UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

PROYECTO FIN DE CARRERA

RECONOCIMIENTO DE PALABRAS CLAVE EN CONVERSACIONES ESPONTÁNEAS EN CASTELLANO

Verónica Peña García

OCTUBRE 2008

PROYECTO FIN DE CARRERA

Título: *Reconocimiento de palabras clave en conversaciones espontáneas en castellano*

Autor: Dª. Verónica Peña García
Tutor: D. Doroteo Torre Toledano

Tribunal:

Presidente: D. Joaquín González Rodríguez

Vocal: D. José M. Martínez Sánchez

Vocal secretario: D. Doroteo Torre Toledano

Fecha de lectura: **13-10-08**

Calificación:

RECONOCIMIENTO DE PALABRAS CLAVE EN CONVERSACIONES ESPONTÁNEAS EN CASTELLANO

AUTOR: Verónica Peña García
TUTOR: Doroteo Torre Toledano

Área de Tratamiento de Voz y Señales - ATVS
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
OCTUBRE 2008

Resumen

En este proyecto se estudia, implementa y evalúa un sistema de reconocimiento de palabras clave en conversaciones espontáneas en Castellano. Como base de datos para la experimentación se emplea el corpus C-ORAL-ROM, desarrollado por el laboratorio de Lingüística Informática de la UAM.

Tras una introducción al sistema de producción de voz humano, y el estado del arte en reconocimiento de voz, se efectúa un estudio exhaustivo de los distintos enfoques de reconocimiento y los sistemas existentes en la actualidad.

La técnica usada para llevar a cabo la implementación del sistema reconocedor de palabras clave, será la búsqueda basada en lattices fonema, con ésta se desarrollará la tarea de búsqueda de términos de voz.

Para la parte experimental se ha realizado una selección de la base de datos original, con el fin de adaptarla a las pruebas y experimentos a realizar. Se evalúa el sistema para diferentes tipos de habla espontánea analizando las diferencias entre ellos. Finalmente, se presentan las conclusiones y se proponen líneas de trabajo futuras.

Palabras clave:
Wordspotting, Lattice, detector de palabras clave, C-ORAL-ROM, Reconocedor automático de voz

Abstract

In the present project, a spoken term detection system has been studied, implemented and evaluated. The database used in the experimental part is the corpus C-ORAL-ROM developed by the UAM Linguistic Informatics laboratory.

After a brief introduction to the human voice and speech recognition, we do an exhaustive study of the different approaches and current recognition systems.

We examine the task of spoken term detection in Spanish spontaneous speech with a lattice-based approach. We use lattices generated with phonemes to develop the task and we discuss methods for lattice post-processing and system combination.

In the experimental part, it is evaluated the performance of the implemented spoken term detection system using a selection from the C-ORAL-ROM database in order to analyze the system behaviour.

We evaluate the system for different types of spontaneous speech, we make a comparison between these types in order to determine which one achieves better results. Finally, conclusions are drawn, and future lines of work are proposed.

Keywords:
Wordspotting, Lattice, Spoken Term Detection, C-ORAL-ROM, Automatic Speech Recognition

Agradecimientos

No es fácil tarea agradecer todo a tantos sin olvidarse de ninguno... En primer lugar agradezco a mi tutor Doroteo Torre Toledano por concederme la posibilidad de formar parte del grupo de investigación ATVS, por toda su ayuda prestada y el apoyo profesional transmitido.

Estos agradecimientos nunca habrían sido escritos sin la formación ofrecida por la Escuela Politécnica Superior durante estos años, agradezco a todo el personal docente, y en especial a Jesús Bescós porque siempre mantuvo una puerta abierta por la que entraron dudas, consultas, problemas… y hoy, agradecimientos.

Agradezco a todos mis compañeros de laboratorio por hacer de esas cuatro paredes un lugar al que llegar con una sonrisa dibujada, gracias a cada uno de vosotros. A Dani Ramos por tener una palabra perfecta para cada momento, a Abejón por sus múltiples visitas a mi puesto brindándome toda su ayuda, a Ismael porque aún escondido siempre estuvo ahí, a Dani Hernández por compartir aquellas "puestas de sol" proyectando, y Sergio Lucas porque aunque nunca acertó con sus mensajes bomba, le quiero, a los que se fueron, Iñaqui, Víctor, porque realmente se quedaron, a Franco y Harriero porque siempre supieron mirar mi perfil bueno, y en especial a Javi por convertirse en mi principal aliciente para poner el punto y final sobre la última página.

No agradezco a mis compañeros de promoción, sino a mis amigos, por estos años universitarios que tantos y tantos momentos nos han hecho compartir. A Héctor, porque recuerdo el comienzo del camino y me encuentro de su mano, a mis chicos, Pablo, Gus, Chus y Kiko porque nunca se cansaron de arroparme, y por supuesto a mis chicas, a Esther simplemente "quería decirle como le he dicho otras veces… que pase lo que pase estoy aquí", felicitar a Sonso por su habilidad para hacerme sonreír en el momento más triste, a Moni darle las gracias por todo lo que ella me sabe dar (P.K.), también a Bár porque un Erasmus juntas, marca y une demasiado, y con un cariño especial gracias a Ele por haber compartido conmigo tanto los mejores como los peores momentos, porque se acabó mi compañera infatigable de prácticas, días de estudio, idas y venidas, pero no se terminó y siempre seguirá, mi amiga.

Agradezco también a mis amigos de la infancia, Fati, Julián, Loli, Óscar, a muchos montalbeños y más visitantes, que aún estando fuera de las aulas, siempre han sabido estar presentes.

En un plano más familiar he tenido cerca a mucha gente que ha sabido entenderme, apoyarme y animarme en todo momento, y que hoy leyendo estas líneas, sé se sienten orgullosos. Gracias también a ellos por estar ahí, a Lucía por la luz que siempre ha mantenido encendida para mí, a mis padres Pedro y Carmen por su confianza, comprensión, y porque en su día supieron comenzar a hacer de mí, la persona que hoy soy, y en especial, gracias a mi Peter por dejar de ser un hermano para convertirse en un amigo, confidente, protector, compañero y pieza imprescindible en el puzzle de mi vida.

Por supuesto, sí son todos los que están, pero no están todos los que son.
A todos, gracias.

Este proyecto ha sido realizado en el Área de Tratamiento de Voz y Señales (ATVS) en la Escuela Politécnica superior de la Universidad Autónoma de Madrid. El proyecto ha sido financiado parcialmente por el Ministerio de Educación y Ciencia a través del proyecto TEC2006-13170-C02-01.

INDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. Introducción

1.1. Motivación

La comunicación oral es una importante fuente de información, en la actualidad grandes cantidades de datos de audio son creadas y guardadas digitalmente. El procesado de información ha sido y es una actividad económica primaria en el mundo, ésto unido al crecimiento de datos de audio accesibles por ordenador, ha creado una oportunidad y a la vez una necesidad urgente de encontrar un medio de recuperación de información inteligente de archivos de voz.

Hoy en día el éxito de ciertas aplicaciones en búsqueda de texto provoca interés en la búsqueda de otros medios. Entre estas, la búsqueda de habla es probablemente la más interesante, ya que la mayoría de la comunicación sigue ésta modalidad, y a que el habla debido a su naturaleza, es el proceso de comunicación más eficiente y económico de la sociedad. Por esta razón, desde hace mucho tiempo, investigadores en las áreas de la ciencia computacional y del lenguaje se han centrado en el estudio y desarrollo de nuevas técnicas de reconocimiento de voz.

Aunque hay gran cantidad de audio grabado públicamente, la única información que se puede obtener directamente es el título o sumario. Por ejemplo, en el caso en el que se busque información específica, discutida en una reunión de una hora de duración, se necesita emplear mucho tiempo escuchando algo que no es interesante, hasta encontrar lo que realmente se busca. En ésta y muchas otras situaciones, un sistema capaz de buscar en grabaciones de habla, sería de gran ayuda. En general, buscar en habla es necesario incluso cuando es necesario acceder a información desde grabaciones multimedia. Así, hay cantidad de aplicaciones posibles, centros de llamada, procesado de reuniones, análisis de datos multimedia, seguridad y defensa, etc.

El principal interés de un reconocedor de voz es proveer la interacción entre el hombre y los sistemas de computación. La tendencia a largo plazo en la investigación de reconocimiento de voz ha sido progresiva hacia la trascripción de las fuentes cada vez más difíciles.

En los últimos años, el reconocimiento de habla para conversaciones espontáneas ha mejorado en un grado en que, la precisión de transcripción, comparable a lo que anteriormente fue considerado eficaz para la difusión de noticias, puede ahora ser alcanzado para una diversa gama de fuentes. Esto ha inspirado el seguimiento de la investigación sobre la tecnología de búsqueda y la navegación sobre contenidos de audio. Se están desarrollando ensayos en proyectos en todo el mundo, y algunas actividades de evaluación comparativa.

En este ámbito, la detección de términos de voz, tarea y objetivo del trabajo realizado, permite localizar un término específico, definido como una secuencia de una o más palabras, rápida y exactamente, en archivos de audio amplios y heterogéneos, para ser usados finalmente como entradas de las tecnologías de recuperación de información más sofisticadas.

La línea de desarrollo comienza con una visión general sobre el estado del arte en reconocimiento de voz.

1.2. Objetivos

Suponiendo un reconocedor de voz ya listo y entrenado, que devuelve unos lattices fonéticos, el objetivo consiste en desarrollar una aplicación cuya función es detectar todas las ocurrencias de un determinado "término" en un cuerpo dado de datos de voz, siendo término, la secuencia de fonemas que componen la palabra de búsqueda.

La tarea de detección de términos de voz no es formulada como tarea de recuperación, requiere una especificación de los tiempos inicial y final por cada ocurrencia, también, sistemas que proporcionen una puntuación para estas ocurrencias, así como una decisión indicando si la ocurrencia es correcta.

Los sistemas completos de reconocimiento de información vocal suelen implementarse en dos fases: indexado y búsqueda.

En la fase de indexado, el sistema debe procesar los datos de voz sin conocimiento de los términos de consulta. Los índices de salida son guardados, y serán posteriormente accedidos durante la fase de búsqueda para localizar los términos y enlazarlos con el audio original.

La fase de búsqueda, objetivo del presente trabajo, puede ser repetida múltiples veces para cada término de consulta, por lo que la eficiencia de esta implementación es muy importante. El sistema usará los términos, el índice, y opcionalmente el audio, para detectar las ocurrencias de los términos, devolviendo una lista ordenada de las mismas.

