL** (Practical Extraction and Report Language)
Lenguaje de programación que toma características de C y es interpretado por una *shell* (intérprete de comandos).
- **Reconocimiento de locutor**
Modalidad biométrica que utiliza el habla de una persona, una característica influenciada tanto por la estructura física del tracto vocal del individuo como por las características de comportamiento del individuo, para fines de reconocimiento. Existen dos posibles formas de reconocimiento, modo identificación y modo verificación de locutor.
- **Reconocimiento de idioma**
Tecnología empleada en el indexado de contenidos multimedia, enrutamiento en servicios de atención telefónica y configuración de sistema. Consiste en determinar el idioma de los interlocutores de una audición.
- **Solapamiento**
Cobertura de una cosa a otra parcial o totalmente. En audio se habla de un alineamiento de segmentos de audio, normalmente dos, con el fin de no perder características frecuenciales en el análisis posterior.
- **Transcripción**
Escritura física mediante los caracteres propios de una lengua, del habla pronunciado en una locución.
- **Trifonemas**
Conjunto formado por tres fonemas, que identifica la pronunciación de un conjunto de caracteres de una lengua.
- **Umbral**
Valor predeterminado de un usuario para las tareas de verificación o identificación de grupo abierto en los sistemas biométricos. La aceptación o el rechazo de los datos biométricos dependen de si el resultado de coincidencia se encuentra por encima o por debajo de la escala. La escala es ajustable de modo que el sistema biométrico puede ser más o menos estricto según los requisitos de cada aplicación biométrica.



Anexo A: Base de Datos MA2VICMR

SEGMENTACIÓN DE AUDIO Y DE LOCUTORES PARA RECUPERACIÓN DE INFORMACIÓN
MULTIMEDIA Y SU APLICACIÓN A VIDEOS DE INFORMACIÓN TURÍSTICA

ID	ARCHIVO	DURACION	VIDEO	TRANSCRIPCION	IDIOMA	SERIE
1	Alcalá de Henares ciudad Patrimonio de la Humanidad.mp4	00:06:11	Alcala_esp.mp4	Alcala_esp.txt	ESPAÑOL 	España
2	Andalusia. a journey into history and art.mp4	00:02:30	Andalucia_eng.mp4	Andalucia_eng.txt	INGLÉS 	España in sigth
3	Andalucía, un recorrido por la historia y el arte.mp4	00:02:30	Andalucia_esp.mp4	Andalucia_esp.txt	ESPAÑOL 	España in sigth
4	Aranjuez, a unique cultural landscape.mp4	00:03:38	Aranjuez_eng.mp4	Aranjuez_eng.txt	INGLÉS 	España in sigth
5	Aranjuez, un paisaje cultural único.mp4	00:03:38	Aranjuez_esp.mp4	Aranjuez_esp.txt	ESPAÑOL 	España in sigth
6	Ávila ciudad Patrimonio de la Humanidad.mp4	00:05:26	Avila_esp.mp4	Avila_esp.txt	ESPAÑOL 	España
7	Baeza. a World Heritage site.mp4	00:03:58	Baeza_eng.mp4	Baeza_eng.txt	INGLÉS 	España in sigth
8	Baeza. Patrimonio de la Humanidad.mp4	00:03:58	Baeza_esp.mp4	Baeza_esp.txt	ESPAÑOL 	España in sigth
9	Barcelona A tour of the Barrio Gótico and the Ribera neighbourhoods.mp4	00:02:10	Bcngotico_eng.mp4	Bcngotico_eng.txt	INGLÉS 	España in sigth
10	Barcelona recorrido por el Barrio Gótico y el Barrio de la Ribera.mp4	00:02:10	Bcngotico_esp.mp4	Bcngotico_esp.txt	ESPAÑOL 	España in sigth
11	Modernist Barcelona.mp4	00:01:45	Bcnmodernista_eng.mp4	Bcnmodernista_eng.txt	INGLÉS 	España in sigth
12	Barcelona modernista.mp4	00:01:45	Bcnmodernista_esp.mp4	Bcnmodernista_esp.txt	ESPAÑOL 	España in sigth
13	Cáceres ciudad Patrimonio de la Humanidad.mp4	00:05:37	Caceres_esp.mp4	Caceres_esp.txt	ESPAÑOL 	España
14	Camino de Santiago. más que un viaje.mp4	00:03:44	CaminoSantiago_esp.mp4	CaminoSantiago_esp.txt	ESPAÑOL 	España in sigth
15	El Camino de Santiago por Castilla y León.mp4	00:01:47	CastillaLeon_esp.mp4	CastillaLeon_esp.txt	ESPAÑOL 	-
16	Ruta del Cid.mp4	00:03:28	Cid_esp.mp4	Cid_esp.txt	ESPAÑOL 	-
17	Cuenca ciudad Patrimonio de la Humanidad.mp4	00:05:42	Cuenca_esp.mp4	Cuenca_esp.txt	ESPAÑOL 	España
18	Dalí master of Surrealism.mp4	00:03:25	Dalisurre_eng.mp4	Dalisurre_eng.txt	INGLÉS 	España in sigth
19	Dalí maestro del Surrealismo.mp4	00:03:25	Dalisurre_esp.mp4	Dalisurre_esp.txt	ESPAÑOL 	España in sigth
20	San Lorenzo de El Escorial.mp4	00:01:36	Escorial_esp.mp4	Escorial_esp.txt	ESPAÑOL 	-
21	Extremadura. a journey into history.mp4	00:03:24	Extremadura_eng.mp4	Extremadura_eng.txt	INGLÉS 	España in sigth
22	Extremadura. un viaje por la historia.mp4	00:03:24	Extremadura_esp.mp4	Extremadura_esp.txt	ESPAÑOL 	España in sigth
23	Gaudí. fairy-tale architecture.mp4	00:02:08	Gaudi_eng.mp4	Gaudi_eng.txt	INGLÉS 	España in sigth

SEGMENTACIÓN DE AUDIO Y DE LOCUTORES PARA RECUPERACIÓN DE INFORMACIÓN
MULTIMEDIA Y SU APLICACIÓN A VIDEOS DE INFORMACIÓN TURÍSTICA

ID	ARCHIVO	DURACION	VIDEO	TRANSCRIPCION	IDIOMA	SERIE
24	Gaudí arquitectura de ensueño.mp4	00:02:08	Gaudi_esp.mp4	Gaudi_esp.txt	ESPAÑOL 	España in sighth
25	Gaudí a genius in Barcelona.mp4	00:03:31	GaudiGenio_eng.mp4	GaudiGenio_eng.txt	INGLÉS 	España in sighth
26	Gaudí un genio en Barcelona.mp4	00:03:31	GaudiGenio_esp.mp4	GaudiGenio_esp.txt	ESPAÑOL 	España in sighth
27	Ibiza ciudad Patrimonio de la Humanidad.mp4	00:05:57	Ibiza_esp.mp4	Ibiza_esp.txt	ESPAÑOL 	España
28	La Laguna ciudad Patrimonio de la Humanidad.mp4	00:05:31	Laguna_esp.mp4	Laguna_esp.txt	ESPAÑOL 	España
29	La Mancha. the land of Don Quixote.mp4	00:02:49	LaMancha_eng.mp4	LaMancha_eng.txt	ESPAÑOL 	España in sighth
30	La Mancha. tierra de El Quijote.mp4	00:02:49	LaMancha_esp.mp4	LaMancha_esp.txt	ESPAÑOL 	España in sighth
31	Madrid. city of art.mp4	00:01:56	Madrid_eng.mp4	Madrid_eng.txt	INGLÉS 	España in sighth
32	Madrid. una ciudad de arte.mp4	00:01:56	Madrid_esp.mp4	Madrid_esp.txt	ESPAÑOL 	España in sighth
33	A walk around Montjuic.mp4	00:01:21	Montjuic_eng.mp4	Montjuic_eng.txt	INGLÉS 	España in sighth
34	Un paseo por Montjuic.mp4	00:01:21	Montjuic_esp.mp4	Montjuic_esp.txt	ESPAÑOL 	España in sighth
35	Ruta de los pueblos de la Edad Media.mp4	00:01:46	PueblosEdadMedia_esp.mp4	PueblosEdadMedia_esp.txt	ESPAÑOL 	-
36	Ruta del Románico en Aragón.mp4	00:01:34	RomanicoAragon_esp.mp4	RomanicoAragon_esp.txt	ESPAÑOL 	-
37	Salamanca ciudad Patrimonio de la Humanidad.mp4	00:05:29	Salamanca_esp.mp4	Salamanca_esp.txt	ESPAÑOL 	España
38	Santiago de Compostela ciudad Patrimonio de la Humanidad.mp4	00:05:31	SantiagoCompostela_esp.mp4	SantiagoCompostela_esp.txt	ESPAÑOL 	España
39	Segovia World Heritage City.mp4	00:03:47	Segovia_eng.mp4	Segovia_eng.txt	INGLÉS 	España in sighth
40	Segovia Ciudad Patrimonio de la Humanidad.mp4	00:03:47	Segovia_esp.mp4	Segovia_esp.txt	ESPAÑOL 	España in sighth
41	Toledo World Heritage City.mp4	00:03:48	Toledo_eng.mp4	Toledo_eng.txt	INGLÉS 	España in sighth
42	Toledo Ciudad Patrimonio de la Humanidad.mp4	00:03:48	Toledo_esp.mp4	Toledo_esp.txt	ESPAÑOL 	España in sighth
43	Úbeda. a World Heritage site.mp4	00:03:36	Ubeda_eng.mp4	Ubeda_eng.txt	INGLÉS 	España in sighth
44	Úbeda. Patrimonio de la Humanidad.mp4	00:03:36	Ubeda_esp.mp4	Ubeda_esp.txt	ESPAÑOL 	España in sighth
45	Enjoy art in Valencia.mp4	00:03:34	Valencia_eng.mp4	Valencia_eng.txt	INGLÉS 	España in sighth
46	Disfrutar del arte en Valencia.mp4	00:03:34	Valencia_esp.mp4	Valencia_esp.txt	ESPAÑOL 	España in sighth

B

Anexo B: Tablas de Resultados

B1. Tiempos y porcentajes tras la aplicación del Sistema de Segmentación de Audio ATVS-UAM.

ID	VIDEO	DUR (M)	TIEMPOS REALES			REAL	AUDIO SEGMENTATION ATVS-UAM (.fpg)			
			NOVOZ (s)	LOC01 (s)	DUR (s)		TIEMPOS		PORCENTAJES	
							NOVOZ	LOC01	NOVOZ	LOC01
1	Alcala_esp.mp4	00:06:11	141,373	229,627	371,000	RECON				
						NOVOZ	137,372	3,996	97,17%	1,74%
						LOC01	4,001	225,631	2,83%	98,26%
2	Andalucia_eng.mp4	00:02:30	56,778	93,222	150,000	NOVOZ	54,717	2,955	96,37%	3,17%
						LOC01	2,061	90,267	3,63%	96,83%
3	Andalucia_esp.mp4	00:02:30	56,778	93,222	150,000	NOVOZ	55,177	1,016	97,18%	1,09%
						LOC01	1,601	92,206	2,82%	98,91%
4	Aranjuez_eng.mp4	00:03:38	84,686	133,314	218,000	NOVOZ	82,120	1,453	96,97%	1,09%
						LOC01	2,566	131,861	3,03%	98,91%
5	Aranjuez_esp.mp4	00:03:38	84,686	133,314	218,000	NOVOZ	82,730	1,960	97,69%	1,47%
						LOC01	1,956	131,354	2,31%	98,53%
6	Avila_esp.mp4	00:05:26	209,293	116,707	326,000	NOVOZ	202,114	7,166	96,57%	6,14%
						LOC01	7,179	109,541	3,43%	93,86%
7	Baeza_eng.mp4	00:03:58	102,994	135,006	238,000	NOVOZ	97,082	4,739	94,26%	3,51%
						LOC01	5,912	130,267	5,74%	96,49%
8	Baeza_esp.mp4	00:03:58	102,994	135,006	238,000	NOVOZ	97,175	5,522	94,35%	4,09%
						LOC01	5,819	129,484	5,65%	95,91%
9	Bcnngotic_eng.mp4	00:02:10	80,769	49,231	130,000	NOVOZ	77,942	2,329	96,50%	4,73%
						LOC01	2,827	46,902	3,50%	95,27%
10	Bcnngotic_esp.mp4	00:02:10	80,769	49,231	130,000	NOVOZ	79,590	0,862	98,54%	1,75%
						LOC01	1,179	48,369	1,46%	98,25%
11	Bcnmodernista_eng.mp4	00:01:45	58,356	46,644	105,000	NOVOZ	20,279	4,977	34,75%	10,67%
						LOC01	38,077	41,667	65,25%	89,33%
12	Bcnmodernista_esp.mp4	00:01:45	58,356	46,644	105,000	NOVOZ	23,687	1,460	40,59%	3,13%
						LOC01	34,669	45,184	59,41%	96,87%
13	Caceres_esp.mp4	00:05:37	240,305	96,695	337,000	NOVOZ	230,068	10,279	95,74%	10,63%
						LOC01	10,237	86,416	4,26%	89,37%
14	CaminoSantiago_esp.mp4	00:03:44	90,755	133,245	224,000	NOVOZ	80,817	9,940	89,05%	7,46%
						LOC01	9,938	123,305	10,95%	92,54%
15	CastillaLeon_esp.mp4	00:01:47	67,247	39,753	107,000	NOVOZ	67,005	0,239	99,64%	0,60%
						LOC01	0,242	39,514	0,36%	99,40%
16	Cid_esp.mp4	00:03:28	111,004	96,996	208,000	NOVOZ	90,790	20,767	81,79%	21,41%
						LOC01	20,214	76,229	18,21%	78,59%
17	Cuenca_esp.mp4	00:05:42	172,335	169,665	342,000	NOVOZ	166,269	6,040	96,48%	3,56%
						LOC01	6,066	163,625	3,52%	96,44%
18	Dalissurre_eng.mp4	00:03:25	93,267	111,733	205,000	NOVOZ	91,085	2,313	97,66%	2,07%
						LOC01	2,182	109,420	2,34%	97,93%
19	Dalissurre_esp.mp4	00:03:25	93,267	111,733	205,000	NOVOZ	92,288	0,983	98,95%	0,88%
						LOC01	0,979	110,750	1,05%	99,12%
20	Esorial_esp.mp4	00:01:36	24,291	71,709	96,000	NOVOZ	22,530	10,398	92,75%	14,50%
						LOC01	1,761	61,311	7,25%	85,50%
21	Extremadura_eng.mp4	00:03:24	99,726	104,274	204,000	NOVOZ	95,498	4,307	95,76%	4,13%
						LOC01	4,228	99,967	4,24%	95,87%
22	Extremadura_esp.mp4	00:03:24	99,726	104,274	204,000	NOVOZ	90,571	8,519	90,82%	8,17%
						LOC01	9,155	95,755	9,18%	91,83%
23	Gaudi_eng.mp4	00:02:08	73,697	54,303	128,000	NOVOZ	67,050	6,125	90,90%	11,20%
						LOC01	6,647	48,178	9,02%	88,72%
24	Gaudi_esp.mp4	00:02:08	73,697	54,303	128,000	NOVOZ	68,214	5,479	92,56%	10,09%
						LOC01	5,483	48,824	7,44%	89,91%
25	GaudiGenio_eng.mp4	00:03:31	56,470	154,530	211,000	NOVOZ	54,324	2,009	96,20%	1,30%
						LOC01	2,146	152,521	3,80%	98,70%
26	GaudiGenio_esp.mp4	00:03:31	56,470	154,530	211,000	NOVOZ	49,581	6,861	87,80%	4,44%
						LOC01	6,889	147,669	12,20%	95,56%
27	Ibiza_esp.mp4	00:05:57	197,169	159,831	357,000	NOVOZ	195,473	1,694	99,14%	1,06%
						LOC01	1,696	158,137	0,86%	98,94%
28	Laguna_esp.mp4	00:05:31	174,227	156,773	331,000	NOVOZ	166,300	8,027	95,45%	5,12%
						LOC01	7,927	148,746	4,55%	94,88%
29	LaMancha_eng.mp4	00:02:49	88,685	80,315	169,000	NOVOZ	68,438	41,306	77,17%	51,43%
						LOC01	20,247	39,009	22,83%	48,57%
30	LaMancha_esp.mp4	00:02:49	88,685	80,315	169,000	NOVOZ	49,415	38,728	55,72%	48,22%
						LOC01	39,270	41,587	44,28%	51,78%
31	Madrid_eng.mp4	00:01:56	37,545	78,455	116,000	NOVOZ	0,000	4,503	0,00%	5,74%
						LOC01	37,545	73,952	100,00%	94,26%
32	Madrid_esp.mp4	00:01:56	37,545	78,455	116,000	NOVOZ	0,000	1,122	0,00%	1,43%
						LOC01	37,545	77,333	100,00%	98,57%
33	Montjuic_eng.mp4	00:01:21	43,088	37,912	81,000	NOVOZ	34,203	1,619	79,38%	4,27%
						LOC01	8,885	36,293	20,62%	95,73%
34	Montjuic_esp.mp4	00:01:21	43,088	37,912	81,000	NOVOZ	34,755	0,182	80,66%	0,48%
						LOC01	8,333	37,730	19,34%	99,52%
35	PueblosEdadMedia_esp.mp4	00:01:46	36,826	69,174	106,000	NOVOZ	0,000	0,297	0,00%	0,43%
						LOC01	36,826	68,877	100,00%	99,57%
36	RomanicoAragon_esp.mp4	00:01:34	26,353	67,647	94,000	NOVOZ	18,821	2,280	71,42%	3,37%
						LOC01	7,532	65,367	28,58%	96,63%
37	Salamanca_esp.mp4	00:05:29	181,467	147,533	329,000	NOVOZ	176,531	4,942	97,28%	3,35%
						LOC01	4,936	142,591	2,72%	96,65%
38	SantiagoCompostela_esp.mp4	00:05:31	182,655	148,345	331,000	NOVOZ	173,011	9,019	94,72%	6,08%
						LOC01	9,644	139,326	5,28%	93,92%
39	Segovia_eng.mp4	00:03:47	120,469	106,531	227,000	NOVOZ	120,083	0,415	99,68%	0,39%
						LOC01	0,386	106,116	0,32%	99,61%
40	Segovia_esp.mp4	00:03:47	120,469	106,531	227,000	NOVOZ	114,843	5,380	95,33%	5,05%
						LOC01	5,626	101,151	4,67%	94,95%
41	Toledo_eng.mp4	00:03:48	129,071	98,929	228,000	NOVOZ	123,327	5,738	95,55%	5,80%
						LOC01	5,744	93,191	4,45%	94,20%
42	Toledo_esp.mp4	00:03:48	129,071	98,929	228,000	NOVOZ	126,012	2,582	97,63%	2,61%
						LOC01	3,059	96,347	2,37%	97,39%
43	Ubeda_eng.mp4	00:03:36	108,396	107,604	216,000	NOVOZ	104,407	4,175	96,32%	3,88%
						LOC01	3,989	103,429	3,68%	96,12%
44	Ubeda_esp.mp4	00:03:36	108,396	107,604	216,000	NOVOZ	100,613	7,683	92,82%	7,14%
						LOC01	7,783	99,921	7,18%	92,86%
45	Valencia_eng.mp4	00:03:34	76,512	137,488	214,000	NOVOZ	70,238	6,352	91,80%	4,62%
						LOC01	6,274	131,136	8,20%	95,38%
46	Valencia_esp.mp4	00:03:34	76,512	137,488	214,000	NOVOZ	76,030	0,481	99,37%	0,35%
						LOC01	0,482	137,007	0,63%	99,65%