Mientras que en búsqueda de texto los sistemas se tratan directamente como datos, en búsqueda de habla se tiene un proceso más complejo que requiere unos pasos definidos:

- Conversión de la voz a símbolos discretos que puedan ser indexados y buscados mediante un sistema reconocedor.
- Prevención frente a errores inherentes del sistema reconocedor.
- Determinación de una consulta.
- Capacidad para buscar palabras menos comunes, como nombres propios, términos técnicos y palabras mal pronunciadas.
- Mecanismo eficiente y rápido para obtener resultados de búsqueda en un tiempo razonable, incluso con grandes cantidades de datos.

Una aproximación clásica consiste en la conversión de habla en transcripciones de palabra usando una herramienta de Large Vocabulary Continuous Speech Recognition (LVCSR). En la década pasada, la mayoría de los esfuerzos de investigación en recuperación de datos de voz se han enfocado hacia las transcripciones de palabra.

Un significante inconveniente de éstas es la búsqueda de términos fuera de vocabulario que devolverán un resultado nulo, OOV (Out-of-Vocabulary) son palabras perdidas en el sistema de vocabulario del Reconocedor Automático de Voz (RAV), y remplazadas en las transcripciones de salida por alternativas probables, dado un modelo de reconocimiento acústico y un modelo de lenguaje.

La aproximación desarrollada consiste en la conversión de habla en transcripciones fonéticas y la representación de las consultas como una secuencia de fonemas. El inconveniente de esta vía es la tasa de error que se asocia a las transcripciones de manera inherente.

1.3. Organización de la memoria

La memoria consta de los siguientes capítulos:

• **Estado del arte** en reconocimiento de voz, introduciendo conceptos sobre el sonido y la fonética, una visión de los distintos enfoques en reconocimiento de voz y sus aplicaciones. Posteriormente se hace un estudio de distintos sistemas de reconocimiento existentes, y por último un análisis de los algoritmos de búsqueda en reconocimiento de voz.

• **Diseño** del proyecto, describiendo los medios disponibles para el desarrollo (Bases de datos, *Software* y *Hardware*), el sistema general y las distintas fases que lo componen.

• **Desarrollo** del proyecto, describiendo el procedimiento seguido para la implementación del sistema, y el funcionamiento de cada uno de los componentes desarrollados.

• **Pruebas y resultados** obtenidos tras el desarrollo, para la evaluación, análisis y extracción de conclusiones del sistema implementado.

• **Conclusiones y trabajo futuro** que resulte interesante a la vista de los resultados obtenidos para mejorar el sistema implementado.

• **Referencias** de las distintas fuentes de información utilizadas para el desarrollo de esta memoria**.**

• **Glosario.**

2. Estado del arte

2.1. Reconocimiento de Voz

La investigación de tecnologías en reconocimiento de voz comenzó a finales de los años 50 con la llegada de la era de la computación digital. Esto combinado con las herramientas que permiten la captura y análisis de la voz permitió a investigadores encontrar nuevos métodos de representación de las características acústicas que muestran las diferentes propiedades de las palabras.

Uno de los pioneros en este campo fue AT&T. El sistema desarrollado por esta compañía se entreno para reconocer el discurso de manera dependiente del locutor.

En la época de los años 60 la segmentación automática de voz logró avanzar en unidades lingüísticas relevantes (fonemas, palabras y sílabas), así como en la clasificación y reconocimiento de patrones. Inicialmente los investigadores subestimaron la dificultad de la tarea, sin embargo pronto comenzó la tendencia a la simplificación, con aplicaciones dependientes de locutor y con vocabularios pequeños.

En los años 70 surgieron un número de técnicas fomentadas en su mayoría por la Agencia DARPA (Defense Advanced Research Projects Agency) [Baeck96]. Se desarrollaron reconocedores basados en patrones que manejaban un dominio de reconocimiento mayor.

Los reconocedores estaban capacitados para aceptar un vocabulario más extenso. Durante esta época se logró una mejora con respecto al reconocimiento para palabras aisladas y continuas. Se desarrollaron técnicas como Dynamic Time Warping, modelado probabilístico y el algoritmo de retropropagación [Rabin93], que se describirá en una sección posterior.

Los años 80 se caracterizaron por el fuerte avance que se obtuvo en el reconocimiento de voz. Se empezaron a desarrollar aplicaciones con vocabularios grandes y se impulsó el uso de modelos probabilísticas y redes neuronales, los cuales poco a poco mejoraron su desempeño.

Para los 90 el progreso de los sistemas de reconocimiento de voz fue notable gracias a la mejora de la tecnología. Los investigadores realizaron vocabularios grandes para usarse en el entrenamiento, desarrollo y pruebas de los sistemas.

En la actualidad se continúa la mejora de las técnicas que comenzaron a desarrollarse hace algunos años, y se sigue evolucionando en busca de mejores reconocedores, nos encontramos ante un tiempo adecuado para examinar ampliamente las posibles sinergias que pueden ayudar a dar forma a la agenda de investigación de recuperación de la información.

2.1.1. Proceso de producción y percepción del habla

La producción de la voz se puede dividir en dos partes:

- *El proceso psicológico.* El individuo formula en su mente el mensaje que desea transmitir. A continuación, el mensaje es convertido en palabras o frases adecuadas que lo representan.

- *El proceso fisiológico.* El cerebro humano produce órdenes nerviosas necesarias para mover los músculos que forman los distintos órganos del tracto vocal. Los sonidos se producen cuando el aire es exhalado por los pulmones a través de la tráquea generando vibraciones en las cuerdas vocales. Estas cuerdas vocales son dos músculos gemelos y elásticos que vibran cuando el aire pasa por ellas. Después de producir la voz, el aire pasa a la cavidad bucal o nasal donde se configuran y matizan los diversos sonidos en el habla. En este proceso la onda de voz es producida y propagada.

Una vez que la señal de voz es generada y transmitida, comienza el proceso de percepción. En el oído, la señal de voz activa los órganos auditivos para producir impulsos nerviosos. Dichos impulsos son enviados al cerebro donde se logra el entendimiento del mensaje.

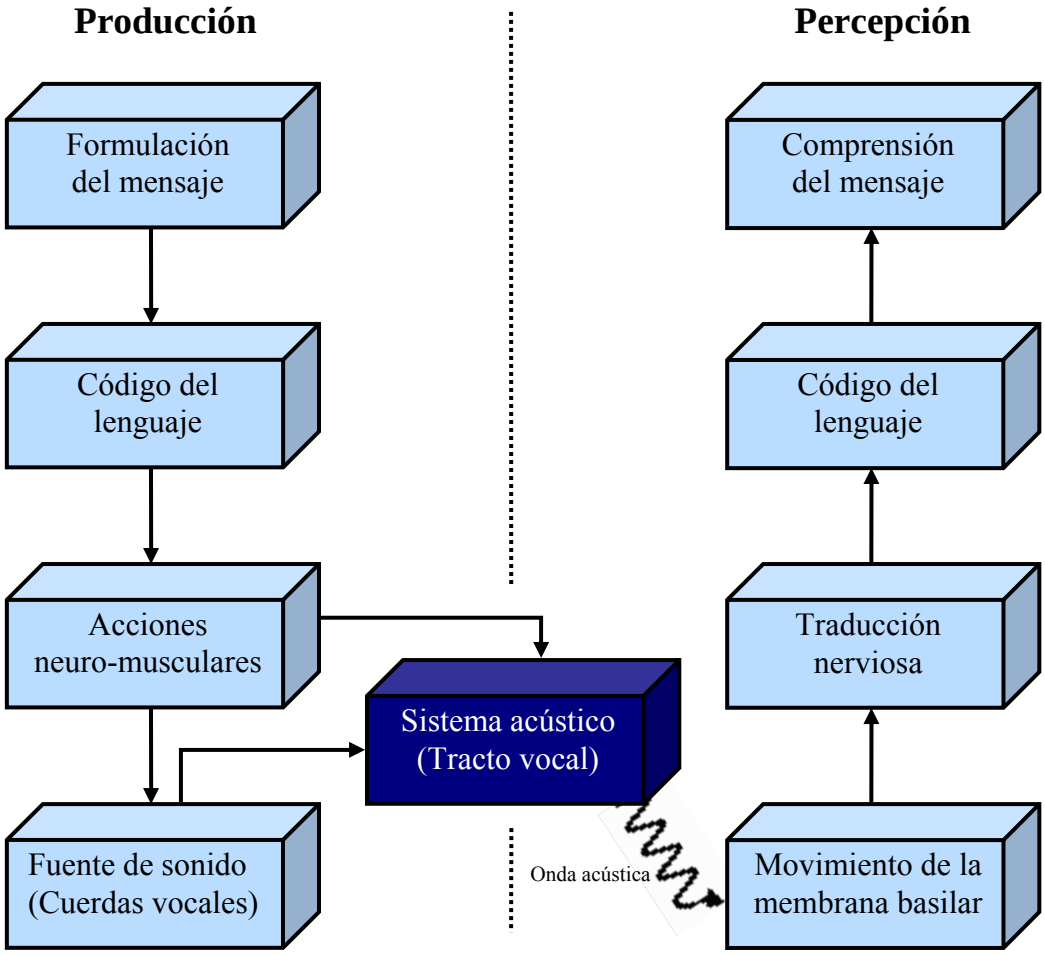


Figura 1 . Esquema de los procesos de producción y percepción del habla

2.1.2. Características acústicas

La señal de voz es un flujo continuo de sonidos y silencios. Esta señal se encuentra constituida por palabras, que a su vez se dividen en fonemas, los cuales son la unidad básica del habla y en conjunto éstos pueden determinar los sonidos con los que construir las palabras de cualquier lenguaje.

2.1.2.1. Fono y fonema

Los fonemas no son sonidos con entidad física, sino abstracciones mentales o abstracciones formales de los sonidos del habla. En este sentido, un fonema puede ser representado por una familia o clase de equivalencia de sonidos, denominados fonos, que los hablantes asocian a un sonido específico durante la producción o la percepción del habla.

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

El número de fonemas de una lengua es finito y limitado, en cada lengua el número de fonos potencialmente definibles, especialmente si son especificados rasgos fonéticos muy sutiles, es potencialmente ilimitado y varían según el contexto fonético y la articulación individual de los hablantes. En cuanto al número de fonemas no tiene por qué ser fijo, y puede cambiar con el cambio lingüístico, de hecho en un instante dado, se pueden construir dos sistemas fonológicos con diferente número de fonemas si se introducen reglas de pronunciación más complejas. En castellano el número de unidades está en torno a 24 (5 vocales y 19 consonantes), aunque no todas las variedades de castellano tienen el mismo número.

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 además es portador de una intención significativa especial.

- **Indivisible**: no se puede descomponer en unidades menores.

- **Abstracta**: no son sonidos, sino modelos o tipos ideales de sonidos.

2.1.2.2. Clasificación de los fonemas

Acústicamente, un fonema es un sonido que se distingue por un patrón característico. Por otro lado, un alófono es una de las diferentes pronunciaciones que puede llegar a tener el fonema. El conjunto de fonemas se divide de acuerdo a su manera y lugar de articulación en varios grupos:

- **Fonemas consonánticos:**

Los sonidos de las consonantes son producidos cuando el aire al salir de los pulmones encuentra un obstáculo (parcial o total), debido generalmente a la lengua y en algunos casos los labios o el velo, dejándole un espacio pequeño por el que pasar, o incluso bloqueando totalmente el paso.

La clasificación de las consonantes depende de la forma de articulación de los sonidos. Se entiende por articulación a los movimientos o configuración de los órganos vocales que producen los sonidos. Por lo tanto, las consonantes se clasifican en: oclusivas, nasales, vibrantes, fricativas y laterales.

	Bilabial		Labiodental		Interdental		Dental		Alveolar		Palatal		Velar	
	sor.	*son.*	*sor.*	*son.*	*sor.*	*son.*	*sor.*	*son.*	*sor.*	*son.*	*sor.*	*son.*	*sor.*	*son.*
Oclusiva	p	b					t	d					k	g
Nasal		m								n		ɲ		
Vibrante simple										ɾ				
Vibrante múltiple										r				
Fricativa			f		θ				s			ʝ	x	
Lateral										l		ʎ		

Tabla 1. Clasificación de fonemas consonánticos

- **Fonemas vocálicos**:

La producción del sonido de una vocal se genera cuando el aire pasa de los pulmones a la laringe y después a la boca (o nariz y boca) sin ninguna obstrucción. En las vocales, la posición de la lengua y la forma en que se abre o cierra la boca determinan el timbre, el tamaño y la forma de la onda sonora.

Las vocales se identifican por sus formantes, las cuales son muy marcadas durante todo el fonema. Esta característica las hace fácilmente distinguibles cuando se analiza su espectrograma. Además se puede decir que las vocales generalmente son de mayor duración que las consonantes.

De acuerdo a la posición de la lengua con respecto a la altura y eje horizontal, las vocales se clasifican en: cerrada, media y abierta; anterior, central y posterior:

	Anterior	Central	Posterior
Cerrada	*i*		*u*
Media	*e*		*o*
Abierta		*A*	

Tabla 2. Clasificación fonemas vocálicos

1	/a/	Fonema vocálico de apertura máxima (alófonos: [a], [ɑ]).
2	/B/	Fonema obstruyente bilabial sonoro (grafías: b, v y w, alófonos: [b], [β]).
3	/č/	Fonema africado palatal (grafía ch).
4	/D/	Fonema obstruyente coronal-alveolar sonoro (alófonos: [d], [ð]).
5	/e/	Fonema vocálico palatal de apertura media (alófonos: [e], [ɛ]).
6	/f/	Fonema fricativo labio-dental, en muchas zonas se realiza fricativo bilabial[ɸ].
7	/G/	Fonema obstruyente velar sonoro (grafías g y gu, alófonos: [g], [ɣ]).
8	/i/	Fonema vocálico palatal y apertura mínima.
9	/x/	Fonema fricativo velar (grafías g y j, alófonos: [x], [χ])
10	/k/	Fonema oclusivo velar sordo (grafías c y qu).
11	/l/	Fonema lateral (coronal-) alveolar.
12	/m/	Fonema nasal labial.
13	/n/	Fonema nasal (coronal-) alveolar.
14	/ñ/	Fonema nasal palatal.
15	/o/	Fonema vocálico velar de apertura media (alófonos: [o], [ɔ]).
16	/p/	Fonema oclusivo (bi) labial sordo.
17	/r/	Fonema vibrante simple (grafía -r-, -r).
18	/r/(rr)	Fonema vibrante múltiple (grafía -rr-, r-).
19	/s/	Fonema fricativo (coronal-) alveolar (grafía s)
20	/t/	Fonema oclusivo (coronal-) alveolar sordo.
21	/u/	Fonema vocálico velar de apertura mínima.
22	/y/	Fonema sonoro palatal (grafía y). A principio de palabra se realiza como africada palatal, en interior de palabra [ʝ] o fricativa [ʃ] y [ ]).

Tabla 3. Fonemas Castellano

2.1.3. Elementos de un reconocedor de voz

El reconocimiento de voz generalmente es utilizado como una interfaz entre el humano y la máquina, por lo que debe cumplir tres tareas clave [Mark96]:

- Preprocesamiento
- Reconocimiento

2.1.3.1. Preprocesamiento

Los sonidos consisten en cambios de presión del aire a través del tiempo, y a frecuencias que pueden ser percibidas por el oído humano. Estos sonidos pueden ser digitalizados por un micrófono o cualquier otro medio que convierte la presión del aire a pulsos eléctricos. La voz es un subconjunto de los sonidos generados por el tracto vocal.

En el preprocesamiento de la señal se extraen las características que utilizará posteriormente el reconocedor, en el proceso de extracción se divide la señal de voz en una colección de segmentos. Posteriormente, se obtiene una representación de características acústicas más distintivas para cada segmento. Con estas características obtenidas, se construye un conjunto de vectores que constituyen la entrada al siguiente módulo. Una de las representaciones más usadas son los coeficientes Linear Predictive Coding (LPC) y los coeficientes Mel-Frecuency Cepstrum Coefficients (MFCC).

Una manera de representar el sonido es graficándolo en forma de onda (waveform). En el eje horizontal representa el tiempo y el vertical la amplitud. Una limitación crucial es el hecho de que no describe explícitamente el contenido de la señal de voz en términos de propiedades.

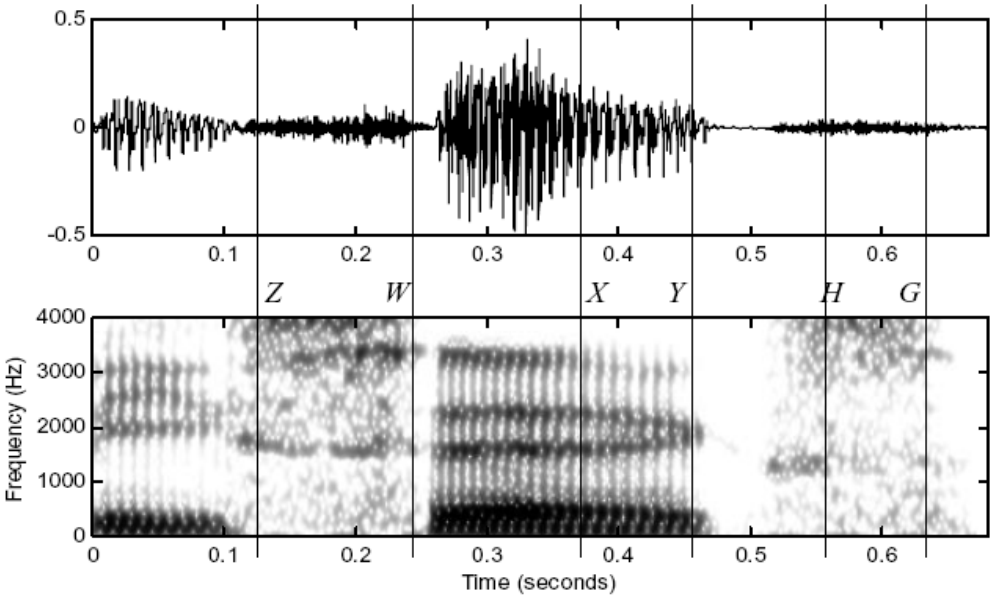


Figura 2. Señal de voz y su correspondiente espectrograma de ancho de banda [Huang01].

Una representación más adecuada para el análisis de la señal de voz son los espectrogramas. Un espectrograma es una representación de la señal de voz de acuerdo a las variaciones de la energía, con respecto al tiempo y la frecuencia. El espectrograma contiene mucha información y revela las características acústicas específicas del habla.

Las bandas oscuras corresponden a las concentraciones de energía y son llamadas formantes. Los formantes son las frecuencias en las que ocurre la resonancia de las vibraciones vocales. Las formantes son colocadas de menor a mayor frecuencia tales como: F1 (500Hz) o primera formante, F2 (1500Hz) o segunda formante, F3 (2500Hz) o tercera formante y así sucesivamente. Aunque existen formantes más altas, habitualmente las tres primeras son las necesarias para identificar un fonema.

Los espectrogramas son útiles para un análisis visual de la señal. Sin embargo un reconocedor debe extraer de la señal acústica sólo la información que requiere para poder reconocer una frase. Para ello la señal se muestrea a cierta frecuencia, se cuantifica y posteriormente se crean vectores de características. Estas últimas son las utilizadas por el reconocedor.

2.1.3.2. Reconocimiento

En esta etapa se traduce la señal de entrada a su texto correspondiente. Durante el proceso se busca clasificar los vectores de características de la señal de entrada para obtener las unidades lingüísticas de las que esta formada. Posteriormente, se realiza una búsqueda para encontrar la secuencia de segmentos con mayor probabilidad de ser reconocidos.

2.1.4. Enfoques del reconocimiento de voz

Destacan dos enfoques principales a la hora de plantear un sistema de reconocimiento de voz [Dell93]. Estos son:

- ***Enfoque acústico-fonético***: Éste engloba aquellos procesos destinados a realizar una decodificación de palabras, a partir de las características diferenciadoras de la voz, y un conjunto de reglas existentes en el habla.

- ***Enfoque de patrones*****:** Son técnicas basadas en el reconocimiento de patrones. A partir de un conjunto de modelos captados automáticamente en fases de entrenamiento, la pronunciación es decodificada. Las dos fases que componen éste método, son: la fase de entrenamiento, donde se generan los modelos de referencia, empleando un conjunto de bases de datos de voces grabadas con la suficiente variabilidad; y la fase de reconocimiento, donde se realiza una comparación entre las pronunciaciones y las referencias obtenidas, eligiendo la secuencia de palabras cuya distancia a los modelos de referencia sea menor.

2.1.4.1. Enfoque Acústico-Fonético

El enfoque acústico-fonético se apoya en el número finito y diferenciado de unidades fonéticas que contiene el lenguaje hablado. Los fonemas se estructuran en palabras, y éstas en frases que modelan las ideas. Este enfoque se basa en el diseño de sistemas expertos que gobiernan el lenguaje mediante reglas, a partir del análisis de la señal de voz.