B2. Tiempos y porcentajes tras la aplicación del Reconocedor fonético para segmentación de audio.

ID	VIDEO	DUR (M)	TIEMPOS REALES			REAL	RECONOCIMIENTO FONÉTICO (recout.train.mlf)			
			NOVOZ (s)	LOC01 (s)	DUR (s)		TIEMPOS		PORCENTAJES	
							NOVOZ	LOC01	NOVOZ	LOC01
1	Alcala_esp.mp4	00:06:11	141,373	229,627	371,000	NOVOZ	135,364	96,636	94,37%	42,03%
						LOC01	8,073	131,927	5,63%	57,97%
2	Andalucia_eng.mp4	00:02:30	56,778	93,222	150,000	NOVOZ	63,558	4,442	89,92%	5,60%
						LOC01	7,126	74,874	10,08%	94,40%
3	Andalucia_esp.mp4	00:02:30	56,778	93,222	150,000	NOVOZ	52,265	11,735	92,05%	12,59%
						LOC01	4,513	81,487	7,95%	87,41%
4	Aranjuez_eng.mp4	00:03:38	84,686	133,314	218,000	NOVOZ	79,537	52,463	89,16%	40,73%
						LOC01	9,671	76,329	10,84%	59,27%
5	Aranjuez_esp.mp4	00:03:38	84,686	133,314	218,000	NOVOZ	77,937	26,063	90,96%	19,62%
						LOC01	7,742	106,758	9,04%	80,38%
6	Avila_esp.mp4	00:05:26	209,293	116,707	326,000	NOVOZ	208,557	63,443	99,65%	54,36%
						LOC01	0,736	53,264	0,35%	45,64%
7	Baeza_eng.mp4	00:03:58	102,994	135,006	238,000	NOVOZ	77,129	24,871	85,11%	16,88%
						LOC01	13,490	122,510	14,89%	83,12%
8	Baeza_esp.mp4	00:03:58	102,994	135,006	238,000	NOVOZ	96,464	23,536	93,01%	17,53%
						LOC01	7,249	110,751	6,99%	82,47%
9	Bcngotico_eng.mp4	00:02:10	80,769	49,231	130,000	NOVOZ	69,981	2,019	85,63%	4,18%
						LOC01	11,740	46,260	14,37%	95,82%
10	Bcngotico_esp.mp4	00:02:10	80,769	49,231	130,000	NOVOZ	73,145	4,855	90,56%	9,86%
						LOC01	7,624	44,376	9,44%	90,14%
11	Bcnmodernista_eng.mp4	00:01:45	58,356	46,644	105,000	NOVOZ	44,221	4,779	78,58%	9,81%
						LOC01	12,057	43,943	21,42%	90,19%
12	Bcnmodernista_esp.mp4	00:01:45	58,356	46,644	105,000	NOVOZ	53,371	3,629	91,46%	7,78%
						LOC01	4,985	43,015	8,54%	92,22%
13	Caceres_esp.mp4	00:05:37	240,305	96,695	337,000	NOVOZ	237,655	51,345	98,90%	53,10%
						LOC01	2,650	45,350	1,10%	46,90%
14	CaminoSantiago_esp.mp4	00:03:44	90,755	133,245	224,000	NOVOZ	81,338	66,662	89,62%	50,03%
						LOC01	9,417	66,583	10,38%	49,97%
15	CastillaLeon_esp.mp4	00:01:47	67,247	39,753	107,000	NOVOZ	67,114	7,886	99,80%	19,84%
						LOC01	0,133	31,867	0,20%	80,16%
16	Cid_esp.mp4	00:03:28	111,004	96,996	208,000	NOVOZ	109,012	26,988	98,21%	27,82%
						LOC01	1,992	70,008	1,79%	72,18%
17	Cuenca_esp.mp4	00:05:42	172,335	169,665	342,000	NOVOZ	168,828	59,172	97,71%	34,97%
						LOC01	3,955	110,045	2,29%	65,03%
18	Dalisure_eng.mp4	00:03:25	93,267	111,733	205,000	NOVOZ	96,355	18,645	88,27%	19,45%
						LOC01	12,808	77,192	11,73%	80,55%
19	Dalisure_esp.mp4	00:03:25	93,267	111,733	205,000	NOVOZ	87,036	15,964	92,91%	14,34%
						LOC01	6,640	95,360	7,09%	85,66%
20	Escorial_esp.mp4	00:01:36	24,291	71,709	96,000	NOVOZ	20,486	21,514	78,75%	30,74%
						LOC01	5,528	48,472	21,25%	69,26%
21	Extremadura_eng.mp4	00:03:24	99,726	104,274	204,000	NOVOZ	95,478	20,522	94,66%	19,90%
						LOC01	5,382	82,618	5,34%	80,10%
22	Extremadura_esp.mp4	00:03:24	99,726	104,274	204,000	NOVOZ	94,405	9,595	94,42%	9,22%
						LOC01	5,580	94,420	5,58%	90,78%
23	Gaudi_eng.mp4	00:02:08	73,697	54,303	128,000	NOVOZ	64,056	25,944	89,43%	46,02%
						LOC01	7,569	30,431	10,57%	53,98%
24	Gaudi_esp.mp4	00:02:08	73,697	54,303	128,000	NOVOZ	71,367	32,633	96,84%	60,09%
						LOC01	2,330	21,670	3,16%	39,91%
25	GaudiGenio_eng.mp4	00:03:31	56,470	154,530	211,000	NOVOZ	48,437	83,563	85,93%	53,69%
						LOC01	7,930	72,070	14,07%	46,31%
26	GaudiGenio_esp.mp4	00:03:31	56,470	154,530	211,000	NOVOZ	51,651	59,349	91,11%	38,46%
						LOC01	5,039	94,961	8,89%	61,54%
27	Ibiza_esp.mp4	00:05:57	197,169	159,831	357,000	NOVOZ	197,169	147,831	100,00%	92,49%
						LOC01	0,000	12,000	0,00%	7,51%
28	Laguna_esp.mp4	00:05:31	174,227	156,773	331,000	NOVOZ	172,464	94,536	98,99%	60,30%
						LOC01	1,763	62,237	1,01%	39,70%
29	LaMancha_eng.mp4	00:02:49	88,685	80,315	169,000	NOVOZ	92,753	59,247	90,27%	90,80%
						LOC01	10,000	6,000	9,73%	9,20%
30	LaMancha_esp.mp4	00:02:49	88,685	80,315	169,000	NOVOZ	87,301	63,699	98,44%	79,31%
						LOC01	1,384	16,616	1,56%	20,69%
31	Madrid_eng.mp4	00:01:56	37,545	78,455	116,000	NOVOZ	35,558	6,442	74,11%	9,47%
						LOC01	12,425	61,575	25,89%	90,53%
32	Madrid_esp.mp4	00:01:56	37,545	78,455	116,000	NOVOZ	33,502	18,498	89,23%	23,58%
						LOC01	4,043	59,957	10,77%	76,42%
33	Montjuic_eng.mp4	00:01:21	43,088	37,912	81,000	NOVOZ	36,489	2,511	83,80%	6,70%
						LOC01	7,055	34,945	16,20%	93,30%
34	Montjuic_esp.mp4	00:01:21	43,088	37,912	81,000	NOVOZ	41,183	1,817	95,58%	4,79%
						LOC01	1,905	36,095	4,42%	95,21%
35	PueblosEdadMedia_esp.mp4	00:01:46	36,826	69,174	106,000	NOVOZ	36,762	43,238	99,83%	62,51%
						LOC01	0,064	25,936	0,17%	37,49%
36	RomanicoAragon_esp.mp4	00:01:34	26,353	67,647	94,000	NOVOZ	26,446	65,554	100,00%	97,04%
						LOC01	0,000	2,000	0,00%	2,96%
37	Salamanca_esp.mp4	00:05:29	181,467	147,533	329,000	NOVOZ	176,052	54,948	96,54%	37,47%
						LOC01	6,311	91,689	3,46%	62,53%
38	SantiagoCompostela_esp.mp4	00:05:31	182,655	148,345	331,000	NOVOZ	179,677	99,323	98,16%	67,13%
						LOC01	3,377	48,623	1,84%	32,87%
39	Segovia_eng.mp4	00:03:47	120,469	106,531	227,000	NOVOZ	116,224	26,776	94,07%	25,88%
						LOC01	7,329	76,671	5,93%	74,12%
40	Segovia_esp.mp4	00:03:47	120,469	106,531	227,000	NOVOZ	118,892	60,108	98,69%	56,42%
						LOC01	1,577	46,423	1,31%	43,58%
41	Toledo_eng.mp4	00:03:48	129,071	98,929	228,000	NOVOZ	122,491	51,509	94,38%	52,45%
						LOC01	7,298	46,702	5,62%	47,55%
42	Toledo_esp.mp4	00:03:48	129,071	98,929	228,000	NOVOZ	128,405	57,595	98,80%	58,75%
						LOC01	1,561	40,439	1,20%	41,25%
43	Ubeda_eng.mp4	00:03:36	108,396	107,604	216,000	NOVOZ	101,019	24,981	91,13%	23,76%
						LOC01	9,836	80,164	8,87%	76,24%
44	Ubeda_esp.mp4	00:03:36	108,396	107,604	216,000	NOVOZ	101,978	18,022	94,08%	16,75%
						LOC01	6,418	89,592	5,92%	83,25%
45	Valencia_eng.mp4	00:03:34	76,512	137,488	214,000	NOVOZ	73,234	9,245	87,29%	6,77%
						LOC01	10,667	127,333	12,71%	93,23%
46	Valencia_esp.mp4	00:03:34	76,512	137,488	214,000	NOVOZ	69,724	22,276	90,06%	16,31%
						LOC01	7,698	114,302	9,94%	83,69%

B3. Tiempos y porcentajes tras la aplicación del Reconocedor fonético con adaptación al locutor para segmentación de audio.

ID	VIDEO	DUR (M)	TIEMPOS REALES			REAL	ADAPTACION AL LOCUTOR RECONOCIMIENTO FONETICO (recout_adapt.train.mlf)			
			NOVOZ (s)	LOC01 (s)	DUR (s)		TIEMPOS		PORCENTAJES	
							NOVOZ	LOC01	NOVOZ	LOC01
1	Alcala_esp.mp4	00:06:11	141,373	229,627	371,000	NOVOZ	123,985	15,102	86,71%	6,63%
						LOC01	19,000	212,549	13,29%	93,37%
2	Andalucia_eng.mp4	00:02:30	56,778	93,222	150,000	NOVOZ	47,703	0,297	67,49%	0,37%
						LOC01	22,981	79,019	32,51%	99,63%
3	Andalucia_esp.mp4	00:02:30	56,778	93,222	150,000	NOVOZ	38,000	0,000	66,93%	0,00%
						LOC01	18,778	93,222	33,07%	100,00%
4	Aranjuez_eng.mp4	00:03:38	84,686	133,314	218,000	NOVOZ	72,561	7,439	81,34%	5,78%
						LOC01	16,647	121,353	18,66%	94,22%
5	Aranjuez_esp.mp4	00:03:38	84,686	133,314	218,000	NOVOZ	71,105	2,895	83,48%	2,18%
						LOC01	14,074	129,926	16,52%	97,82%
6	Avila_esp.mp4	00:05:26	209,293	116,707	326,000	NOVOZ	189,453	6,547	90,52%	5,61%
						LOC01	19,840	110,160	9,48%	94,39%
7	Baeza_eng.mp4	00:03:58	102,994	135,006	238,000	NOVOZ	76,784	1,216	84,73%	0,83%
						LOC01	13,835	146,165	15,27%	99,17%
8	Baeza_esp.mp4	00:03:58	102,994	135,006	238,000	NOVOZ	76,412	1,588	73,68%	1,18%
						LOC01	27,301	132,699	26,32%	98,82%
9	Bcngotico_eng.mp4	00:02:10	80,769	49,231	130,000	NOVOZ	70,851	1,149	86,70%	2,38%
						LOC01	10,870	47,130	13,30%	97,62%
10	Bcngotico_esp.mp4	00:02:10	80,769	49,231	130,000	NOVOZ	62,000	0,000	76,76%	0,00%
						LOC01	18,769	49,231	23,24%	100,00%
11	Bcnmodernista_eng.mp4	00:01:45	58,356	46,644	105,000	NOVOZ	40,978	0,022	72,81%	0,05%
						LOC01	15,300	48,700	27,19%	99,95%
12	Bcnmodernista_esp.mp4	00:01:45	58,356	46,644	105,000	NOVOZ	33,000	0,000	56,55%	0,00%
						LOC01	25,356	46,644	43,45%	100,00%
13	Caceres_esp.mp4	00:05:37	240,305	96,695	337,000	NOVOZ	157,223	15,777	65,43%	16,32%
						LOC01	83,082	80,918	34,57%	83,68%
14	CaminoSantiago_esp.mp4	00:03:44	90,755	133,245	224,000	NOVOZ	78,420	7,580	86,41%	5,69%
						LOC01	12,335	125,665	13,59%	94,31%
15	CastillaLeon_esp.mp4	00:01:47	67,247	39,753	107,000	NOVOZ	51,650	1,350	76,81%	3,40%
						LOC01	15,597	38,403	23,19%	96,60%
16	Cid_esp.mp4	00:03:28	111,004	96,996	208,000	NOVOZ	94,323	2,677	84,97%	2,76%
						LOC01	16,681	94,319	15,03%	97,24%
17	Cuenca_esp.mp4	00:05:42	172,335	169,665	342,000	NOVOZ	156,063	15,937	90,32%	9,42%
						LOC01	16,720	153,280	9,68%	90,58%
18	Dalissurre_eng.mp4	00:03:25	93,267	111,733	205,000	NOVOZ	91,262	3,738	83,60%	3,90%
						LOC01	17,901	92,099	16,40%	96,10%
19	Dalissurre_esp.mp4	00:03:25	93,267	111,733	205,000	NOVOZ	69,638	1,362	74,34%	1,22%
						LOC01	24,038	109,962	25,66%	98,78%
20	Escorial_esp.mp4	00:01:36	24,291	71,709	96,000	NOVOZ	15,200	3,800	58,43%	5,43%
						LOC01	10,814	66,186	41,57%	94,57%
21	Extremadura_eng.mp4	00:03:24	99,726	104,274	204,000	NOVOZ	86,969	1,031	86,23%	1,00%
						LOC01	13,891	102,109	13,77%	99,00%
22	Extremadura_esp.mp4	00:03:24	99,726	104,274	204,000	NOVOZ	88,373	1,627	88,39%	1,66%
						LOC01	11,612	102,388	11,61%	98,44%
23	Gaudi_eng.mp4	00:02:08	73,697	54,303	128,000	NOVOZ	61,860	6,140	86,37%	10,89%
						LOC01	9,765	50,235	13,63%	89,11%
24	Gaudi_esp.mp4	00:02:08	73,697	54,303	128,000	NOVOZ	65,593	12,407	89,00%	22,85%
						LOC01	8,104	41,896	11,00%	77,15%
25	GaudiGenio_eng.mp4	00:03:31	56,470	154,530	211,000	NOVOZ	35,621	22,379	63,19%	14,38%
						LOC01	20,746	133,254	36,81%	85,62%
26	GaudiGenio_esp.mp4	00:03:31	56,470	154,530	211,000	NOVOZ	29,970	14,030	52,87%	9,09%
						LOC01	26,720	140,280	47,13%	90,91%
27	Ibiza_esp.mp4	00:05:57	197,169	159,831	357,000	NOVOZ	188,731	60,269	95,72%	37,71%
						LOC01	8,438	99,562	4,28%	62,29%
28	Laguna_esp.mp4	00:05:31	174,227	156,773	331,000	NOVOZ	159,789	25,211	91,71%	16,08%
						LOC01	14,438	131,562	8,29%	83,92%
29	LaMancha_eng.mp4	00:02:49	88,685	80,315	169,000	NOVOZ	80,722	33,278	78,56%	51,00%
						LOC01	22,031	31,969	21,44%	49,00%
30	LaMancha_esp.mp4	00:02:49	88,685	80,315	169,000	NOVOZ	69,305	33,695	78,15%	41,95%
						LOC01	19,380	46,620	21,85%	58,05%
31	Madrid_eng.mp4	00:01:56	37,545	78,455	116,000	NOVOZ	31,343	0,657	65,32%	0,97%
						LOC01	16,640	67,360	34,68%	99,03%
32	Madrid_esp.mp4	00:01:56	37,545	78,455	116,000	NOVOZ	26,000	0,000	69,25%	0,00%
						LOC01	11,545	36,360	30,75%	100,00%
33	Montjuic_eng.mp4	00:01:21	43,088	37,912	81,000	NOVOZ	21,000	0,000	48,23%	0,00%
						LOC01	22,544	37,456	51,77%	100,00%
34	Montjuic_esp.mp4	00:01:21	43,088	37,912	81,000	NOVOZ	25,000	0,000	58,02%	0,00%
						LOC01	18,088	37,192	41,98%	100,00%
35	PueblosEdadMedia_esp.mp4	00:01:46	36,826	69,174	106,000	NOVOZ	31,970	9,030	86,81%	13,05%
						LOC01	4,856	60,144	13,19%	86,95%
36	RomanicoAragon_esp.mp4	00:01:34	26,353	67,647	94,000	NOVOZ	21,326	35,674	80,64%	52,81%
						LOC01	5,120	31,880	19,36%	47,19%
37	Salamanca_esp.mp4	00:05:29	181,467	147,533	329,000	NOVOZ	156,678	16,322	85,92%	11,13%
						LOC01	25,685	130,315	14,08%	88,87%
38	SantiagoCompostela_esp.mp4	00:05:31	182,655	148,345	331,000	NOVOZ	169,092	25,908	92,37%	17,51%
						LOC01	13,962	122,038	7,63%	82,49%
39	Segovia_eng.mp4	00:03:47	120,469	106,531	227,000	NOVOZ	108,611	6,389	87,91%	6,18%
						LOC01	14,942	97,058	12,09%	93,82%
40	Segovia_esp.mp4	00:03:47	120,469	106,531	227,000	NOVOZ	105,987	11,013	87,98%	10,34%
						LOC01	14,482	95,518	12,02%	89,66%
41	Toledo_eng.mp4	00:03:48	129,071	98,929	228,000	NOVOZ	120,547	21,453	92,88%	21,84%
						LOC01	9,242	76,758	7,12%	78,16%
42	Toledo_esp.mp4	00:03:48	129,071	98,929	228,000	NOVOZ	108,513	11,487	83,49%	11,72%
						LOC01	21,453	86,547	16,51%	88,28%
43	Ubeda_eng.mp4	00:03:36	108,396	107,604	216,000	NOVOZ	92,117	1,883	83,10%	1,79%
						LOC01	18,738	103,262	16,90%	98,21%
44	Ubeda_esp.mp4	00:03:36	108,396	107,604	216,000	NOVOZ	93,500	2,500	86,26%	2,32%
						LOC01	14,896	105,104	13,74%	97,68%
45	Valencia_eng.mp4	00:03:34	76,512	137,488	214,000	NOVOZ	72,198	31,802	91,45%	23,55%
						LOC01	6,746	103,254	8,55%	76,45%
46	Valencia_esp.mp4	00:03:34	76,512	137,488	214,000	NOVOZ	59,778	0,222	77,21%	0,16%
						LOC01	17,644	136,356	22,79%	99,84%

B4. Tiempos y porcentajes tras la combinación de resultados de los sistemas de Segmentación de Audio y Rec. fonético con la función lógica AND.

ID	VIDEO	DUR (M)	TIEMPOS REALES			AND				
			NOVOZ (s)	LOC01 (s)	DUR (s)	REAL RECON	TIEMPOS		PORCENTAJES	
							NOVOZ	LOC01	NOVOZ	LOC01
1	Alcala_esp.mp4	00:06:11	141,373	229,627	371,000	NOVOZ	130,230	70,151	92,12%	30,55%
						LOC01	11,143	159,476	7,88%	69,45%
2	Andalucia_eng.mp4	00:02:30	56,778	93,222	150,000	NOVOZ	49,533	7,970	87,24%	8,56%
						LOC01	7,245	85,252	12,76%	91,46%
3	Andalucia_esp.mp4	00:02:30	56,778	93,222	150,000	NOVOZ	50,686	9,509	89,27%	10,20%
						LOC01	6,092	83,713	10,73%	89,80%
4	Aranjuez_eng.mp4	00:03:38	84,686	133,314	218,000	NOVOZ	81,548	59,774	96,30%	44,84%
						LOC01	3,138	73,540	3,71%	55,16%
5	Aranjuez_esp.mp4	00:03:38	84,686	133,314	218,000	NOVOZ	77,689	14,198	91,74%	10,65%
						LOC01	6,997	119,116	8,26%	89,35%
6	Avila_esp.mp4	00:05:26	209,293	116,707	326,000	NOVOZ	204,500	40,427	97,71%	34,64%
						LOC01	4,793	76,280	2,29%	65,36%
7	Baeza_eng.mp4	00:03:58	102,994	135,006	238,000	NOVOZ	100,379	30,330	97,46%	22,47%
						LOC01	2,615	104,676	2,54%	77,53%
8	Baeza_esp.mp4	00:03:58	102,994	135,006	238,000	NOVOZ	97,298	21,520	94,47%	16,94%
						LOC01	5,696	113,486	5,53%	84,06%
9	Bcngotico_eng.mp4	00:02:10	80,769	49,231	130,000	NOVOZ	78,980	8,724	97,79%	17,72%
						LOC01	1,789	40,507	2,22%	82,28%
10	Bcngotico_esp.mp4	00:02:10	80,769	49,231	130,000	NOVOZ	76,157	7,197	94,29%	14,62%
						LOC01	4,612	42,034	5,71%	85,38%
11	Bcnmodernista_eng.mp4	00:01:45	58,356	46,644	105,000	NOVOZ	39,799	9,005	68,20%	19,31%
						LOC01	18,557	37,639	31,80%	80,70%
12	Bcnmodernista_esp.mp4	00:01:45	58,356	46,644	105,000	NOVOZ	32,358	18,121	55,45%	38,85%
						LOC01	25,998	28,523	44,55%	61,15%
13	Caceres_esp.mp4	00:05:37	240,305	96,695	337,000	NOVOZ	238,565	49,527	99,28%	51,22%
						LOC01	1,740	47,168	0,72%	48,78%
14	CaminoSantiago_esp.mp4	00:03:44	90,755	133,245	224,000	NOVOZ	78,793	47,635	86,82%	35,75%
						LOC01	11,962	85,610	13,18%	64,25%
15	CastillaLeon_esp.mp4	00:01:47	67,247	39,753	107,000	NOVOZ	65,129	6,599	96,85%	16,60%
						LOC01	2,118	33,154	3,15%	83,40%
16	Cid_esp.mp4	00:03:28	111,004	96,996	208,000	NOVOZ	108,318	31,776	97,58%	32,76%
						LOC01	2,686	65,220	2,42%	67,24%
17	Cuenca_esp.mp4	00:05:42	172,335	169,665	342,000	NOVOZ	166,648	43,689	96,70%	25,75%
						LOC01	5,687	125,976	3,30%	74,25%
18	Dalissime_eng.mp4	00:03:25	93,267	111,733	205,000	NOVOZ	89,023	28,067	95,45%	25,12%
						LOC01	4,244	83,666	4,55%	74,88%
19	Dalissime_esp.mp4	00:03:25	93,267	111,733	205,000	NOVOZ	83,706	12,592	89,75%	11,27%
						LOC01	9,561	99,141	10,25%	88,73%
20	Escorial_esp.mp4	00:01:36	24,291	71,709	96,000	NOVOZ	18,709	13,651	77,02%	19,04%
						LOC01	5,582	58,058	22,98%	80,96%
21	Extremadura_eng.mp4	00:03:24	99,726	104,274	204,000	NOVOZ	91,337	23,086	91,59%	22,14%
						LOC01	8,389	81,188	8,41%	77,86%
22	Extremadura_esp.mp4	00:03:24	99,726	104,274	204,000	NOVOZ	91,939	19,059	92,19%	18,28%
						LOC01	7,787	85,215	7,81%	81,72%
23	Gaudi_eng.mp4	00:02:08	73,697	54,303	128,000	NOVOZ	69,503	25,474	94,31%	46,91%
						LOC01	4,194	28,829	5,69%	53,09%
24	Gaudi_esp.mp4	00:02:08	73,697	54,303	128,000	NOVOZ	68,680	21,110	93,19%	38,87%
						LOC01	5,017	33,193	6,81%	61,13%
25	GaudiGenio_eng.mp4	00:03:31	56,470	154,530	211,000	NOVOZ	49,504	86,073	87,66%	55,70%
						LOC01	6,966	68,457	12,34%	44,30%
26	GaudiGenio_esp.mp4	00:03:31	56,470	154,530	211,000	NOVOZ	45,930	49,278	81,34%	31,89%
						LOC01	10,540	105,252	18,66%	68,11%
27	Ibiza_esp.mp4	00:05:57	197,169	159,831	357,000	NOVOZ	194,008	133,906	98,40%	83,78%
						LOC01	3,161	25,925	1,60%	16,22%
28	Laguna_esp.mp4	00:05:31	174,227	156,773	331,000	NOVOZ	169,737	78,111	97,42%	49,82%
						LOC01	4,490	78,662	2,58%	50,18%
29	LaMancha_eng.mp4	00:02:49	88,685	80,315	169,000	NOVOZ	38,631	62,402	43,56%	77,70%
						LOC01	50,054	17,913	56,44%	22,30%
30	LaMancha_esp.mp4	00:02:49	88,685	80,315	169,000	NOVOZ	85,785	56,614	96,73%	70,49%
						LOC01	2,900	23,701	3,27%	29,51%
31	Madrid_eng.mp4	00:01:56	37,545	78,455	116,000	NOVOZ	28,574	9,461	76,11%	12,06%
						LOC01	8,971	68,994	23,90%	87,94%
32	Madrid_esp.mp4	00:01:56	37,545	78,455	116,000	NOVOZ	28,684	9,022	76,40%	11,50%
						LOC01	8,861	69,433	23,60%	88,50%
33	Montjuic_eng.mp4	00:01:21	43,088	37,912	81,000	NOVOZ	33,445	9,838	77,62%	25,95%
						LOC01	9,643	28,074	22,38%	74,05%
34	Montjuic_esp.mp4	00:01:21	43,088	37,912	81,000	NOVOZ	32,620	4,159	75,71%	10,97%
						LOC01	10,468	33,753	24,30%	89,03%
35	PueblosEdadMedia_esp.mp4	00:01:46	36,826	69,174	106,000	NOVOZ	31,424	34,721	85,33%	50,19%
						LOC01	5,402	34,453	14,67%	49,81%
36	RomanicoAragon_esp.mp4	00:01:34	26,353	67,647	94,000	NOVOZ	18,194	59,665	69,04%	88,20%
						LOC01	8,159	7,982	30,96%	11,80%
37	Salamanca_esp.mp4	00:05:29	181,467	147,533	329,000	NOVOZ	172,648	44,629	95,14%	30,25%
						LOC01	8,819	102,904	4,86%	69,75%
38	SantiagoCompostela_esp.mp4	00:05:31	182,655	148,345	331,000	NOVOZ	177,738	75,508	97,31%	50,90%
						LOC01	4,917	72,837	2,69%	49,10%
39	Segovia_eng.mp4	00:03:47	120,469	106,531	227,000	NOVOZ	117,645	31,009	97,66%	29,11%
						LOC01	2,824	75,522	2,34%	70,89%
40	Segovia_esp.mp4	00:03:47	120,469	106,531	227,000	NOVOZ	115,673	36,015	96,02%	33,81%
						LOC01	4,796	70,516	3,98%	66,19%
41	Toledo_eng.mp4	00:03:48	129,071	98,929	228,000	NOVOZ	122,522	53,262	94,93%	53,84%
						LOC01	6,549	45,667	5,07%	46,16%
42	Toledo_esp.mp4	00:03:48	129,071	98,929	228,000	NOVOZ	124,489	41,090	96,45%	41,54%
						LOC01	4,582	57,839	3,55%	58,47%
43	Ubeda_eng.mp4	00:03:36	108,396	107,604	216,000	NOVOZ	106,613	30,555	98,36%	28,40%
						LOC01	1,783	77,049	1,65%	71,60%
44	Ubeda_esp.mp4	00:03:36	108,396	107,604	216,000	NOVOZ	102,472	12,006	94,54%	11,16%
						LOC01	5,924	95,598	5,47%	88,84%
45	Valencia_eng.mp4	00:03:34	76,512	137,488	214,000	NOVOZ	73,948	72,045	96,65%	52,40%
						LOC01	2,564	65,443	3,35%	47,60%
46	Valencia_esp.mp4	00:03:34	76,512	137,488	214,000	NOVOZ	69,471	13,797	90,80%	10,04%
						LOC01	7,041	123,691	9,20%	89,97%

B5. Tiempos y porcentajes tras la combinación de resultados de los sistemas de Segmentación de Audio y Rec. fonético con la función lógica OR.