En un reconocedor de voz basado en el enfoque acústico-fonético se pueden distinguir un conjunto de bloques de procesado:

· *Análisis acústico de la señal de voz*: Se tiene una señal de voz, obtenida a través de un micrófono, a partir de las variaciones de presión, sobre ella se aplica un procesado inicial, que es básicamente la transformación de la señal de su dominio temporal a frecuencial haciendo uso de la transformada de Fourier en un banco de filtros perceptuales. Otros análisis que también se utilizan son la envolvente espectral obtenida a través del cálculo de los coeficientes de predicción lineal, en inglés Linear Predictive Coding (LPC), o el número de cruces por cero. Las características obtenidas se denominan *características acústicas*.

· *Detección de características fonéticas*: Tras el análisis acústico de la señal de voz, se realiza la extracción de los parámetros denominados *características fonéticas*. Una característica fonética es la propiedad mínima que presenta un fragmento de la señal de voz y que diferencia a dos unidades diferentes. Entre las características fonéticas más usuales se encuentran, el cálculo de la frecuencia fundamental de las cuerdas vocales, los valores de los formantes de la voz, la detección del grado de sonoridad, etc.

· *Fase de segmentación y etiquetado*: La señal de voz es dividida en regiones en las que las características fonéticas son similares, para que puedan ser asignadas a una o varias categorías fonéticas, posteriormente utilizadas para realizar la decodificación de la secuencia de palabras.

· *Fase de discriminación*: Su función es la decodificación de las palabras pronunciadas a partir del conjunto de categorías fonéticas obtenidas en la fase previa. Usa reglas sintácticas y semánticas obtenidas mediante el estudio de la señal de voz, y que imponen restricciones para llevar a cabo la discriminación entre las palabras pronunciadas.

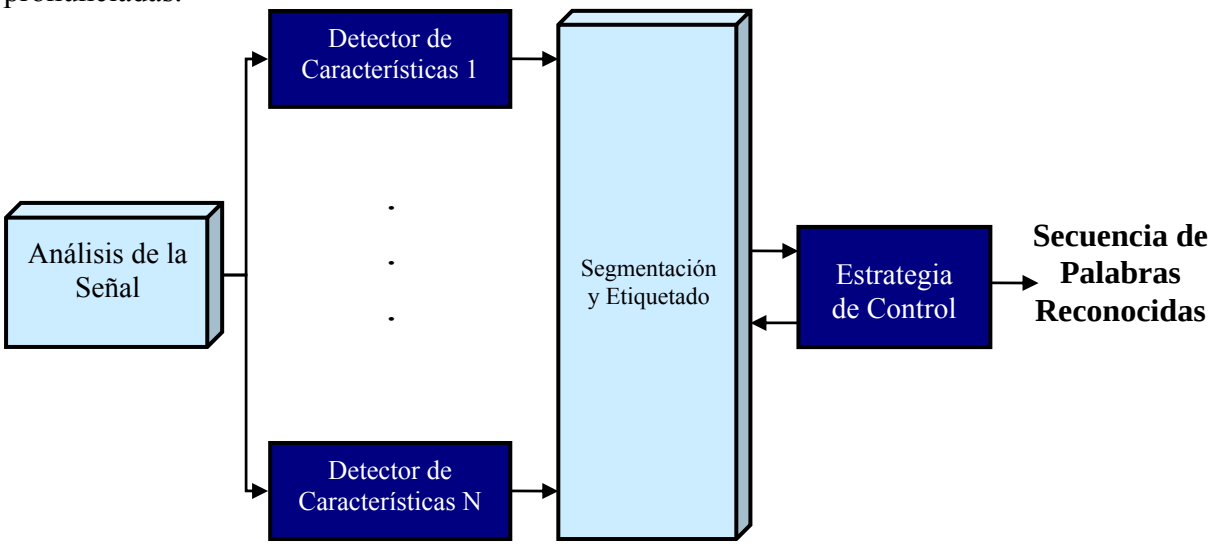


Figura 3. Bloques de un reconocedor de voz basado en un enfoque acústico-fonético.

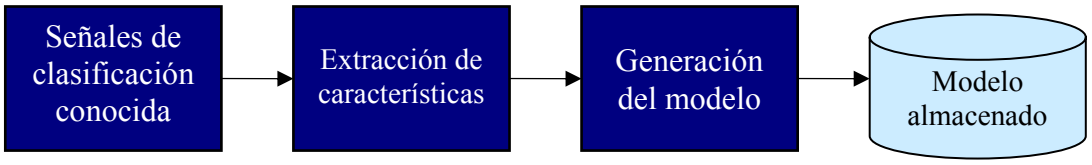
2.1.4.2. Enfoque de Patrones

El enfoque basado en el reconocimiento de patrones cuenta en la actualidad con un mayor desarrollo y en general ofrece mejores resultados. Se pueden diferenciar dos modos de funcionamiento dentro de un reconocedor de patrones [Cox88]:

· *Modo de entrenamiento*: Cada clase a ser entrenada posee su modelo creado a partir de ejemplos que sirven como referencia.

· *Modo de reconocimiento*: Se compara mediante una determinada métrica, el patrón a reconocer con todos los modelos de clase y pasa a identificarse con el más próximo.

Modo de entrenamiento



Modo de reconocimiento

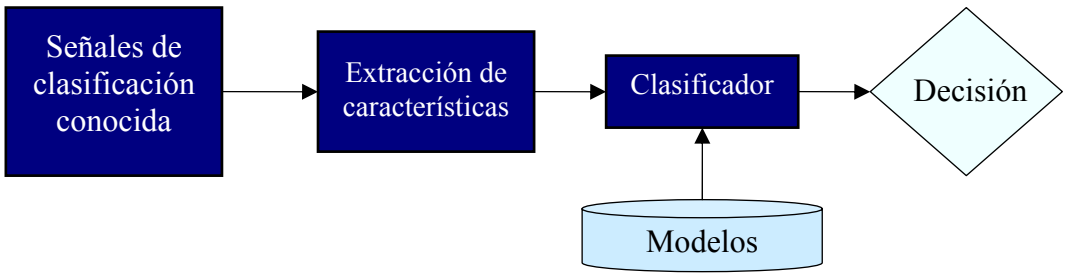


Figura 4. Sistema de clasificación de patrones en modo de entrenamiento y de reconocimiento.

En reconocimiento de voz resulta muy importante la selección de las características, ya que la precisión de reconocimiento depende en gran medida del tipo y número de parámetros usados.

Es suficiente extraer las características entre 10 o 20*ms* debido a la lenta variación de las articulaciones utilizadas en la producción de la voz.

Se puede hablar de distintas metodologías para el reconocimiento de voz basado en patrones, éstas son:

- *Reconocimiento determinista*: la metodología más usada es la *alineación dinámica de patrones*, en ingles *Dynamic Time Warping* (*DTW*) [Vins68].
- *Reconocimiento estocástico*: son los más representativos son los modelos ocultos de Markov, más conocidos por su denominación en inglés como *Hidden Markov Models* (*HMM*) [Baker75a] [Jelin76], tienen mayor capacidad de modelado que los *DTW*.
- Reconocimiento basado en las denominadas *redes neuronales* (*ANN*: *Artificial Neural Networks*).

2.1.4.2.1. Dynamic Time Warping (DTW)

Dynamic Time Warping (*DTW*), o *alineamiento dinámico de patrones*, se basa en la alineación temporal de las características de una pronunciación respecto a un patrón de referencia, antes de calcular una puntuación general.

La palabra inglesa *warped* da la clave del sistema, significa que la referencia se adapta estirándose o comprimiéndose en el dominio temporal, así los vectores de características se alinean con el patrón antes de hacer la medida de distancia.

La principal ventaja de esta técnica es su sencillez, usa algoritmos muy simples y con un grado de computación reducida, por lo que utilización se extiende en aplicaciones sencillas que obtienen muy buenos resultados.

El principal inconveniente es la dificultad de generalizar, ya que requiere un alto procesado en tareas muy complejas con grandes vocabularios [Dell93].

El *DTW* llegó a ser la metodología más madura y de consecución de los mejores resultados en reconocimiento de palabras antes de la aparición de los modelos de Markov, a partir de este momento, éstos tomaron el protagonismo, dada su mayor flexibilidad, robustez y su capacidad de abordar problemas más complejos.

2.1.4.2.2. Modelos Ocultos de Markov (HMM)

Los modelos ocultos de Markov, Hidden Markov Models (HMM) se han convertido en uno de los métodos estadísticos más potentes para modelar las señales del habla. Sus principios han sido utilizados con éxito en reconocimiento de voz automático, análisis de tonos y formantes, mejora de expresión, síntesis de voz, comprensión de lenguaje hablado, etc [Huan01].

Un HMM es una máquina de estados finita, en la que las observaciones son una función probabilística del estado, es decir, el modelo es un proceso doblemente estocástico formado por un proceso estocástico oculto no observable directamente, que corresponde a las transiciones entre estados y un proceso estocástico observable cuya salida es la secuencia de vectores espectrales.

Después de años de investigación y desarrollo, la precisión del reconocimiento de voz sigue siendo uno de los más importantes retos a la investigación. Una vez conocidos los factores que determinan la exactitud; los más notables son variaciones en su contexto, micrófono, y en medio ambiente. El modelado acústico desempeña un papel fundamental en la mejora de la precisión y podría decirse que es la parte central de cualquier sistema de reconocimiento de voz.

Para una observación acústica dada $X = x_1 x_2 ... x_n$, el objetivo del reconocimiento de voz es averiguar la secuencia de palabra correspondiente $W = w_1 w_2 ... w_n$ que tiene la máxima probabilidad posterior P (W | X) expresada como:

$$\hat{W} = \arg_W \max P(W/X) = \arg_W \max \frac{P(W)P(W/X)}{P(X)} \quad (1)$$

Dado que su maximización se lleva a cabo con la observación de una X fija, es equivalente a la maximización de la ecuación siguiente:

$$\hat{W} = \arg_W \max P(W)P(W/X) \quad (2)$$

El desafío práctico es cómo construir modelos acústicos precisos, $P(X/W)$, y modelos de lenguaje, $P(W)$, que puedan reflejar realmente el lenguaje hablado para poder ser posteriormente reconocidos. Por reconocimiento de voz de gran vocabulario, debido al gran número de palabras, una palabra ha de descomponerse en una secuencia de sub-palabras. Por lo tanto $P(X/W)$ está estrechamente relacionada con el modelado fonético, debería tener en cuenta las variaciones del hablante, de la pronunciación, del medio ambiente, y variaciones fonéticas dependientes del contexto.