ID	VIDEO	DUR (M)	TIEMPOS REALES			OR				
			NOVOZ (s)	LOC01 (s)	DUR (s)	REAL	TIEMPOS		PORCENTAJES	
							NOVOZ	LOC01	NOVOZ	LOC01
1	Alcala_esp.mp4	00:06:11	141,373	229,627	371,000	NOVOZ	118,581	10,007	83,88%	4,36%
						LOC01	22,792	219,620	16,12%	95,64%
2	Andalucia_eng.mp4	00:02:30	56,778	93,222	150,000	NOVOZ	45,644	1,542	80,39%	1,65%
						LOC01	11,134	91,680	19,61%	98,35%
3	Andalucia_esp.mp4	00:02:30	56,778	93,222	150,000	NOVOZ	51,254	4,981	90,27%	5,34%
						LOC01	5,524	88,241	9,73%	94,66%
4	Aranjuez_eng.mp4	00:03:38	84,686	133,314	218,000	NOVOZ	71,678	5,333	84,64%	4,00%
						LOC01	13,008	127,981	15,36%	96,00%
5	Aranjuez_esp.mp4	00:03:38	84,686	133,314	218,000	NOVOZ	68,032	3,737	80,34%	2,80%
						LOC01	16,654	129,577	19,67%	97,20%
6	Avila_esp.mp4	00:05:26	209,293	116,707	326,000	NOVOZ	185,404	10,404	88,59%	8,92%
						LOC01	23,889	106,303	11,41%	91,09%
7	Baeza_eng.mp4	00:03:58	102,994	135,006	238,000	NOVOZ	85,170	4,010	82,69%	2,97%
						LOC01	17,824	130,996	17,31%	97,03%
8	Baeza_esp.mp4	00:03:58	102,994	135,006	238,000	NOVOZ	82,933	6,953	80,52%	5,15%
						LOC01	20,061	128,053	19,48%	94,85%
9	Bcngotico_eng.mp4	00:02:10	80,769	49,231	130,000	NOVOZ	69,251	3,096	85,74%	6,29%
						LOC01	11,518	46,135	14,26%	93,71%
10	Bcngotico_esp.mp4	00:02:10	80,769	49,231	130,000	NOVOZ	69,663	1,226	86,25%	2,49%
						LOC01	11,106	48,005	13,75%	97,51%
11	Bcnmodernista_eng.mp4	00:01:45	58,356	46,644	105,000	NOVOZ	45,418	7,790	77,83%	16,70%
						LOC01	12,938	38,854	22,17%	83,30%
12	Bcnmodernista_esp.mp4	00:01:45	58,356	46,644	105,000	NOVOZ	40,102	1,096	68,72%	2,35%
						LOC01	18,254	45,548	31,28%	97,65%
13	Caceres_esp.mp4	00:05:37	240,305	96,695	337,000	NOVOZ	210,483	7,800	87,59%	8,07%
						LOC01	29,822	88,895	12,41%	91,93%
14	CaminoSantiago_esp.mp4	00:03:44	90,755	133,245	224,000	NOVOZ	65,761	7,835	72,46%	5,88%
						LOC01	24,994	125,410	27,54%	94,12%
15	CastillaLeon_esp.mp4	00:01:47	67,247	39,753	107,000	NOVOZ	61,811	1,618	91,92%	4,07%
						LOC01	5,436	38,135	8,08%	95,93%
16	Cid_esp.mp4	00:03:28	111,004	96,996	208,000	NOVOZ	98,694	5,645	88,91%	5,82%
						LOC01	12,310	91,351	11,09%	94,18%
17	Cuenca_esp.mp4	00:05:42	172,335	169,665	342,000	NOVOZ	155,291	11,732	90,11%	6,92%
						LOC01	17,044	157,933	9,89%	93,09%
18	Dalissurre_eng.mp4	00:03:25	93,267	111,733	205,000	NOVOZ	76,610	7,615	82,14%	6,82%
						LOC01	16,657	104,118	17,86%	93,19%
19	Dalissurre_esp.mp4	00:03:25	93,267	111,733	205,000	NOVOZ	75,968	6,682	81,45%	5,98%
						LOC01	17,299	105,051	18,55%	94,02%
20	Escorial_esp.mp4	00:01:36	24,291	71,709	96,000	NOVOZ	11,067	2,538	45,56%	3,54%
						LOC01	13,224	69,171	54,44%	96,46%
21	Extremadura_eng.mp4	00:03:24	99,726	104,274	204,000	NOVOZ	89,335	6,632	89,58%	6,36%
						LOC01	10,391	97,642	10,42%	93,64%
22	Extremadura_esp.mp4	00:03:24	99,726	104,274	204,000	NOVOZ	77,479	4,789	77,69%	4,59%
						LOC01	22,247	99,485	22,31%	95,41%
23	Gaudi_eng.mp4	00:02:08	73,697	54,303	128,000	NOVOZ	52,726	1,553	71,54%	2,86%
						LOC01	20,971	52,750	28,46%	97,14%
24	Gaudi_esp.mp4	00:02:08	73,697	54,303	128,000	NOVOZ	58,067	0,132	78,79%	0,24%
						LOC01	15,630	54,171	21,21%	99,76%
25	GaudiGenio_eng.mp4	00:03:31	56,470	154,530	211,000	NOVOZ	32,679	4,719	57,87%	3,05%
						LOC01	23,791	149,811	42,13%	96,95%
26	GaudiGenio_esp.mp4	00:03:31	56,470	154,530	211,000	NOVOZ	33,576	7,294	59,46%	4,72%
						LOC01	22,894	147,236	40,54%	95,28%
27	Ibiza_esp.mp4	00:05:57	197,169	159,831	357,000	NOVOZ	175,342	6,857	88,93%	4,29%
						LOC01	21,827	152,974	11,07%	95,71%
28	Laguna_esp.mp4	00:05:31	174,227	156,773	331,000	NOVOZ	149,591	9,203	85,86%	5,87%
						LOC01	24,636	147,570	14,14%	94,13%
29	LaMancha_eng.mp4	00:02:49	88,685	80,315	169,000	NOVOZ	82,140	36,945	92,62%	46,00%
						LOC01	6,545	43,370	7,38%	54,00%
30	LaMancha_esp.mp4	00:02:49	88,685	80,315	169,000	NOVOZ	69,902	30,584	78,82%	38,08%
						LOC01	18,783	49,731	21,18%	61,92%
31	Madrid_eng.mp4	00:01:56	37,545	78,455	116,000	NOVOZ	26,965	9,948	71,82%	12,68%
						LOC01	10,580	68,507	28,18%	87,32%
32	Madrid_esp.mp4	00:01:56	37,545	78,455	116,000	NOVOZ	20,417	6,553	54,38%	8,35%
						LOC01	17,128	71,902	45,62%	91,65%
33	Montjuic_eng.mp4	00:01:21	43,088	37,912	81,000	NOVOZ	36,888	2,093	85,61%	5,52%
						LOC01	6,200	35,819	14,39%	94,48%
34	Montjuic_esp.mp4	00:01:21	43,088	37,912	81,000	NOVOZ	39,330	2,007	91,28%	5,29%
						LOC01	3,758	35,905	8,72%	94,71%
35	PueblosEadadlMedia_esp.mp4	00:01:46	36,826	69,174	106,000	NOVOZ	0,037	17,397	0,10%	25,15%
						LOC01	36,789	51,777	99,90%	74,85%
36	RomanicoAragon_esp.mp4	00:01:34	26,353	67,647	94,000	NOVOZ	19,496	5,074	73,98%	7,50%
						LOC01	6,857	62,573	26,02%	92,50%
37	Salamanca_esp.mp4	00:05:29	181,467	147,533	329,000	NOVOZ	152,977	7,804	84,30%	5,29%
						LOC01	28,490	139,729	15,70%	94,71%
38	SantiagoCompostela_esp.mp4	00:05:31	182,655	148,345	331,000	NOVOZ	162,088	8,812	88,74%	5,94%
						LOC01	20,567	139,533	11,26%	94,06%
39	Segovia_eng.mp4	00:03:47	120,469	106,531	227,000	NOVOZ	109,223	5,849	90,67%	5,49%
						LOC01	11,246	100,682	9,34%	94,51%
40	Segovia_esp.mp4	00:03:47	120,469	106,531	227,000	NOVOZ	114,048	11,101	94,67%	10,42%
						LOC01	6,421	95,430	5,33%	89,58%
41	Toledo_eng.mp4	00:03:48	129,071	98,929	228,000	NOVOZ	115,890	12,414	89,79%	12,55%
						LOC01	13,181	86,515	10,21%	87,45%
42	Toledo_esp.mp4	00:03:48	129,071	98,929	228,000	NOVOZ	108,625	6,193	84,16%	6,26%
						LOC01	20,446	92,736	15,84%	93,74%
43	Ubeda_eng.mp4	00:03:36	108,396	107,604	216,000	NOVOZ	95,367	6,720	87,98%	6,25%
						LOC01	13,029	100,884	12,02%	93,76%
44	Ubeda_esp.mp4	00:03:36	108,396	107,604	216,000	NOVOZ	90,453	5,875	83,45%	5,46%
						LOC01	17,943	101,729	16,55%	94,54%
45	Valencia_eng.mp4	00:03:34	76,512	137,488	214,000	NOVOZ	65,616	10,490	85,76%	7,63%
						LOC01	10,896	126,998	14,24%	92,37%
46	Valencia_esp.mp4	00:03:34	76,512	137,488	214,000	NOVOZ	65,018	4,730	84,98%	3,44%
						LOC01	11,494	132,758	15,02%	96,56%

B6. Puntuaciones y resultados obtenidos tras el reconocimiento de Idioma y decisiones tomadas con el modelo de HMM de 1 Gaussiana.

MODELO 1 GAUSS						
ID	VIDEO	IDIOMA ORIGINAL	SCORE SPANISH	SCORE ENGLISH	IDIOMA RECONOCIDO	PORCENTAJE
1	Alcala_esp.mp4	SPANISH	-126,3931	-128,2880	SPANISH	100%
2	Andalucia_eng.mp4	ENGLISH	-127,1220	-127,1857	SPANISH	0%
3	Andalucia_esp.mp4	SPANISH	-127,4496	-128,1953	SPANISH	100%
4	Aranjuez_eng.mp4	ENGLISH	-125,8373	-126,9293	SPANISH	0%
5	Aranjuez_esp.mp4	SPANISH	-126,7649	-128,7211	SPANISH	100%
6	Avila_esp.mp4	SPANISH	-125,2769	-125,6968	SPANISH	100%
7	Baeza_eng.mp4	ENGLISH	-127,8295	-128,1161	SPANISH	0%
8	Baeza_esp.mp4	SPANISH	-126,8779	-126,9831	SPANISH	100%
9	Bcngotico_eng.mp4	ENGLISH	-128,2460	-127,8897	ENGLISH	100%
10	Bcngotico_esp.mp4	SPANISH	-129,3607	-128,1561	ENGLISH	0%
11	Bcnmodernista_eng.mp4	ENGLISH	-128,2929	-128,5050	SPANISH	0%
12	Bcnmodernista_esp.mp4	SPANISH	-128,7249	-128,4172	ENGLISH	0%
13	Caceres_esp.mp4	SPANISH	-126,8382	-127,4699	SPANISH	100%
14	CaminoSantiago_esp.mp4	SPANISH	-126,0589	-127,1577	SPANISH	100%
15	CastillaLeon_esp.mp4	SPANISH	-125,2989	-123,7925	ENGLISH	0%
16	Cid_esp.mp4	SPANISH	-123,1664	-121,6432	ENGLISH	0%
17	Cuenca_esp.mp4	SPANISH	-126,8512	-126,9151	SPANISH	100%
18	Dalisure_eng.mp4	ENGLISH	-125,6722	-124,8229	ENGLISH	100%
19	Dalisure_esp.mp4	SPANISH	-126,9849	-126,9097	ENGLISH	0%
20	Escorial_esp.mp4	SPANISH	-126,8929	-127,0808	SPANISH	100%
21	Extremadura_eng.mp4	ENGLISH	-126,6164	-126,2399	ENGLISH	100%
22	Extremadura_esp.mp4	SPANISH	-126,8560	-126,9716	SPANISH	100%
23	Gaudi_eng.mp4	ENGLISH	-125,8999	-126,6743	SPANISH	0%
24	Gaudi_esp.mp4	SPANISH	-126,1004	-126,8183	SPANISH	100%
25	GaudiGenio_eng.mp4	ENGLISH	-128,1352	-130,4590	SPANISH	0%
26	GaudiGenio_esp.mp4	SPANISH	-128,9343	-131,0378	SPANISH	100%
27	Ibiza_esp.mp4	SPANISH	-125,8052	-128,1030	SPANISH	100%
28	Laguna_esp.mp4	SPANISH	-124,3769	-125,8724	SPANISH	100%
29	LaMancha_eng.mp4	ENGLISH	-125,8840	-126,4603	SPANISH	0%
30	LaMancha_esp.mp4	SPANISH	-125,0277	-125,4676	SPANISH	100%
31	Madrid_eng.mp4	ENGLISH	-126,4352	-125,8015	ENGLISH	100%
32	Madrid_esp.mp4	SPANISH	-126,6005	-127,1093	SPANISH	100%
33	Montjuic_eng.mp4	ENGLISH	-130,7432	-130,7099	ENGLISH	100%
34	Montjuic_esp.mp4	SPANISH	-130,6053	-130,2718	ENGLISH	0%
35	PueblosEdadMedia_esp.mp4	SPANISH	-124,6824	-125,5928	SPANISH	100%
36	RomanicoAragon_esp.mp4	SPANISH	-126,9135	-128,3833	SPANISH	100%
37	Salamanca_esp.mp4	SPANISH	-126,9446	-126,9713	SPANISH	100%
38	SantiagoCompostela_esp.mp4	SPANISH	-125,6445	-126,9341	SPANISH	100%
39	Segovia_eng.mp4	ENGLISH	-124,8808	-123,5902	ENGLISH	100%
40	Segovia_esp.mp4	SPANISH	-123,8944	-123,6967	ENGLISH	0%
41	Toledo_eng.mp4	ENGLISH	-124,8974	-123,5120	ENGLISH	100%
42	Toledo_esp.mp4	SPANISH	-124,6275	-124,4267	ENGLISH	0%
43	Ubeda_eng.mp4	ENGLISH	-127,8635	-128,0189	SPANISH	0%
44	Ubeda_esp.mp4	SPANISH	-127,2265	-127,2456	SPANISH	100%
45	Valencia_eng.mp4	ENGLISH	-126,7275	-127,6083	SPANISH	0%
46	Valencia_esp.mp4	SPANISH	-127,2212	-128,8314	SPANISH	100%

MODELO 1 GAUSSIANA		
AUDIOS	REAL	
RECONOCIDO	SPANISH	ENGLISH
SPANISH	22	9
ENGLISH	8	7

PORCENTAJES		
RECONOCIDO	REAL	
SPANISH	SPANISH	ENGLISH
SPANISH	73,33%	56,25%
ENGLISH	26,67%	43,75%

B7. Puntuaciones y resultados obtenidos tras el reconocimiento de Idioma y decisiones tomadas con el modelo de HMM de 2 Gaussianas.

ID	VIDEO	IDIOMA ORIGINAL	MODELO 2 GAUSS			
			SCORE SPANISH	SCORE ENGLISH	IDIOMA RECONOCIDO	PORCENTAJE
1	Alcala_esp.mp4	SPANISH	-125,9483	-128,2171	SPANISH	100%
2	Andalucia_eng.mp4	ENGLISH	-126,7506	-126,9180	SPANISH	0%
3	Andalucia_esp.mp4	SPANISH	-127,0865	-127,9860	SPANISH	100%
4	Aranjuez_eng.mp4	ENGLISH	-125,1972	-126,8162	SPANISH	0%
5	Aranjuez_esp.mp4	SPANISH	-126,1580	-128,5841	SPANISH	100%
6	Avila_esp.mp4	SPANISH	-124,7394	-125,6290	SPANISH	100%
7	Baeza_eng.mp4	ENGLISH	-127,3125	-127,8085	SPANISH	0%
8	Baeza_esp.mp4	SPANISH	-126,3940	-126,6771	SPANISH	100%
9	Bcngotico_eng.mp4	ENGLISH	-127,9953	-127,7316	ENGLISH	100%
10	Bcngotico_esp.mp4	SPANISH	-129,0415	-127,8680	ENGLISH	0%
11	Bcnmodernista_eng.mp4	ENGLISH	-128,0259	-128,3480	SPANISH	0%
12	Bcnmodernista_esp.mp4	SPANISH	-128,3639	-128,1172	ENGLISH	0%
13	Caceres_esp.mp4	SPANISH	-126,5905	-127,3630	SPANISH	100%
14	CaminoSantiago_esp.mp4	SPANISH	-125,1589	-127,0072	SPANISH	100%
15	CastillaLeon_esp.mp4	SPANISH	-124,8496	-123,3141	ENGLISH	0%
16	Cid_esp.mp4	SPANISH	-122,3795	-121,0959	ENGLISH	0%
17	Cuenca_esp.mp4	SPANISH	-126,5931	-126,7566	SPANISH	100%
18	Dalisure_eng.mp4	ENGLISH	-124,9643	-124,2811	ENGLISH	100%
19	Dalisure_esp.mp4	SPANISH	-126,3341	-126,4938	SPANISH	100%
20	Escorial_esp.mp4	SPANISH	-126,3848	-126,8159	SPANISH	100%
21	Extremadura_eng.mp4	ENGLISH	-126,1125	-125,9827	ENGLISH	100%
22	Extremadura_esp.mp4	SPANISH	-126,4194	-126,7375	SPANISH	100%
23	Gaudi_eng.mp4	ENGLISH	-124,9898	-126,5605	SPANISH	0%
24	Gaudi_esp.mp4	SPANISH	-125,1479	-126,7578	SPANISH	100%
25	GaudiGenio_eng.mp4	ENGLISH	-127,5656	-130,4628	SPANISH	0%
26	GaudiGenio_esp.mp4	SPANISH	-128,3516	-131,0386	SPANISH	100%
27	Ibiza_esp.mp4	SPANISH	-125,2004	-128,1102	SPANISH	100%
28	Laguna_esp.mp4	SPANISH	-123,7271	-125,8414	SPANISH	100%
29	LaMancha_eng.mp4	ENGLISH	-125,3516	-126,3262	SPANISH	0%
30	LaMancha_esp.mp4	SPANISH	-124,1172	-125,3729	SPANISH	100%
31	Madrid_eng.mp4	ENGLISH	-125,7748	-125,3920	ENGLISH	100%
32	Madrid_esp.mp4	SPANISH	-126,0097	-126,8093	SPANISH	100%
33	Montjuic_eng.mp4	ENGLISH	-130,5982	-130,5991	SPANISH	0%
34	Montjuic_esp.mp4	SPANISH	-130,3403	-130,0882	ENGLISH	0%
35	PueblosEdadMedia_esp.mp4	SPANISH	-124,0935	-125,3736	SPANISH	100%
36	RomanicoAragon_esp.mp4	SPANISH	-126,4116	-128,2313	SPANISH	100%
37	Salamanca_esp.mp4	SPANISH	-126,3463	-126,6725	SPANISH	100%
38	SantiagoCompostela_esp.mp4	SPANISH	-125,0388	-126,9355	SPANISH	100%
39	Segovia_eng.mp4	ENGLISH	-124,3094	-123,2034	ENGLISH	100%
40	Segovia_esp.mp4	SPANISH	-123,2578	-123,4705	SPANISH	100%
41	Toledo_eng.mp4	ENGLISH	-124,3999	-123,2395	ENGLISH	100%
42	Toledo_esp.mp4	SPANISH	-124,0081	-124,2267	SPANISH	100%
43	Ubeda_eng.mp4	ENGLISH	-127,5223	-127,8014	SPANISH	0%
44	Ubeda_esp.mp4	SPANISH	-126,8974	-127,0213	SPANISH	100%
45	Valencia_eng.mp4	ENGLISH	-126,2361	-127,4479	SPANISH	0%
46	Valencia_esp.mp4	SPANISH	-126,6911	-128,7456	SPANISH	100%

MODELO 2 GAUSSIANAS		
AUDIOS	REAL	
RECONOCIDO	SPANISH	ENGLISH
SPANISH	25	10
ENGLISH	5	6

PORCENTAJES	REAL	
RECONOCIDO	SPANISH	ENGLISH
SPANISH	83.33%	62.50%
ENGLISH	16.67%	37.50%

B8. Puntuaciones y resultados obtenidos tras el reconocimiento de Idioma y decisiones tomadas con el modelo de HMM de 3 Gaussianas.

ID	VIDEO	IDIOMA ORIGINAL	MODELO 3 GAUSS			
			SCORE SPANISH	SCORE ENGLISH	IDIOMA RECONOCIDO	PORCENTAJE
1	Alcala_esp.mp4	SPANISH	-125,6671	-128,2574	SPANISH	100%
2	Andalucia_eng.mp4	ENGLISH	-126,5642	-126,8992	SPANISH	0%
3	Andalucia_esp.mp4	SPANISH	-126,8953	-126,8992	SPANISH	100%
4	Aranjuez_eng.mp4	ENGLISH	-124,9405	-126,8142	SPANISH	0%
5	Aranjuez_esp.mp4	SPANISH	-125,8595	-128,8522	SPANISH	100%
6	Avila_esp.mp4	SPANISH	-124,4902	-125,6194	SPANISH	100%
7	Baeza_eng.mp4	ENGLISH	-127,0504	-127,7819	SPANISH	0%
8	Baeza_esp.mp4	SPANISH	-126,0808	-126,6330	SPANISH	100%
9	Bcngotico_eng.mp4	ENGLISH	-127,8750	-127,7680	ENGLISH	100%
10	Bcngotico_esp.mp4	SPANISH	-128,7572	-127,9295	ENGLISH	0%
11	Bcnmodernista_eng.mp4	ENGLISH	-127,8676	-128,4079	SPANISH	0%
12	Bcnmodernista_esp.mp4	SPANISH	-128,1460	-128,0901	ENGLISH	0%
13	Caceres_esp.mp4	SPANISH	-126,5116	-127,4771	SPANISH	100%
14	CaminoSantiago_esp.mp4	SPANISH	-124,6395	-126,9627	SPANISH	100%
15	CastillaLeon_esp.mp4	SPANISH	-124,6322	-123,1809	ENGLISH	0%
16	Cid_esp.mp4	SPANISH	-122,0465	-120,9757	ENGLISH	0%
17	Cuenca_esp.mp4	SPANISH	-126,4360	-126,8390	SPANISH	100%
18	Dalisure_eng.mp4	ENGLISH	-124,7285	-124,1534	ENGLISH	100%
19	Dalisure_esp.mp4	SPANISH	-126,0800	-126,4171	SPANISH	100%
20	Escorial_esp.mp4	SPANISH	-126,1187	-126,8402	SPANISH	100%
21	Extremadura_eng.mp4	ENGLISH	-127,8708	-125,9869	ENGLISH	100%
22	Extremadura_esp.mp4	SPANISH	-126,1878	-126,7472	SPANISH	100%
23	Gaudi_eng.mp4	ENGLISH	-124,5839	-126,5408	SPANISH	0%
24	Gaudi_esp.mp4	SPANISH	-124,6614	-126,7549	SPANISH	100%
25	GaudiGenio_eng.mp4	ENGLISH	-127,3471	-130,5416	SPANISH	0%
26	GaudiGenio_esp.mp4	SPANISH	-128,0131	-131,1670	SPANISH	100%
27	Ibiza_esp.mp4	SPANISH	-124,8714	-128,1798	SPANISH	100%
28	Laguna_esp.mp4	SPANISH	-123,3606	-125,8604	SPANISH	100%
29	LaMancha_eng.mp4	ENGLISH	-125,1273	-126,4118	SPANISH	0%
30	LaMancha_esp.mp4	SPANISH	-123,6857	-125,3785	SPANISH	100%
31	Madrid_eng.mp4	ENGLISH	-125,6095	-125,2874	ENGLISH	100%
32	Madrid_esp.mp4	SPANISH	-125,7935	-126,7599	SPANISH	100%
33	Montjuic_eng.mp4	ENGLISH	-130,4568	-130,7234	SPANISH	0%
34	Montjuic_esp.mp4	SPANISH	-130,0311	-130,1427	SPANISH	100%
35	PueblosEdadMedia_esp.mp4	SPANISH	-123,7853	-125,2007	SPANISH	100%
36	RomanicoAragon_esp.mp4	SPANISH	-126,1887	-128,3152	SPANISH	100%
37	Salamanca_esp.mp4	SPANISH	-125,9704	-126,6650	SPANISH	100%
38	SantiagoCompostela_esp.mp4	SPANISH	-124,7393	-126,9892	SPANISH	100%
39	Segovia_eng.mp4	ENGLISH	-124,1154	-123,1767	ENGLISH	100%
40	Segovia_esp.mp4	SPANISH	-122,9574	-123,4206	SPANISH	100%
41	Toledo_eng.mp4	ENGLISH	-124,1984	-123,3796	ENGLISH	100%
42	Toledo_esp.mp4	SPANISH	-123,7130	-124,2255	SPANISH	100%
43	Ubeda_eng.mp4	ENGLISH	-127,4133	-127,8980	SPANISH	0%
44	Ubeda_esp.mp4	SPANISH	-126,7225	-127,0757	SPANISH	100%
45	Valencia_eng.mp4	ENGLISH	-125,9774	-127,4578	SPANISH	0%
46	Valencia_esp.mp4	SPANISH	-126,3266	-128,8042	SPANISH	100%

MODELO 3 GAUSSIANAS		
AUDIOS	REAL	
RECONOCIDO	SPANISH	ENGLISH
SPANISH	26	10
ENGLISH	4	6

PORCENTAJES		
RECONOCIDO	REAL	
SPANISH	SPANISH	ENGLISH
SPANISH	86,67%	62,50%
ENGLISH	13,33%	37,50%

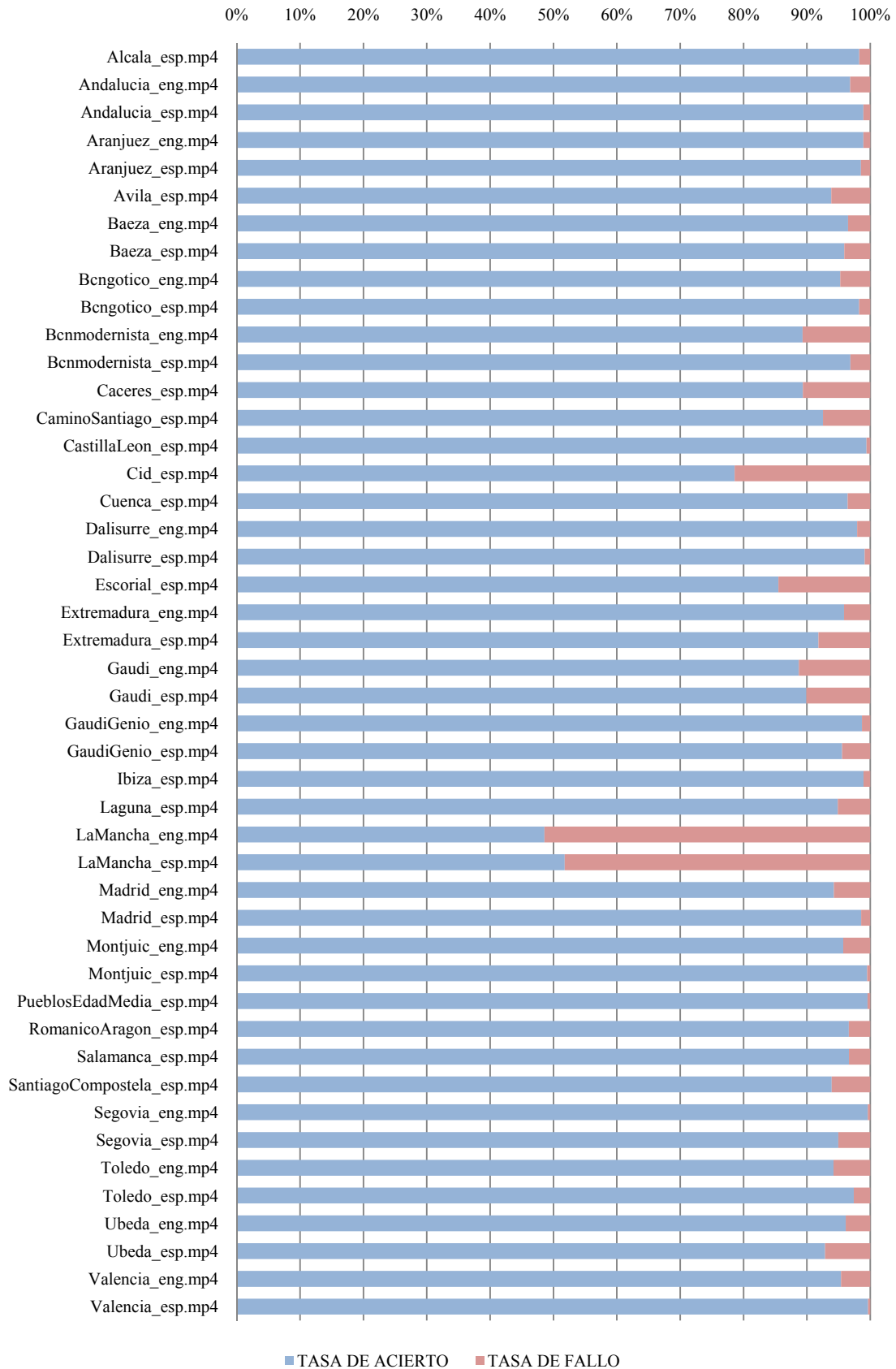
B9. Puntuaciones y resultados obtenidos tras el reconocimiento de Idioma y decisiones tomadas con el modelo de HMM de 4 Gaussianas.

MODELO 4 GAUSS						
ID	VIDEO	IDIOMA ORIGINAL	SCORE SPANISH	SCORE ENGLISH	IDIOMA RECONOCIDO	PORCENTAJE
1	Alcala_esp.mp4	SPANISH	-125,6160	-128,4250	SPANISH	100%
2	Andalucia_eng.mp4	ENGLISH	-126,5611	-127,0301	SPANISH	0%
3	Andalucia_esp.mp4	SPANISH	-126,9100	-128,1523	SPANISH	100%
4	Aranjuez_eng.mp4	ENGLISH	-124,9153	-127,0067	SPANISH	0%
5	Aranjuez_esp.mp4	SPANISH	-125,7864	-128,7700	SPANISH	100%
6	Avila_esp.mp4	SPANISH	-124,4904	-125,8153	SPANISH	100%
7	Baeza_eng.mp4	ENGLISH	-127,0113	-127,9998	SPANISH	0%
8	Baeza_esp.mp4	SPANISH	-126,0271	-126,8517	SPANISH	100%
9	Bcngotico_eng.mp4	ENGLISH	-127,8460	-128,0178	SPANISH	0%
10	Bcngotico_esp.mp4	SPANISH	-128,6674	-128,0736	ENGLISH	0%
11	Bcnmodernista_eng.mp4	ENGLISH	-127,8640	-128,5445	SPANISH	0%
12	Bcnmodernista_esp.mp4	SPANISH	-128,0640	-128,1830	SPANISH	100%
13	Caceres_esp.mp4	SPANISH	-126,5592	-127,6783	SPANISH	100%
14	CaminoSantiago_esp.mp4	SPANISH	-124,5375	-127,0939	SPANISH	100%
15	CastillaLeon_esp.mp4	SPANISH	-124,8535	-123,3616	ENGLISH	0%
16	Cid_esp.mp4	SPANISH	-122,1226	-121,2085	ENGLISH	0%
17	Cuenca_esp.mp4	SPANISH	-126,4665	-127,0010	SPANISH	100%
18	Dalisure_eng.mp4	ENGLISH	-124,8599	-124,2744	ENGLISH	100%
19	Dalisure_esp.mp4	SPANISH	-126,1403	-126,5171	SPANISH	100%
20	Escorial_esp.mp4	SPANISH	-126,1535	-126,9685	SPANISH	100%
21	Extremadura_eng.mp4	ENGLISH	-125,9138	-126,1689	SPANISH	0%
22	Extremadura_esp.mp4	SPANISH	-126,1829	-126,9022	SPANISH	100%
23	Gaudi_eng.mp4	ENGLISH	-124,4964	-126,7647	SPANISH	0%
24	Gaudi_esp.mp4	SPANISH	-124,5675	-127,0014	SPANISH	100%
25	GaudiGenio_eng.mp4	ENGLISH	-127,2955	-130,8167	SPANISH	0%
26	GaudiGenio_esp.mp4	SPANISH	-127,8886	-131,3857	SPANISH	100%
27	Ibiza_esp.mp4	SPANISH	-124,8204	-128,4097	SPANISH	100%
28	Laguna_esp.mp4	SPANISH	-123,3161	-126,0667	SPANISH	100%
29	LaMancha_eng.mp4	ENGLISH	-125,1573	-126,6836	SPANISH	0%
30	LaMancha_esp.mp4	SPANISH	-123,6350	-125,6502	SPANISH	100%
31	Madrid_eng.mp4	ENGLISH	-125,7301	-125,4841	ENGLISH	100%
32	Madrid_esp.mp4	SPANISH	-125,7930	-126,9111	SPANISH	100%
33	Montjuic_eng.mp4	ENGLISH	-130,4098	-130,8037	SPANISH	0%
34	Montjuic_esp.mp4	SPANISH	-129,9389	-130,2308	SPANISH	100%
35	PueblosEdadMedia_esp.mp4	SPANISH	-123,9043	-125,3628	SPANISH	100%
36	RomanicoAragon_esp.mp4	SPANISH	-126,2075	-128,5440	SPANISH	100%
37	Salamanca_esp.mp4	SPANISH	-125,9396	-126,8530	SPANISH	100%
38	SantiagoCompostela_esp.mp4	SPANISH	-124,7026	-127,2017	SPANISH	100%
39	Segovia_eng.mp4	ENGLISH	-124,2132	-123,4632	ENGLISH	100%
40	Segovia_esp.mp4	SPANISH	-122,9906	-123,6943	SPANISH	100%
41	Toledo_eng.mp4	ENGLISH	-124,2846	-123,5893	ENGLISH	100%
42	Toledo_esp.mp4	SPANISH	-123,7415	-124,4492	SPANISH	100%
43	Ubeda_eng.mp4	ENGLISH	-127,4136	-128,0646	SPANISH	0%
44	Ubeda_esp.mp4	SPANISH	-126,6761	-127,2394	SPANISH	100%
45	Valencia_eng.mp4	ENGLISH	-125,9614	-127,6283	SPANISH	0%
46	Valencia_esp.mp4	SPANISH	-126,2225	-128,9580	SPANISH	100%