Elementos de un HMM

Suponiendo un HMM discreto en que las observaciones posibles pertenecen a un conjunto discreto [Huan01], entonces el HMM vendrá dado por:

- N: el número de estados del modelo, donde q_t denota el estado en el instante de tiempo t. Los HMMs habitualmente están compuestos por X estados. Ni el estado 1 ni el X generan salida.

$$S=\{s_1,s_2,...,s_N\} \quad (3)$$

- La dimensión del conjunto de observaciones distintas de salida M, es decir el tamaño del alfabeto.

$$V=\{v_1,v_2,...,v_M\} \quad (4)$$

- La distribución de probabilidad de transición entre estados $A=\{a_{ij}\}$:

$$a_{ij}=P(q_t=s_j \mid q_t\text{-}1=si)\ 1\leq i,j \leq N \quad (5)$$

- La distribución de probabilidades de emisión de símbolos entre estados $B = \{b_j(k)\}$, donde O_k es un símbolo perteneciente a V:

$$b_j(O_k)=P(O_k \mid q_t = sj)\ 1\leq j \leq N,\ 1\leq k \leq M \quad (6)$$

- Distribución del estado inicial $\pi = \{\pi_i\}$:

$$\pi_i = P(q_0 = si)\ 1\leq i \leq N \quad (7)$$

Con todo esto, un HMM se describe como $\lambda = \{A, B, \pi\}$.

Problemas a resolver para utilizar HMMs

Dada la definición anterior para un HMM, existen tres problemas básicos a resolver para que el modelo pueda ser usado en aplicaciones reales [Huan01]:

> **1. Problema de Evaluación**: Dada una secuencia de observaciones y un modelo, se busca cómo calcular la probabilidad de que la secuencia observada haya sido producida por dicho modelo.
> **2. Problema de Decodificación**: Dada una secuencia de observaciones y un modelo, se busca cómo elegir una secuencia de estados que sea óptima en algún sentido.
> **3. Problema de aprendizaje:** Dada una secuencia de observaciones de entrenamiento, se busca cómo obtener los parámetros del modelo de forma óptima.

2.1.4.2.3. Redes Neuronales

Las denominadas *redes neuronales* (*ANN*: *Artificial Neural Networks*) se complementan con otras estrategias, como por ejemplo *HMM*, formando *sistemas híbridos*. Existen problemas concretos que se resuelven mediante este enfoque dada su sencillez y eficiencia, como por ejemplo el diseño de un cuantificador vectorial [Beal90].

Las ANN son estructuras con capacidad de clasificación y discriminación, intrínsecamente no lineales, que les permite aprender una determinada tarea a partir de pares observación-objetivo, sin necesidad de suposiciones sobre el modelo subyacente.

El elemento fundamental de estas técnicas es la *neurona,* que presenta un funcionamiento similar aunque muy simplificado, a las neuronas de cualquier sistema nervioso. Son definidas a partir de un conjunto de entradas y de salidas conectadas entre sí, creando una estructura que almacena la información.

Una ventaja es la sencillez con la que la red de neuronas puede ser entrenada de forma automática mediante datos de referencia, es decir su *aprendizaje*, así como la posibilidad de procesado en paralelo, lo que provoca el requerimiento de equipos rápidos y potentes.

2.1.5. Aplicaciones del reconocimiento de voz

El objetivo principal de los sistemas de reconocimiento de voz es desarrollar interfaces centradas en las necesidades del usuario aprovechando una de las capacidades que tiene el hombre para comunicarse, la expresión oral. Estos sistemas han probado su utilidad para ciertas aplicaciones. Uno de los medios más populares para las aplicaciones de voz es el teléfono por razones de facilidad de uso, disponibilidad y coste.

Las aplicaciones basadas en este tipo de reconocedores son servicios financieros, asistencia de directorio, llamadas a cobro revertido (operadora automática), transferencia de llamadas telefónicas, consultas de información (clima, tráfico, reservas). Las ventajas que presenta este tipo de aplicaciones son que al interactuar el usuario utiliza la eficiencia del habla (rápida, flexible, natural) y está libre de movimientos de las manos en caso de que las tenga ocupadas.

Existen otras aplicaciones que no se basan en el teléfono, por ejemplo el dictado automático. También el reconocimiento de voz es usado en compañías en donde la entrada de datos o comandos por voz es requerida tales como desarrollo de inventarios, control de robots, etc.

2.2. Sistemas de reconocimiento

2.2.1. Reconocimiento Automático de Voz (RAV)

El área del reconocimiento automático de voz (RAV), plantea como objetivo fundamental la transcripción automática de la señal de voz mediante máquinas, la conversión de habla en texto.

El habla es sin duda el método de comunicación más intuitivo y natural para los seres humanos. La Tecnología del Habla, en sentido general, está gozando de un interés cada vez mayor por parte de la comunidad científica internacional y la sociedad en general y, dentro de ésta, el reconocimiento automático de voz presenta uno de los campos más atractivos de investigación.

Hasta hace pocos años, los sistemas RAV de mediana y gran complejidad, estaban disponibles casi únicamente como prototipos de laboratorio. En la actualidad, empresas como Philips, IBM, Lernout & Hauspie y Dragon Systems han ganado terreno en el mercado con productos para dictado de texto de gran calidad.

También existen sistemas comerciales de reconocimiento automático de voz incluyendo Nexidia/FastTalk, Virage/AudioLogger, Convera así como sistemas de investigación como AT&T DVL, AT&T ScanMail, BNN rough'n'Ready, CMU Informedia, SpeechBot, entre otros.

Los sistemas RAV han sido aplicados en gran parte a la tarea de búsqueda de audio. Existen distintas tipologías, pero la mayoría de estos sistemas producen una trascripción de los datos de audio, y a partir de ésta, aplican métodos de búsqueda basada en texto.

La mayoría de los sistemas RAV tienen un vocabulario cerrado, debido a la cantidad de datos finita que se usa para entrenar los modelos de lenguaje. Normalmente el vocabulario está formado por palabras aparecidas en corpus de entrenamiento, y en ocasiones se reduce para que incluya las palabras más frecuentes del corpus. Las palabras que no están dentro del vocabulario definido son palabras –out of vocabulary (OOV)- fuera de vocabulario, no serán reconocidas por los sistemas RAV, contribuyendo a errores de reconocimiento. Usar búsqueda fonética ayuda a la recuperación de palabras OOV.

2.2.1.1. Modelado

Los modelos ocultos de Markov (HMM) constituyen la técnica de modelado más utilizada en los laboratorios de todo el mundo, desde los trabajos pioneros de Baker [Baker75b] y Jelinek [Jelin76]. Son lo suficientemente potentes como para modelar adecuadamente la mayor parte de las fuentes de variabilidad presentes en el habla[Lee89].

Se puede hacer una clasificación en función de la naturaleza de las distribuciones que modelan las observaciones acústicas:

- Distribuciones en un espacio discreto de símbolos, *Modelos discretos de Markov* (DDHMM) [Huang90b] [Hasan90]. Las observaciones acústicas son símbolos pertenecientes a un alfabeto finito, usan técnicas de cuantificación vectorial para transformar el vector de parámetros de entrada [Gray84] [Rabin86].

- Distribuciones de probabilidad de las observaciones en un espacio continuo, *Modelos Continuos de Markov* (CDHMM) [Soudo86][Huang90b]. Necesitan aplicar restricciones para limitar la complejidad de la estimación y cálculo de probabilidades. El mecanismo más usual caracteriza cada modelo como una mezcla de funciones del mismo tipo (generalmente gausianas).

- Modelos Semicontinuos de Markov (SCDHMM) [Huang89] [Huang90a], una unificación de los dos anteriores, comparten el mismo conjunto de funciones de densidad de probabilidad para distintos modelos, variando únicamente los pesos de ponderación aplicados a cada una.

La segunda alternativa en éxito y utilización de cara a abordar el modelado acústico en RAV, la constituyen las redes neuronales (NN) [Morg91], poseen características que las hacen atractivas para su aplicación en sistemas RAV. Pero como inconvenientes requieren un elevado tiempo de entrenamiento y son complejas para el modelado temporal.

Se han desarrollado arquitecturas y soluciones concretas para reducir en lo posible sus defectos en el modelado temporal (redes neuronales recurrentes (RNN) [Robin91] y de retardo temporal (TDNN) [Lang90]) y el coste de entrenamiento [Menén94]. En los últimos años, enfoques híbridos basados en HMM y NN han mostrado éxitos notables [Bourl93] [Renal94] [Menén94] [Hoch94] [Robin95] [Morg95].

2.2.1.2. Arquitecturas

Ante el proceso de diseño de un sistema RAV, uno de los planteamientos es la especificación de la arquitectura modular del sistema.

En la Figura 5. se muestra el esquema genérico de un sistema RAV, la señal de entrada se representa generalmente por una secuencia temporal de vectores de parámetros, que son calculados mediante análisis localizado. En el caso general, la idea es transformar dicha representación inicial en un conjunto de parámetros más elaborados, más adaptados a la tarea de discriminación posterior.

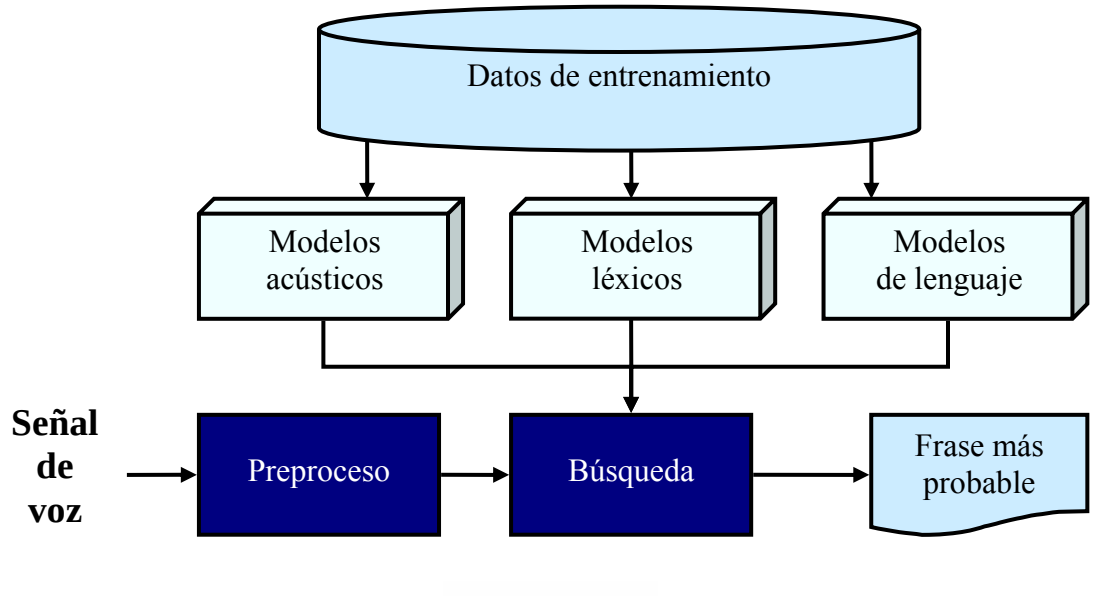


Figura 5. Esquema genérico de un RAV

La selección de los parámetros a usar sigue siendo un problema de difícil solución [Bourl96], aunque la característica común a la mayoría de los enfoques, pasa por un proceso de filtrado de la secuencia de entrada [Soong86]. En [Nadeu97] y [Cheng97] pueden encontrarse desarrollos de la formulación del proceso de parametrización.