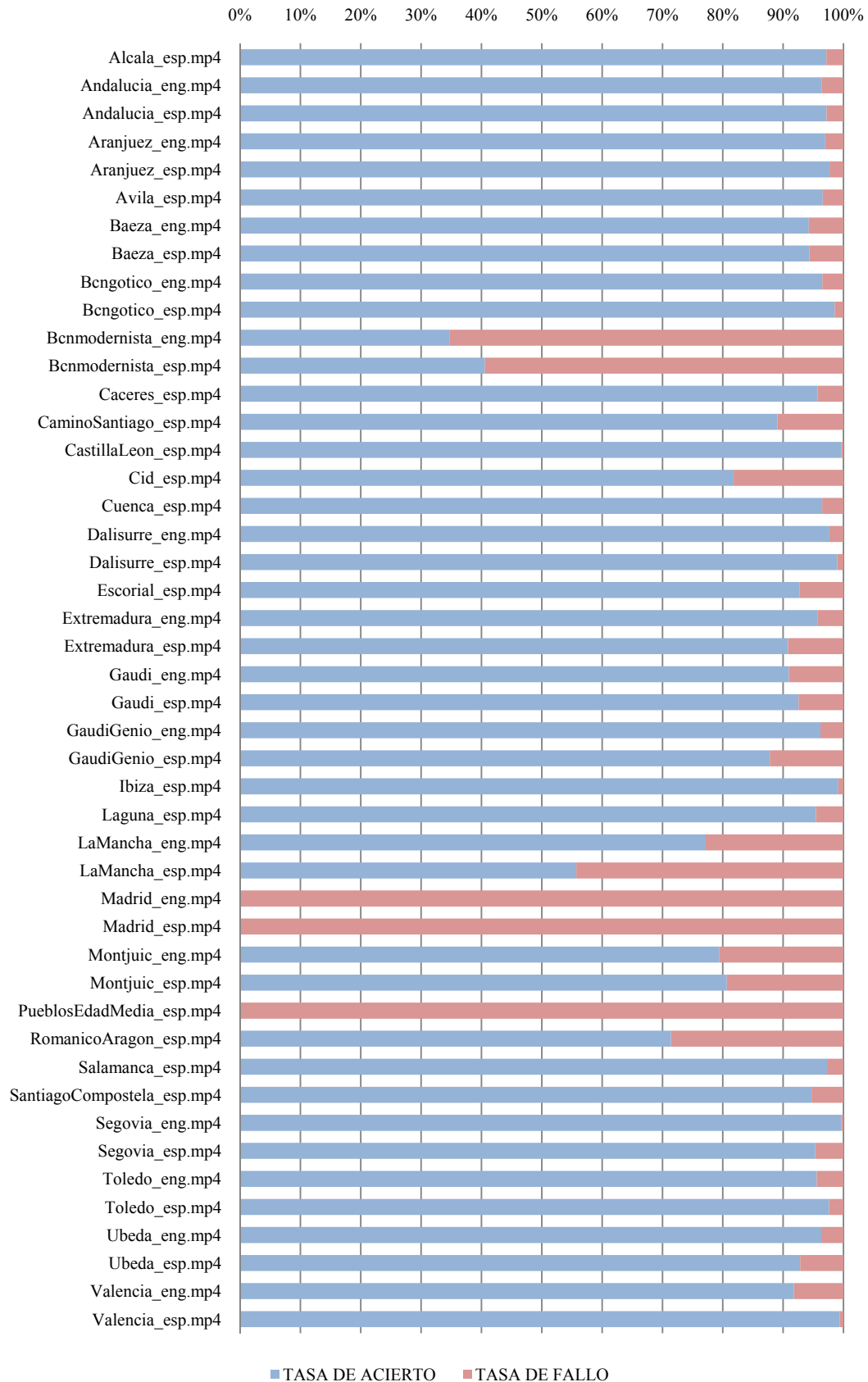
MODELO 4 GAUSSIANAS		
AUDIOS	REAL	
RECONOCIDO	SPANISH	ENGLISH
SPANISH	27	12
ENGLISH	3	4

PORCENTAJES		
RECONOCIDO	REAL	
SPANISH	SPANISH	ENGLISH
SPANISH	90.00%	75.00%
ENGLISH	10.00%	25.00%

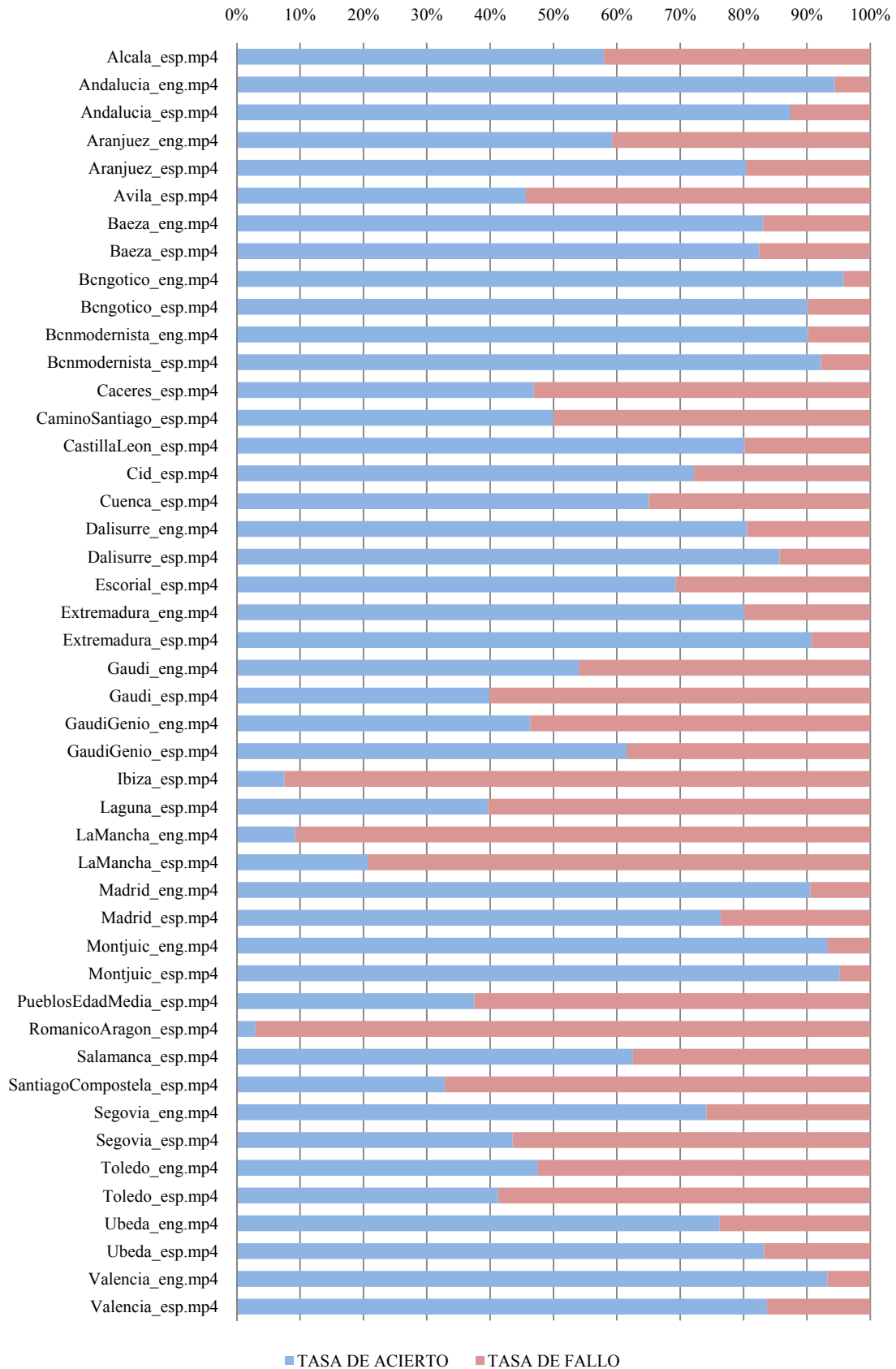
B10. Diagramas de barras de las tasas de acierto y fallo, para segmentos de VOZ en el reconocimiento con el sistema ATVS-UAM.



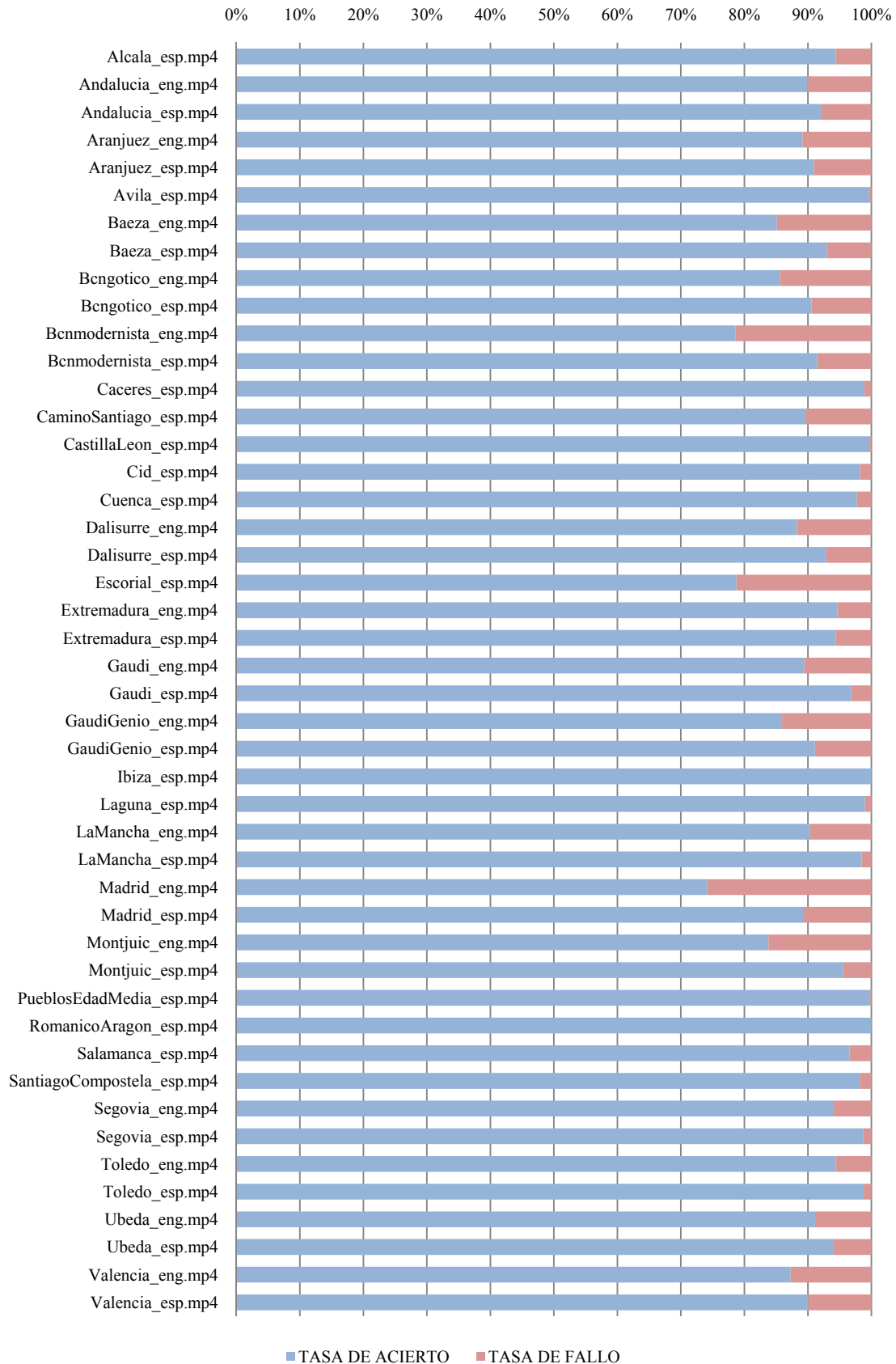
B11. Diagramas de barras de las tasas de acierto y fallo, para segmentos de NOVOZ en el reconocimiento con el sistema ATVS-UAM.



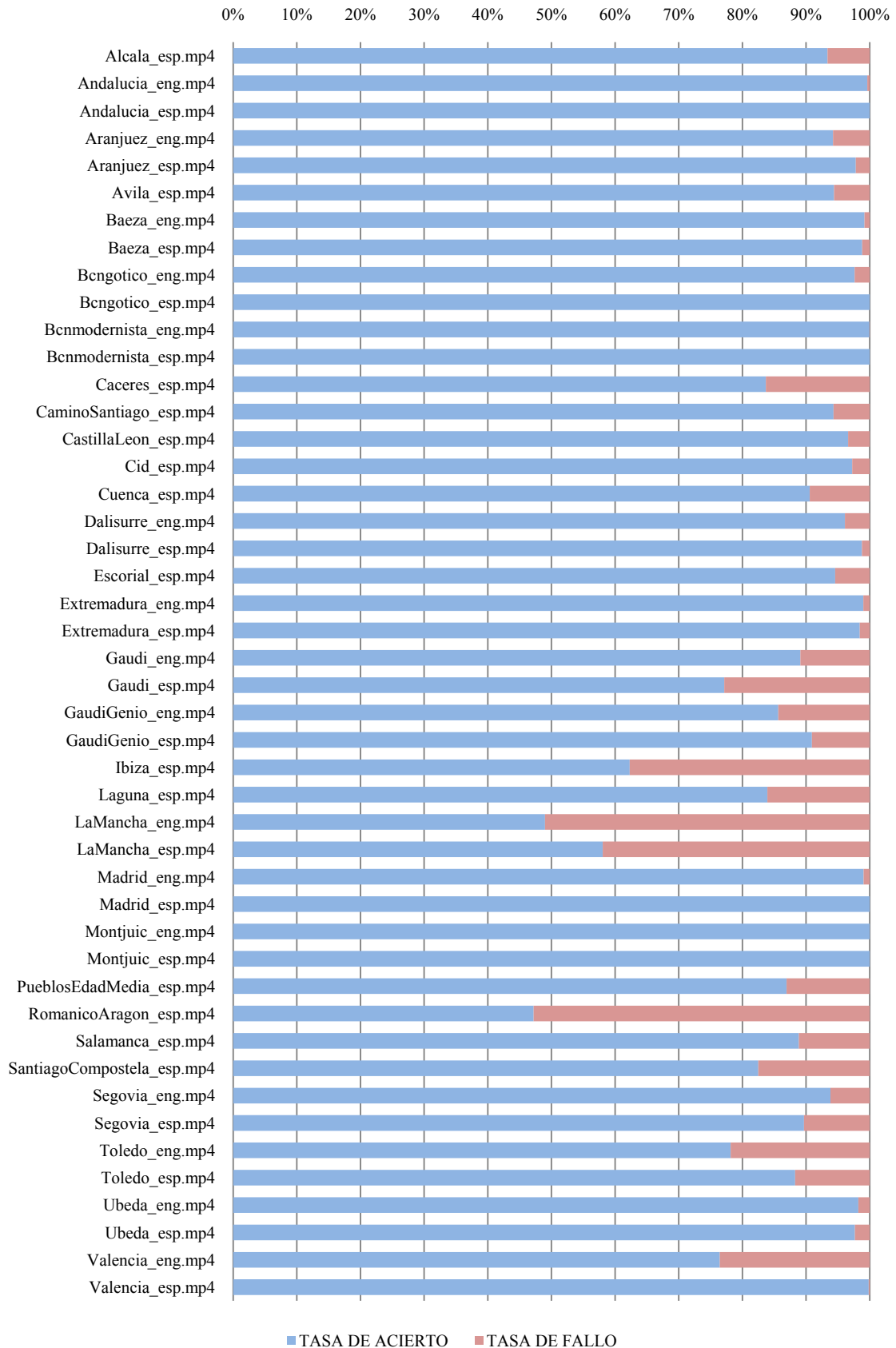
B12. Diagramas de barras de las tasas de acierto y fallo, para segmentos de VOZ en reconocimiento con Reconocedor fonético.