2.2.1.3. Alfabetos y diccionarios

Se entiende por alfabeto, el conjunto de unidades utilizadas para modelar el habla, teniendo un modelo acústico para cada uno de ellos, y por diccionario, la lista de palabras válidas en un entorno o aplicación, que quedan representadas a partir de las unidades que conforman el alfabeto.

Independientemente del tipo de unidades seleccionadas, deben cumplir, una serie de propiedades [Holt98]:

1. **Consistencia**, diferentes realizaciones de la misma unidad deberán tener características similares.
2. **Entrenabilidad**, la base de datos de entrenamiento dispone de un número suficientemente elevado de ejemplos de cada unidad, como para conseguir modelos acústicos fiables.

3. **Economía**, mantener su número dentro de ciertos límites, para evitar problemas de aumento de carga computacional, o dificultad de entrenamiento.
4. **Cobertura**, las unidades seleccionadas han de cubrir razonablemente todos los eventos acústicos posibles.
5. **Concatenabilidad**, con la idea de que las palabras puedan ser descompuestas como una secuencia de dichas unidades.

2.2.2. Sistemas de detección de atributos de voz

Hay un nuevo paradigma en detección basada en RAV, propuesto para dirigir algunas de las limitaciones de los sistemas modernos y para estrechar el salto entre RAV y reconocimiento de habla humana.

Específicamente, se desarrollan métodos de detección de diseño en la transcripción automática de atributos de habla, se incorporan detectores de varios atributos del habla, características distintivas, fonológicas o acústico-fonéticas, en una aproximación a sistemas RAV. Son conocidos distintos sistemas:

- Multi-Layer Perceptrons (MLPs),
- Hidden Markov Models (HMM),
- SupportVector Machines (SVMs) .

2.2.2.1. Detectores de atributos

2.2.2.1.1. Detectores de atributos basados en MLP

Una red neuronal Multi-Layered Perceptron (MLP) está capacitada para estimar la probabilidad a posteriori de fonemas con evidencias acústicas en sus entradas [Joel07]. Un MLP es capaz de estimar un fonema identidad con suficiente precisión porque es entrenado utilizando suficiente contexto temporal y aprende a discriminar entre los fonemas. Por otra parte, los errores de un MLP son sistemáticos y pueden ser capturados en forma de matriz confusión. Una desventaja es que un desajuste entre la cadena fonética de la palabra clave (obtenida a partir de una búsqueda el diccionario) y los fonemas posteriores del MLP, provocaría un fallo en la detección de palabras clave. Este desfase se debe, al error del hablante que no pronuncia la palabra correctamente, o al error de la máquina debido a una confusión acústica.

2.2.2.1.2. Detectores de atributos basados en HMM.

Se modela cada atributo fonético 'target' y su 'anti-target' con HMMs. Ambos HMMs son entonces Viterbis alineados a cada segmento. Para una observación O, el detector de puntuaciones es computado como el radio de probabilidad logarítmica (Log-Likelihood Ratio).
$LLR(\mathrm{O}) = \log \mathrm{L}(\mathrm{O}|\lambda_o) - \log \mathrm{L}(\mathrm{O}|\lambda_1)$ donde $L(\mathrm{O}|\lambda_o)$ y $L(\mathrm{O}|\lambda_1)$ son probabilidades acústicas de modelos target y anti-target respectivamente [Jiny05] [Sabat06].

Así este ratio de probabilidad es computado cuando se analiza un test de observación, y contrastándolo con una decisión de umbral, se decide cual de las dos hipótesis ha de ser aceptada.

2.2.2.1.3. Detectores de atributos basados en SVM

Las máquinas kernel están incrementando en gran medida la familia de métodos para reconocimiento de patrones. Entre las máquinas Kernel, Support Vector Machines (SVM), son las más usadas y han sido aplicadas a muchos problemas de reconocimiento de patrones, incluido el reconocimiento de habla. Los métodos de máquinas Kernel tienen la capacidad de usar técnicas de clasificación lineal para envolver productos interiores en un espacio no lineal, logrando así impresionantes resultados en muchas tareas de clasificación.

La curva DET (Detector Error Trade-off), así como la ROC (Receiver Operating Characteristics) dibujan la posición de un detector exactamente sobre un rango completo de valores umbral. Éste tipo de análisis puede ser usado para examinar la compensación entre el número de falsas alarmas y rechazos que un detector producirá, y es usado extensivamente en el desarrollo de detectores de atributos de habla.

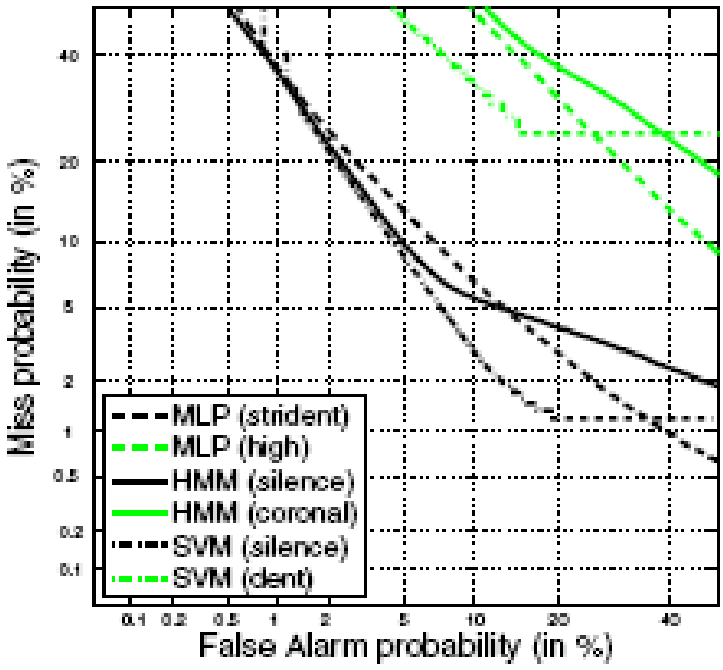


Figura 6. Curvas DET seleccionadas de detectores MLP, HMM y SVM [Joel07].

En la figura superior, ejemplos representativos de los mejores y peores comportamientos de los atributos para detectores HMM, MLP y SVM. HMMs y SVMs son los mejores para detectar silencios mientras que MLPs detectan mejor los atributos estridentes.

2.2.2.2. Diseño de un detector de atributos

Un ejemplo de front-end de un transcriptor automático de atributos de habla basado en detección de sistemas RAV se muestra en la Figura 7. La señal de habla es primero analizada por un banco de detectores, cada uno produce un resultado de confianza y probabilidad a posteriori, perteneciendo a algún atributo acústico-fonético. El diseño de estos detectores, la optimización de sus parámetros y la selección del conjunto de atributos a detectar, son problemas críticos de diseño para el paradigma de la detección basada en RAV [Ilan07].

La salida de los detectores pasará a un módulo de combinación que forma parte de la fase de decodificación, una vez combinados, se llevará a cabo la tarea de reconocimiento.