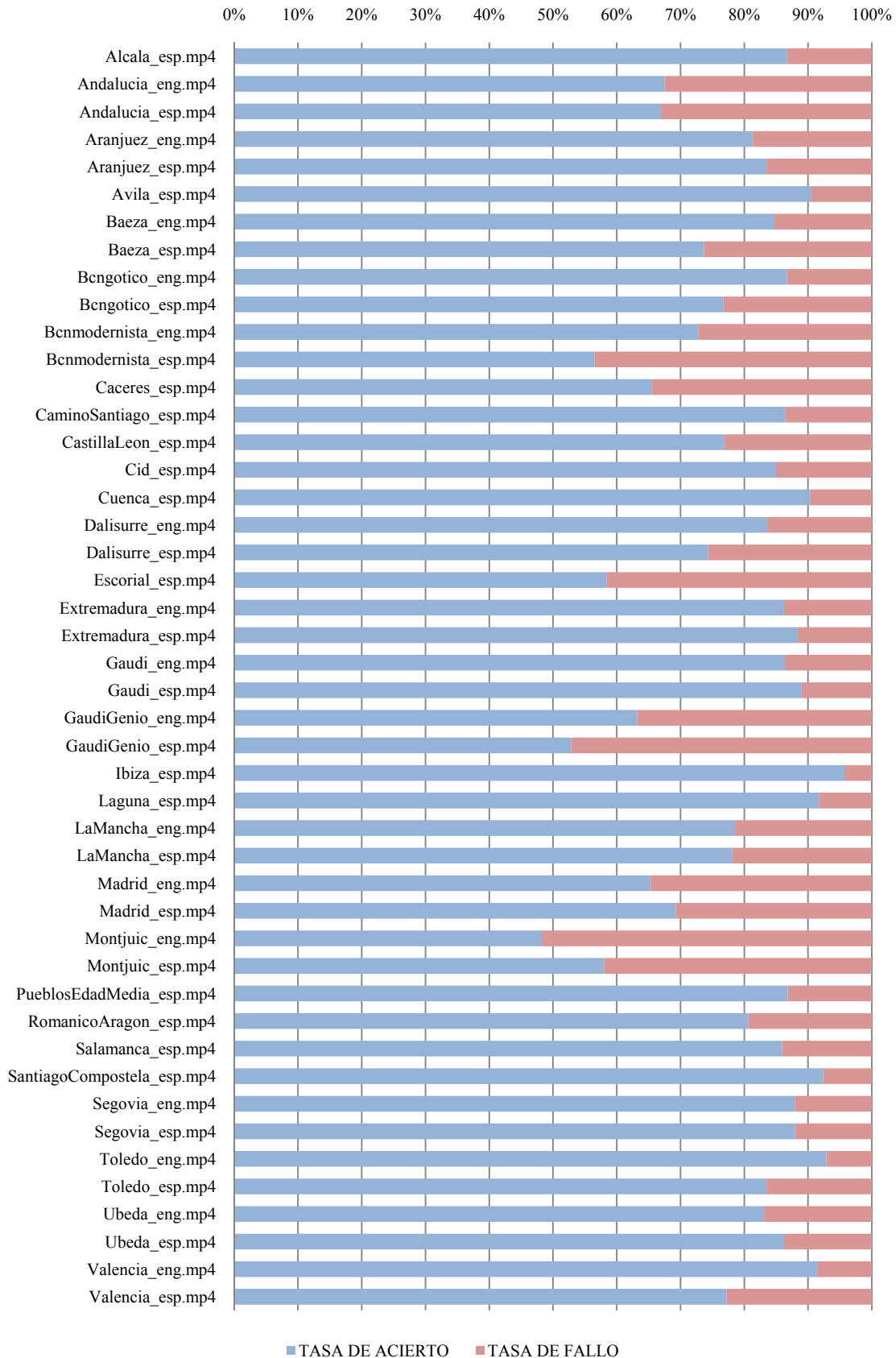
B13. Diagramas de barras de las tasas de acierto y fallo, para segmentos de NOVOZ en reconocimiento con Reconocedor fonético.



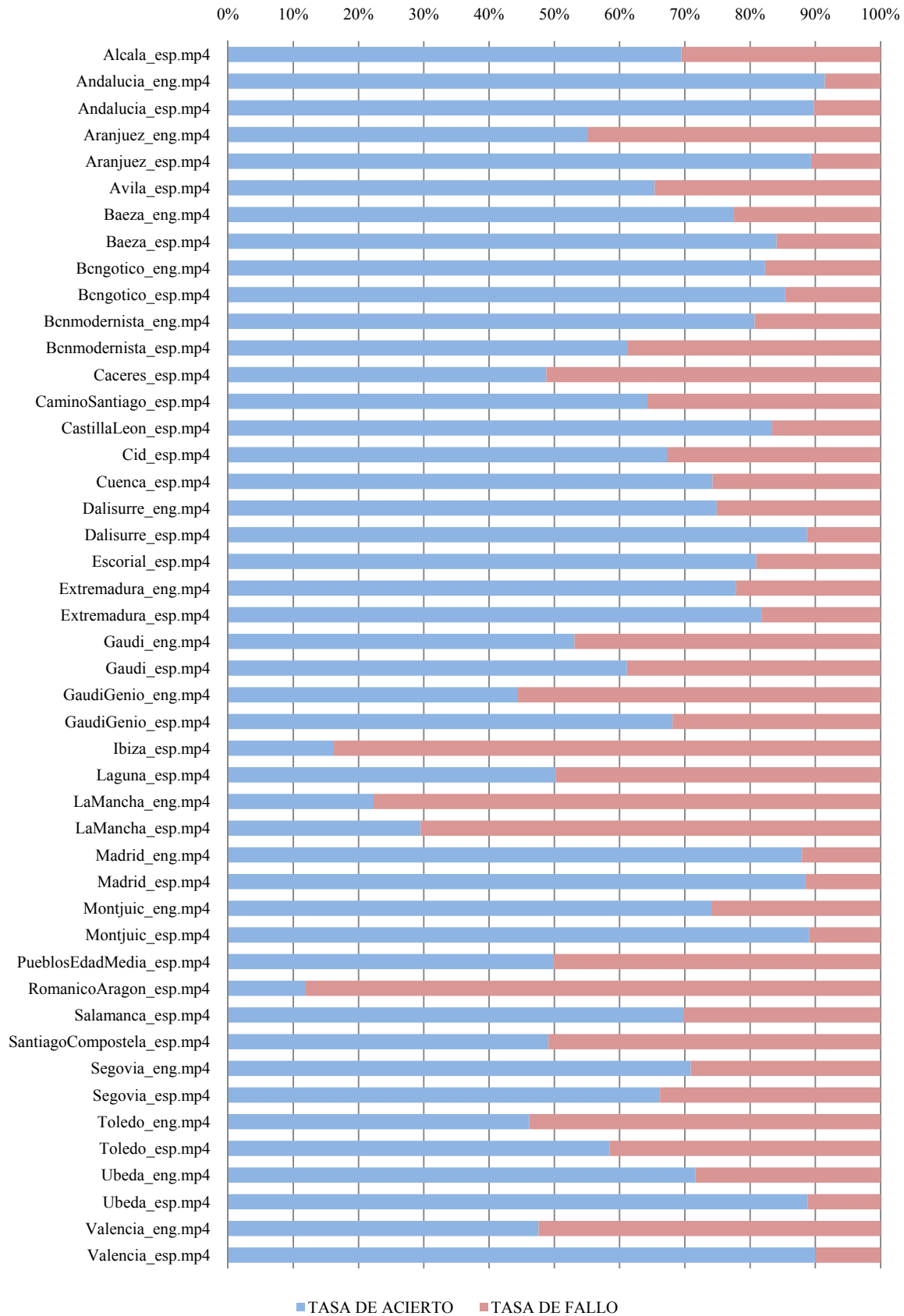
B14. Diagramas de barras de las tasas de acierto y fallo, para segmentos de VOZ en reconocimiento con Rec. fonético con adaptación al locutor.



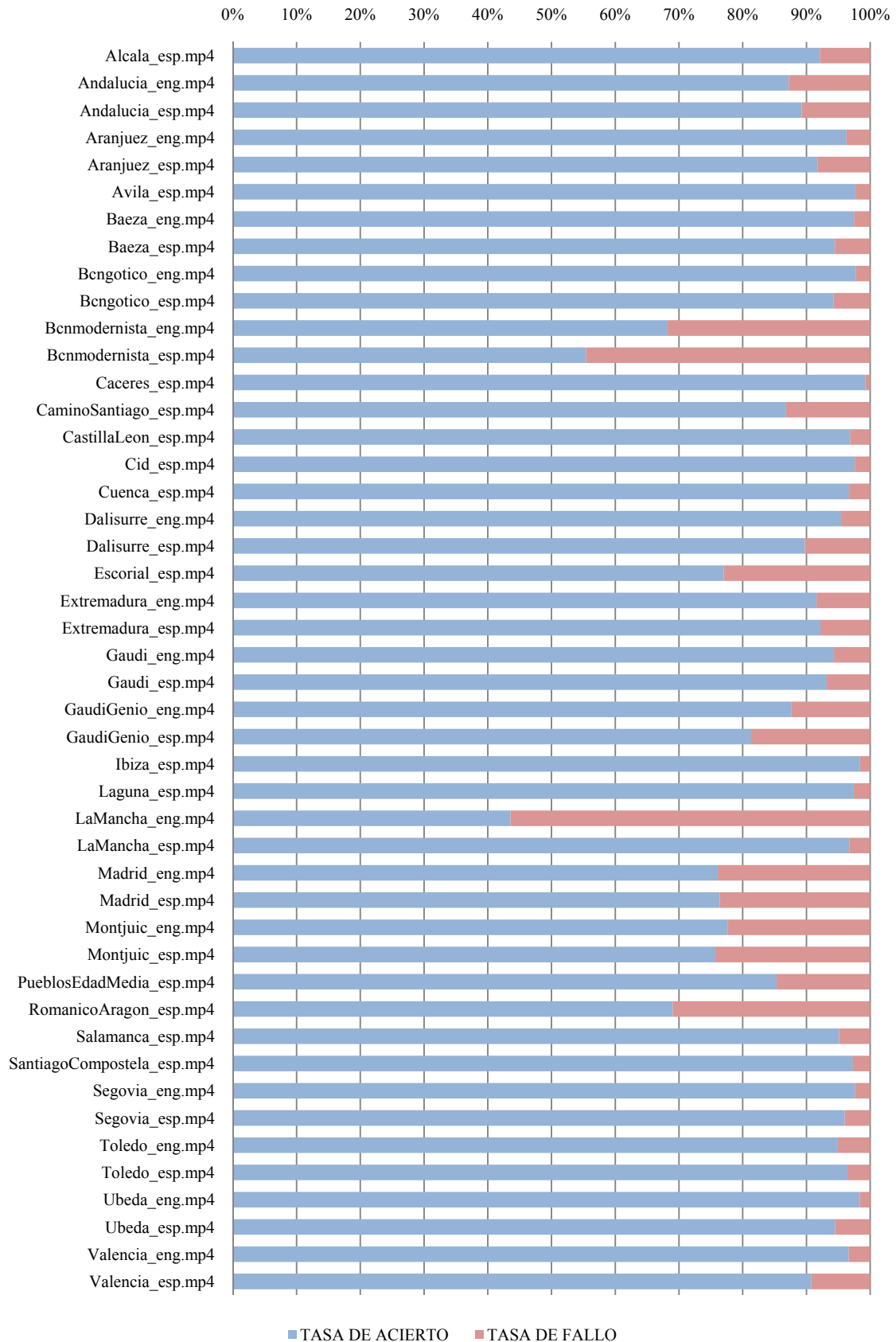
B15. Diagramas de barras de las tasas de acierto y fallo, para segmentos de NOVOZ en reconocimiento con Rec. fonético con adaptación al locutor.



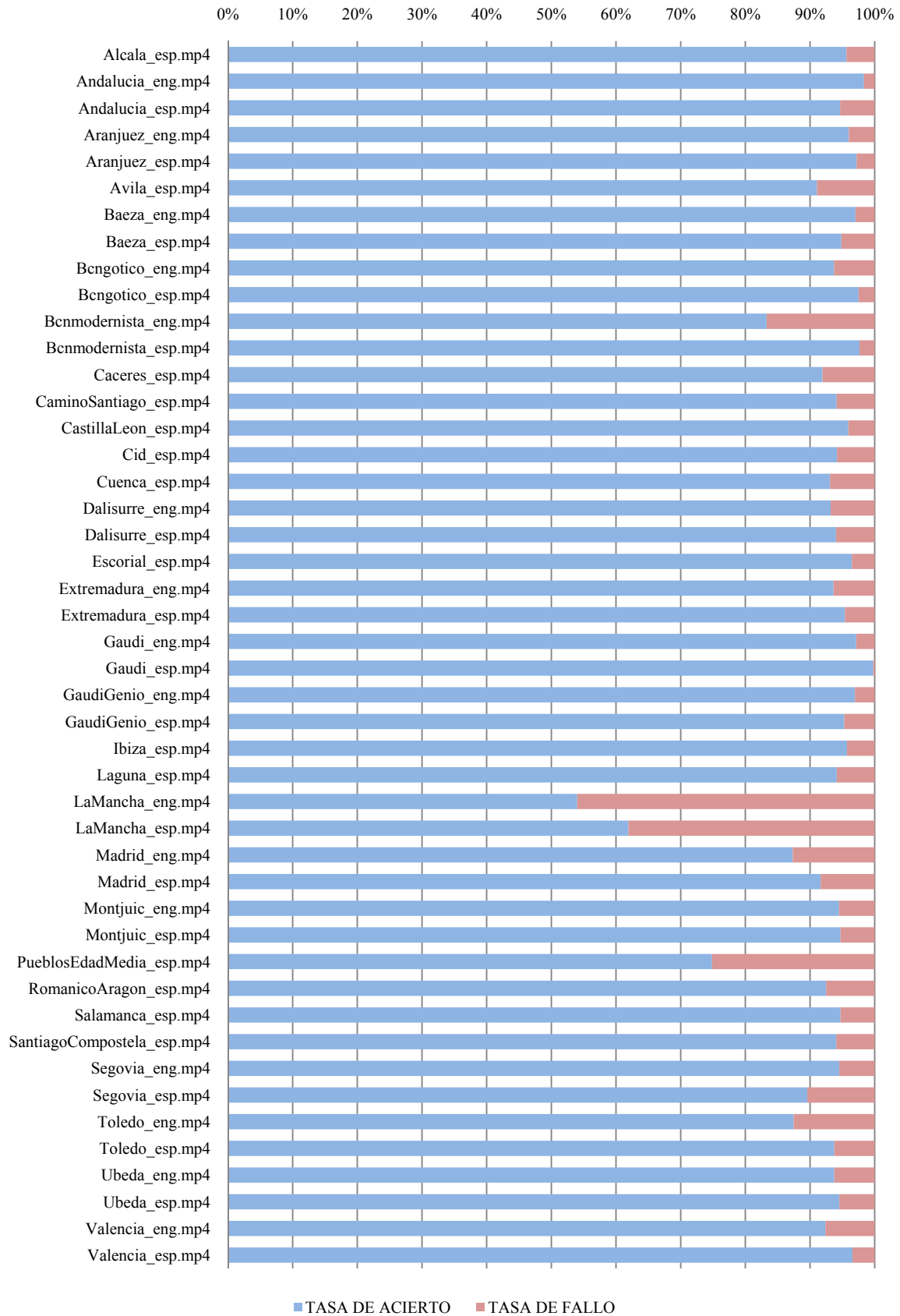
B16. Diagramas de barras de las tasas de acierto y fallo, para segmentos de VOZ en combinación de sistemas con función lógica AND.



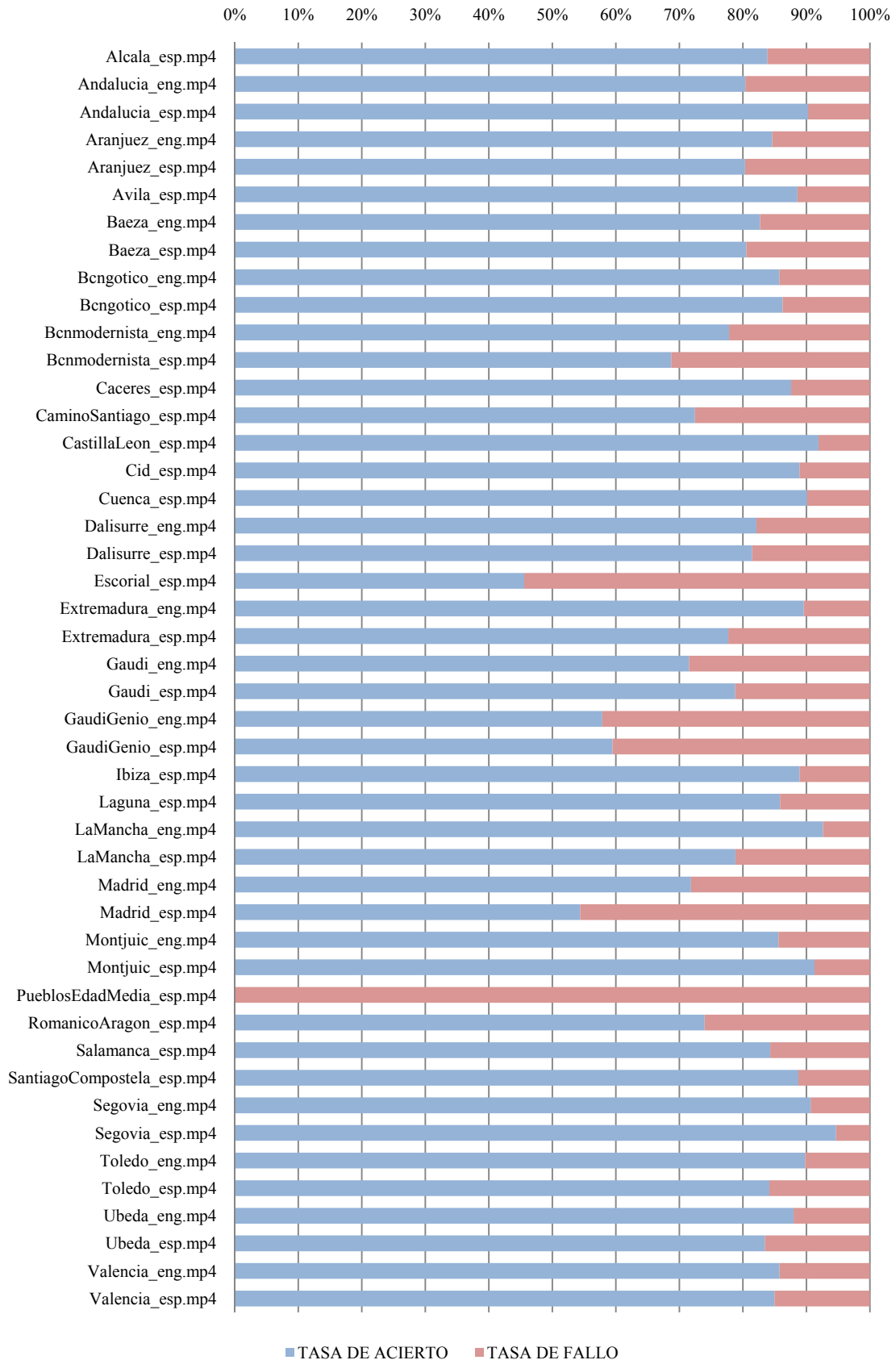
B17. Diagramas de barras de las tasas de acierto y fallo, para segmentos de NOVOZ en combinación de sistemas con función lógica AND.