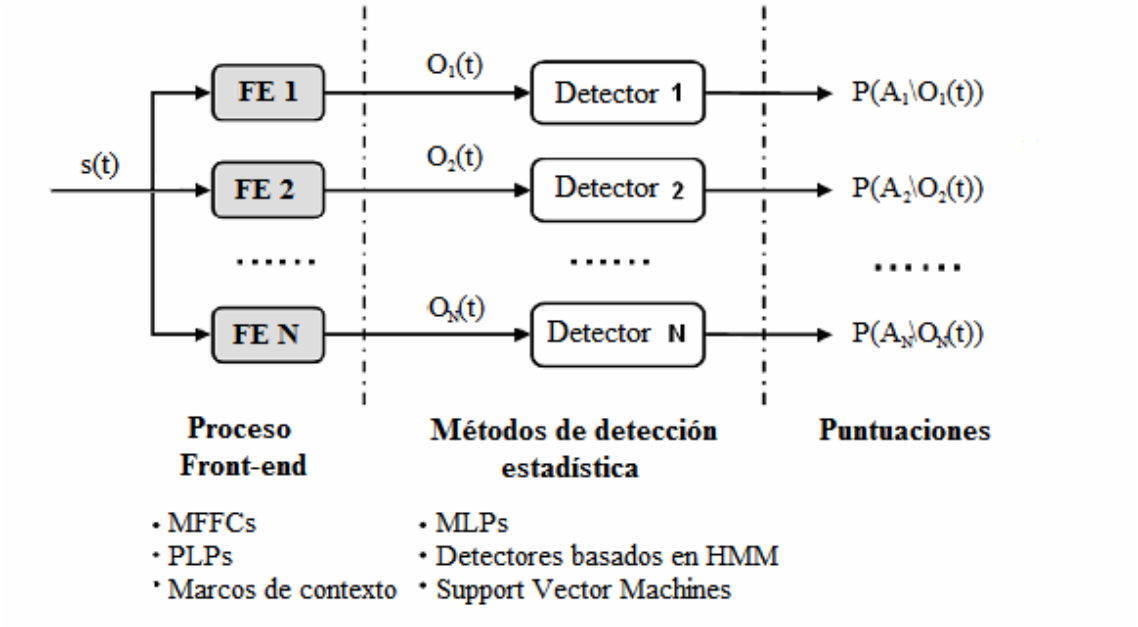


Figura 7. Front end de un detector de atributos

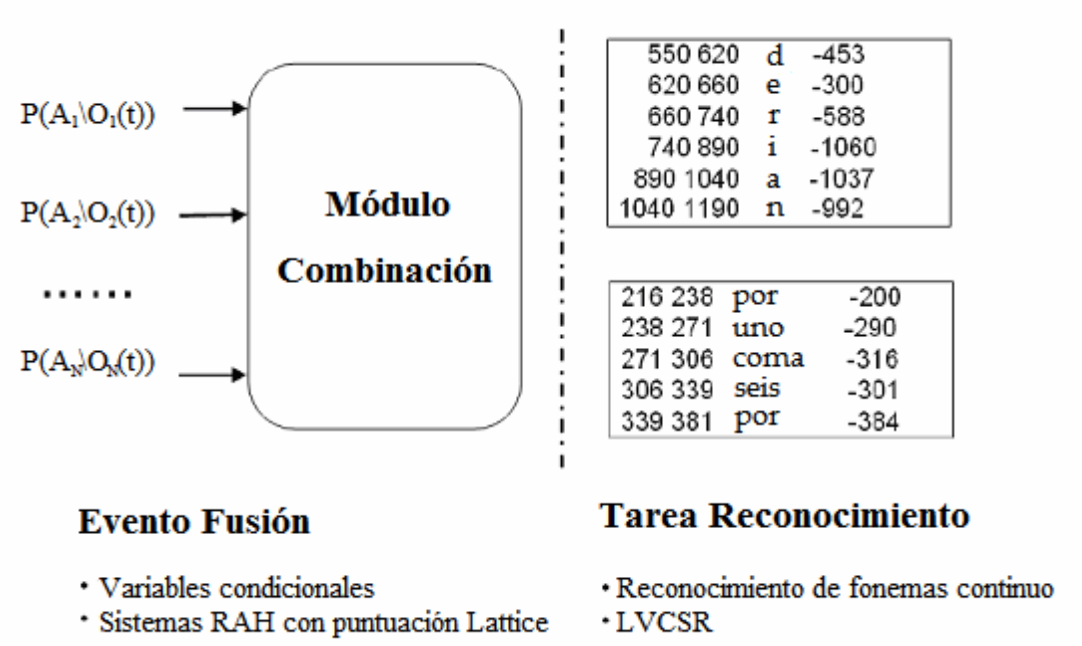


Figura 8. Fase de decodificación de un detector de atributos

2.2.3. Sistemas de detección de palabra simple

Aunque la investigación de reconocimiento automático de voz ha logrado grandes éxitos en las últimas décadas, existen todavía algunos problemas retadores. Entre ellos, como se ha descrito anteriormente, la detección de palabras fuera de vocabulario. Parte de la razón causante de estas dificultades es que el conocimiento sobre fonética, fonología y lingüística no ha sido completamente integrado en sistemas RAV [Lee03][Lee04].

La detección de atributos de habla, fonos y palabras es un componente clave de los sistemas basados en detección RAV [Kirch98].

En el nivel convencional de wordspotting los sistemas directamente trabajan con características acústicas y enfocan sólo palabras contenido, que son generalmente de gran tamaño [James94][Szök05].

2.2.3.1. Wordspotting

Wordspotting es la habilidad para localizar una palabra clave o frase en un contexto de habla fluida. En wordspotting se diferencia entre el reconocimiento de palabras aisladas [Drag90], y el reconocimiento de habla continua, las palabras son reconocidas en un flujo continuo [Bak91].

Es útil en tareas tales como la edición de audio de voz e indexado de archivos. En edición de voz, wordspotting es usado para encontrar automáticamente los límites de palabras clave o frases, y para sustituciones, borrados o inserciones. También proporciona indexado por palabras clave dentro de ficheros de audio grandes, con ello se puede recuperar información específica sin la necesidad de escuchar grabaciones completas.

El wordspotter esta basado en HMM [Rabin89]. Es generalmente dependiente del hablante y requiere un entrenamiento inicial que se realiza en un segmento de habla grabado, durante la fase de entrenamiento un conjunto de unidades acústicas son adquiridas y usadas para modelar el habla. Estas unidades acústicas son obtenidas por grupos sin supervisar, o por un vector de cuantificación [Gray84], de un segmento arbitrario del habla.

2.2.3.2. Diseño de un detector de palabra simple

Un detector de palabra simple es esencialmente un sistema Key-Wordspotting (KWS) que proporciona una detección sencilla de palabras en habla continua. En la investigación convencional KWS, las palabras función son generalmente tratadas como “stopwords” que fueron excluidas de la lista de palabras clave. En general, los métodos KWS se pueden clasificar en distintos grupos. El primero es basado en Large Vocabulary Continuous Speech Recognition (LVCSR), lattice de palabra de vocabulario dependiente, o lattice de fonema de vocabulario independiente. Otro es el método basado en palabras de relleno, y por último el reconocedor de fonemas y búsqueda en lattices.

Investigaciones previas muestran que el método basado en lattices de palabra tiene mejor comportamiento, sin embargo, la desventaja es que el vocabulario y el lenguaje son fijados por adelantado [Szök05].

La métrica es de crucial importancia en un sistema KWS. Ya que los resultados de probabilidad de la clave cambian con frecuencia (dependiendo principalmente de la longitud de la palabra), es necesaria alguna normalización para el cómputo de la confianza. Las medidas más usadas comúnmente son la tasa de probabilidad, la probabilidad posterior local y sus variantes [Rose90].

Dada una secuencia de observación de un segmento de palabra hipotético, $O = \{o_t, t = t_s, \cdots, t_e\}$, el LLR (Log-Likelihood Rate) se define como:

$$LLR(O) = \frac{1}{t_e - t_s} \sum_{t=t_s}^{t_e} \log \frac{p(o_t / \Lambda_k)}{p(o_t / \Lambda_b)} \quad (8)$$

Dónde, t_s y t_e son los índices de tiempos iniciales y finales de un segmento detectado, y Λ_k y Λ_b son los modelos de palabra clave y fondo, respectivamente.

2.2.3.2.1. Large-vocabulary Continuous Speech Recognition(LVCSR)

El propósito de los reconocedores LVCSR es desarrollar todos los aspectos del reconocimiento de voz en el dominio espontáneo, el habla en conversaciones entre humanos. Esto incluye características de extracción, tratado de transcripciones automáticas de atributos de habla, modelado de lenguaje y entendimiento del habla. Se utiliza un reconocedor de palabras completo que transcribe la grabación íntegramente [Toled08].

- ✓ Las búsquedas son muy rápidas (se buscan las palabras en la cadena de texto reconocida).
- ✗ La indexación se realiza con un conjunto de palabras grande pero predefinido: Si la palabra a buscar no se utilizó en la indexación no se puede buscar (Out-Of-Vocabulary, OOV).

2.2.3.2.2. Método basado en la palabra clave de relleno

Selección del modelo de relleno y fondo

Se usan modelos de relleno para ocupar los intervalos de habla sin palabras clave y modelos fondo para calcular la medida de confianza para palabras clave. El modelo más usado es el bucle-fonema.

En la selección del modelo de relleno, pueden usarse distintas clases de modelos fonéticos (vocal, nasal, oclusiva, etc) y proporcionan un mejor comportamiento que un modelo de bucle-fonema. Esto es porque con los modelos de relleno menos precisos, la palabra clave tiene más oportunidades de ocurrir incluso en casos de pesos de penalización inapropiados. El modelo fondo es entrenado usando observaciones acústicas desde todos los fonemas.

Diseño de la red gramatical

Generalmente, se usa una red gramatical de respuesta directa para wordspotting [Szök05]. Los modelos de relleno se colocan en paralelo al modelo de palabra clave. Con una determinada condición acústica y un peso de penalización elegido apropiadamente, esta red puede generar resultados muy buenos y con alta precisión. Sin embargo, el comportamiento de la palabra clave depende en gran medida de la elección de los pesos de penalización.

Figura 9. Red grammatical de respuesta directa

El rendimiento en la detección de una sola palabra mejora en gran medida en relación al sistema convencional KWS. Sólo se intentan reconocer unas palabras determinadas, el resto de la voz se asigna a modelos de relleno.

✓ Es una técnica muy precisa y menos costosa que la anterior.

✘ Debe operar con un conjunto predefinido de palabras clave, si se quiere buscar otra palabra (OOV) es necesario rehacer la indexación.

2.2.3.2.3. Reconocedor de fonemas y búsqueda en lattices

Un tercer método apto para KWS es el que se desarrolla en este proyecto, un reconocedor de fonemas y búsqueda en lattices.

Definición de lattice

Un lattice $L = (N,A, n_{\text{inicio}}, n_{\text{fin}})$ es un grafo dirigido y acíclico en el que N representa el número de nodos, A el número de arcos, y $n_{\text{inicio}}, n_{\text{fin}} \in N$ los nodos inicial y final respectivamente [Meng07].

Cada nodo $n \in$ N tiene asociado un tiempo $t[n]$ y normalmente una acústica y una condición en el contexto de modelo de lenguaje. Los arcos son 4-tuplas $a = (S[a],E[a], I[a], w[a])$. $S[a], E[a] \in N$ denotan el nodo inicial y final del arco. $I[a]$ es la palabra identificada. Por último $w[a]$ debe ser un coste asignado al arco para ser reconocido. Específicamente $w[a] = p_{\text{ac}}(a)1/\lambda P_{\text{LM}}(a)$, con probabilidad acústica $p_{\text{ac}}(a)$, probabilidad LM P_{LM}, y coste LM λ. Normalmente también proporciona la mejor pronunciación por cada arco, cuando existen múltiples pronunciaciones para una palabra $I[a]$.

Además, se definen caminos $\pi = (a_1, \cdots, a_K)$ como las secuencias de arcos conectados. Se usan los símbolos S, E, I, y w para caminos que representan las propiedades respectivas del camino completo, es decir, el nodo inicial del camino $S[\pi] = S[a_1]$, nodo final $E[\pi] = E[a_K]$, secuencia de camino etiquetado $I[\pi] = (I[a_1], \cdots, I[a_K])$, y coste total del camino $w[\pi] = \Pi^K k{=}1\ w[a_k]$.

Una alternativa equivalente de representación de lattices, llamada lattice posterior, es más conveniente en determinados casos [Zhou06]. Se definen los arcos y nodos posteriores como $P_{\text{arco}}[a]$ y $P_{\text{nodo}}[n]$ respectivamente.

$$Parc[a] = \frac{\alpha_{s[a]}.w[a].\beta_{E[a]}}{\alpha_{n_{end}}}; \qquad Pnodo[a] = \frac{\alpha_n.\beta_n}{\alpha_{n_{end}}} \tag{9}$$

con probabilidades hacia delante y hacia atrás definidas como:

$$\alpha_n = \sum_{\pi:S[\pi]=n_{start} \wedge E[n]=n} w[\pi] \quad ; \qquad \beta_n = \sum_{\pi:S[\pi]=n \wedge E[n]=n_{end}} w[\pi] \tag{10}$$

α_n y β_n pueden ser computadas adecuadamente usando la conocida recursión 'forward-backward'.

La representación de la probabilidad a posteriori de los lattices contiene cuatro campos por cada arco: $S[a]$, $E[a]$, $I[a]$, y $P_{\text{arco}}[a]$, y dos campos con cada nodo: $t[n]$, y $P_{\text{nodo}}[a]$.