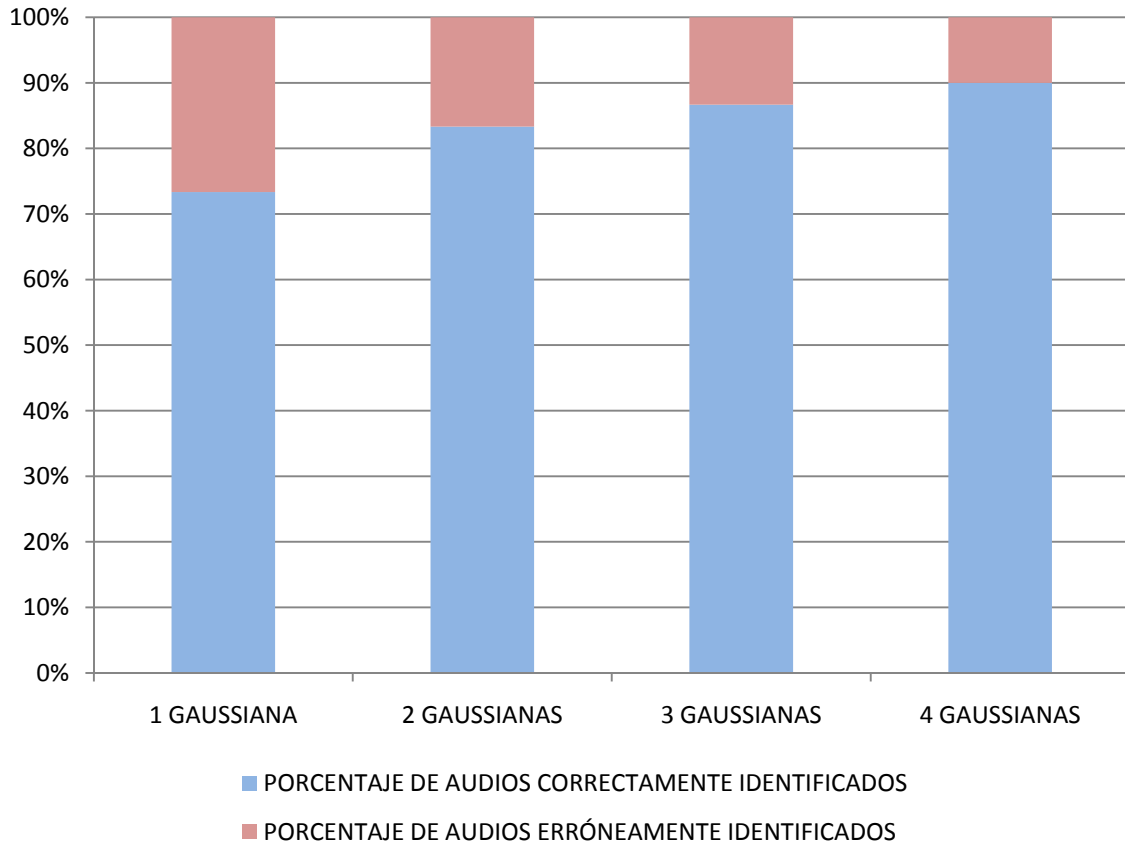
B18. Diagramas de barras de las tasas de acierto y fallo, para segmentos de VOZ en combinación de sistemas con función lógica OR.



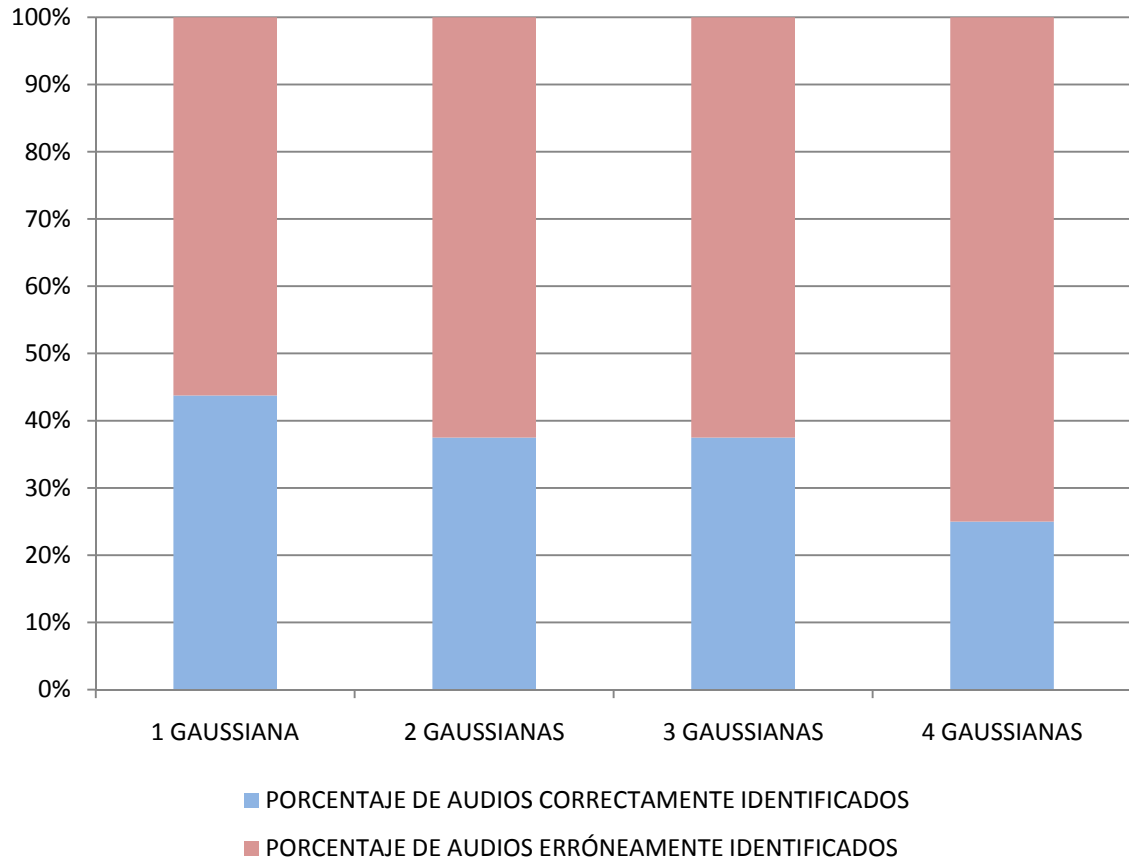
B19. Diagramas de barras de las tasas de acierto y fallo, para segmentos de NOVOZ en combinación de sistemas con función lógica OR.



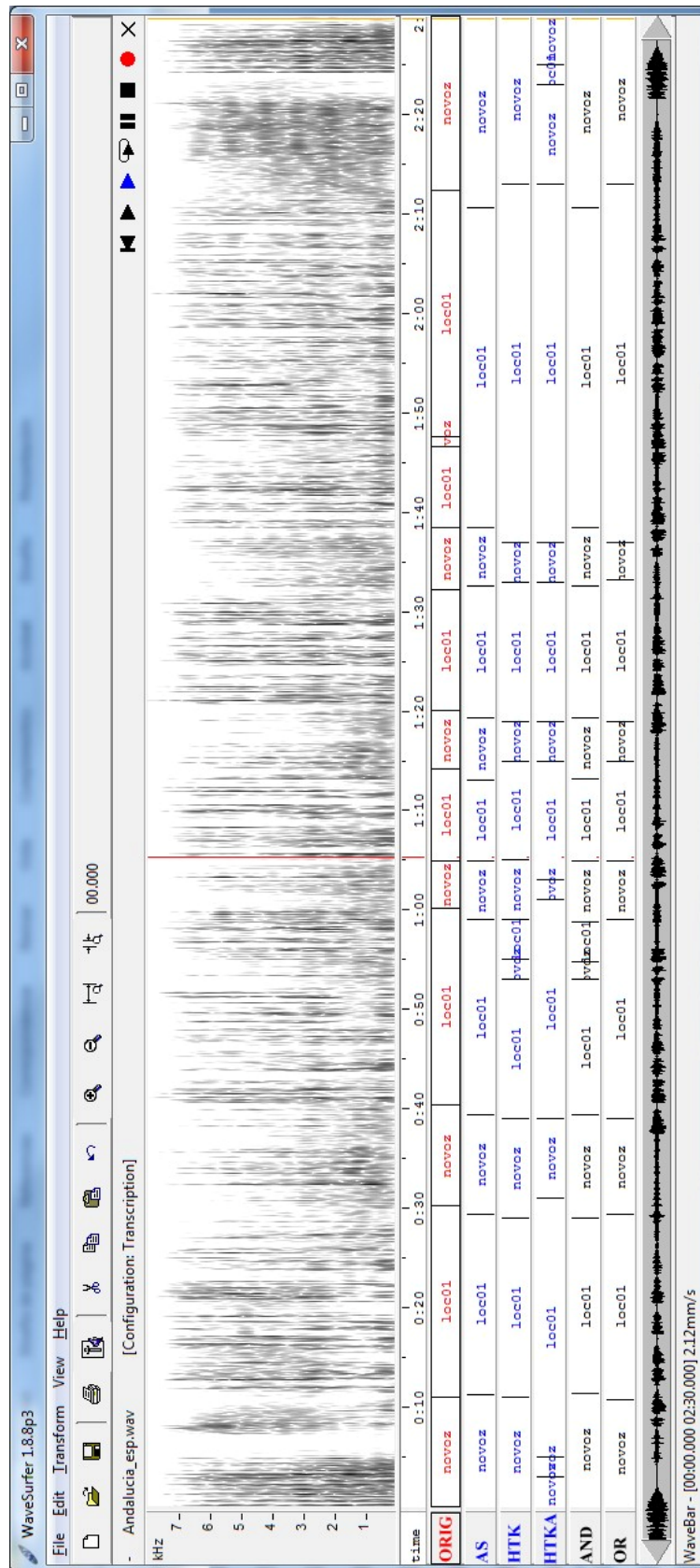
B20. Diagramas de barras de las tasas de acierto y fallo, para identificaciones de idioma de audios en ESPAÑOL.



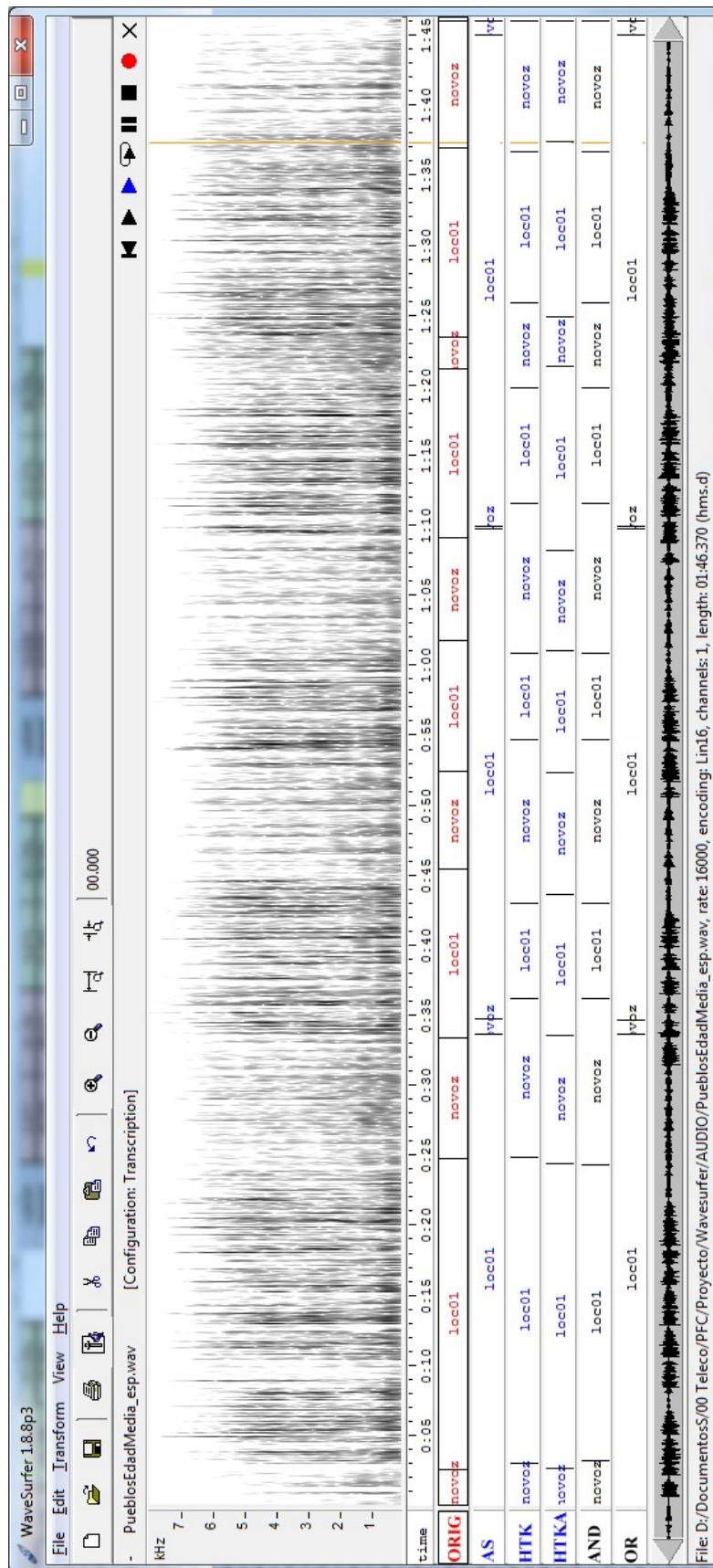
B21. Diagramas de barras de las tasas de acierto y fallo, para identificaciones de idioma de audios en INGLÉS.



B22. Representación espectral, temporal y etiquetado de segmentos del audio 'Andalucia_esp.wav' con software wavesurfer.



B23. Representación espectral, temporal y etiquetado de segmentos del audio 'PueblosEdadMedia_esp.wav' con software wavesurfer.



C

Anexo C: Presupuesto

PRESUPUESTO

1. Ejecución del Material

- Compra de ordenador personal (software incluido) 1000 €
 - Alquiler de impresora láser durante 12 meses 100 €
 - Material de oficina 200 €
- Total de ejecución material **1300 €**

2. Gastos Generales

- sobre los gastos de ejecución Material **208 €**

3. Beneficio Industrial

- sobre los gastos de ejecución Material **78 €**

4. Honorarios

- 860 horas a 10 € /hora **8600 €**

5. Material fungible

- Gastos de impresión 150 €
 - Encuadernación 100€
- Total material fungible **250 €**

6. Subtotal del presupuesto

- Subtotal del presupuesto **10.186 €**

7. IVA Aplicable

- 18% Subtotal del presupuesto **1.833,50 €**

8. Total del presupuesto

- Total del presupuesto **12.019,50 €**

Madrid, Septiembre de 2011
El Ingeniero Jefe de Proyecto

Fdo.: José Antonio Morejón Saravia
Ingeniero Superior de Telecomunicación

D

Anexo D: Pliego de Condiciones

PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un sistema de segmentación de audio e identificación de locutores para recuperación de información multimedia en videos de información turística. 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 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 declinar 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.

12. 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.