Con la representación de lattices posteriores, el cómputo de la probabilidad de una secuencia en el lattice viene dada como:

$$P(*,t_s,Q,t_e,*/O) = \sum_{\substack{\pi=(a_1,\ldots a_K);\\ t[S[\pi]]=t_s \wedge t[E[\pi]]=t_e \wedge t[\pi]=Q}} \frac{P_{arc}[a_1]..P_{arc}[a_K]}{Pnode[S[a_2]]..P_{node}[S[a_K]]} \tag{11}$$

La representación de la probabilidad a posteriori de los lattices hace más fácil segmentar un arco con unidades más largas (por ejemplo, palabras) para descomponerlo en unidades más cortas (por ejemplo, caracteres) y mezclar múltiples arcos y nodos juntos. Se opta por este tipo de representación para la tarea de manipulación de lattices en la fase de búsqueda del desarrollo.

Recuperación e indexado de lattices

Los lattices se guardan con un conjunto de índices, uno por cada etiqueta de arco (fonema), por cada índice: un número de lattice *L[a]*, el estado de entrada *k[a]* de cada arco *a* a través de la probabilidad *f(k[a])* y un índice para el siguiente estado [Sproat07]. Se puede recuperar una etiqueta simple desde un conjunto de lattices representados en un cuerpo de habla, mediante el índice de dicha etiqueta. Los lattices son primero normalizados por ponderación de peso, entonces la probabilidad del conjunto de todos los caminos seguidos desde el arco hasta el estado final es 1. Después de la ponderación por peso, para un arco dado a, la probabilidad de un conjunto de caminos conteniendo este arco viene dada por:

$$p(a) = \sum_{\pi \in L:a \in \pi} p(\pi) = f(k[a])p(a/k[a]) \tag{12}$$

Ésta es la probabilidad de todos los caminos seguidos en el arco, multiplicados por la probabilidad de este mismo arco. Para un lattice l se construye una 'cuenta' C(l/L) para una etiqueta dada l usando la información guardada en el índice I(l) como sigue,

$$C(l/L) = \sum_{\pi \in L} p(\pi)C(l/\pi) = \sum_{\pi \in L}\left(p(\pi)\sum_{a \in \pi} \delta(a,l)\right) = \sum_{\pi \in L}\left(\delta(a,l) \sum_{\pi \in L:a \in \pi} p(\pi)\right) =$$
$$= \sum_{a \in I(l):L[a]=L} p(a) = \sum_{a \in I(l):L[a]=L} f(k[a])p(a/k[a]) \tag{13}$$

Donde C(l\π) es el número de veces que l es visto en el camino π y δ(a,l) es l si el arco a tiene etiqueta l y 0 si no la tiene. La recuperación puede estar limitada a que la unión de las sucesivas cuentas no sea devuelta.

La cuenta C(l/L) puede ser interpretada como una medida de la confianza basada en lattices. El uso de probabilidades a posteriori permite una factorización simple que se hace un indexado eficiente.

En el método del reconocimiento de fonemas y posterior búsqueda en lattices:

- ✓ Las búsquedas son bastante rápidas (se buscan las secuencias de fonemas en el lattice)
- ✗ No tiene problemas de OOVs: una vez realizada la indexación se puede buscar cualquier palabra porque se buscan secuencias de fonemas
- ✗ Es menos precisa que las anteriores, aunque existen formas de mejorarla

2.2.4. Spotting de palabras clave en sistemas híbridos HMM-ANN

Una técnica para generar modelos alternativos para palabras clave en un modelo de Markov oculto híbrido, es una red neuronal artificial (HMM-ANN) [Bour96]. Dada una pronunciación base para una palabra clave desde un diccionario de búsqueda, el algoritmo genera un nuevo modelo de palabra clave, que tiene en cuenta los errores sistemáticos cometidos por la red neuronal y evita estos modelos que pueden confundirse con otras palabras en el lenguaje. La nueva palabra clave mejora la tasa del modelo de detección, mientras que aumenta mínimamente el número de falsas alarmas [Joel07].

La palabra clave está representada por su cadena fonética y cada fonema en la palabra clave es modelado por un HMM. Para definir las probabilidades a posteriori de los fonemas se utilizan las probabilidades de emisión del estado HMM y para detectar la palabra clave se aplica el algoritmo de Viterbi.

2.2.4.1. Spotting de palabra clave acústica

En la búsqueda de una palabra clave acústica, dicha palabra sigue el modelo de su cadena fonética y todas las que no son claves son modeladas por un modelo conectado en paralelo al inicial.

Además, hay una transición desde el principio hasta el final en este modelo adicional para que se pueda detectar más de una palabra clave en un enunciado. El modelo adicional es un modelo genérico que engloba a todo el discurso y cumple las siguientes desigualdades:

$$p(X_{w_i}/M_{w_i}) > p\left(X_{w_i}/M_A\right) \quad (14)$$

$$p(X_A/M_A) > p\left(X_{A_i}/M_{w_i}\right) \quad (15)$$

La ecuación (14) controla la tasa de detección de palabras clave, la ecuación (15) controla el número de falsas alarmas, donde X_{w_i} y M_{w_i} son la prueba acústica y el modelo para la palabra clave W_i, respectivamente. Del mismo modo, X_A es el habla correspondiente a palabras no clave y M_A es el modelo adicional.

La palabra clave modelo es la concatenación de los Modelos de Markov ocultos (HMMs) correspondiente a los fonemas que forman la palabra clave.

2.2.4.2. Sistema Base

El sistema básico consiste en una palabra modelo clave conectada en paralelo al modelo basura, tal y como se muestra en la figura, siendo P_{KW} la probabilidad de entrada de una keyword.

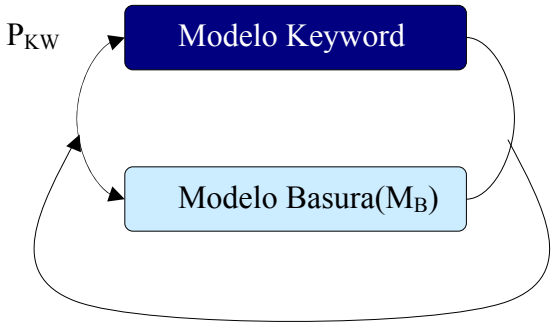


Figura 10. Arquitectura del sistema Keyword Spotting acústico

El HMM para la palabra clave es la concatenación de los HMMs de los fonemas constituyentes. La cadena fonética de la palabra clave se obtiene mediante búsqueda en un diccionario con múltiples pronunciaciones.

2.2.5. Sistema reconocedor completo usando modelos de espacio vectorial

En esta sección se describe un sistema completo de búsqueda. Se usa un método basado en modelos de espacio vectorial, para proporcionar acceso rápido a grandes volúmenes de datos multimedia, está formado por dos etapas, la recuperación de una lista corta de segmentos de audio candidatos como consulta y la búsqueda de éstos mediante índices.

El modelado en el espacio vectorial (VSM) de datos de audio es una eficiente aproximación a búsqueda de audio. Es común en recuperación de información basada en texto, y ha sido aplicada para tareas de reconocimiento y recuperación de voz, como identificación de lenguaje, recuperación por consultas de voz y recuperación por documentos hablados.

Una vista del sistema de búsqueda de audio es dado en la Figura 11. Se extraen los tokens basados en palabras desde una consulta de entrada. Se usa una medida del espacio vectorial para recuperar una lista corta de segmentos candidatos probables para cada consulta. La tarea de detección de términos acepta el conjunto de las transcripciones 1-best así como la lista de segmentos de audio como entrada, la lista de segmentos candidatos es buscada usando indexado basado en palabras, para palabras conocidas y basado en fonemas para entradas fuera de vocabulario OOV.

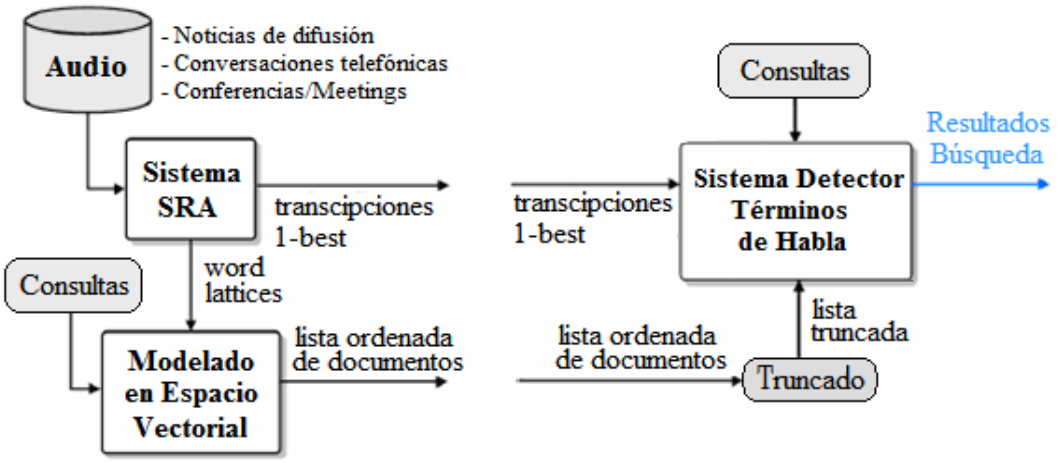


Figura 11. Sistema de búsqueda de audio. Recuperación de segmentos y posterior Detección de términos de habla

2.2.5.1. Recuperación de segmentos de audio por VSM

Vector-space modelling (VSM) aproxima la recuperación de información convirtiendo documentos desde su forma original a vectores numéricos, con frecuencia llamados documentos vector. Las consultas de entrada pueden ser convertidas similarmente en consultas vector y la relevancia de cualquier documento en el conjunto de una consulta de entrada puede ser determinada numéricamente, usando vectores espacio con métricas similares o técnicas de agrupado.

La distancia coseno es comúnmente usada para expresar la similitud entre dos vectores documento, es igual al coseno del ángulo entre dos vectores. Para dos vectores x_1 y x_2 el coseno es dado por :

$$SIM_{\cos}(x^1, x^2) = \frac{x^1 \cdot x^2}{|x^1||x^2|} \qquad (16)$$

Tras la fase de recuperación vendría la de detección de términos de habla, en la que puede ser usado cualquier sistema descrito en la sección anterior.

2.2.6. Investigaciones recientes en este campo

En los últimos años se han llevado a cabo investigaciones relacionadas con el campo de la recuperación de voz, de gran importacia e interés.

- Witbrock y Hauptmann [Witb97], presentan un sistema donde una transcripción fonética es obtenida desde la transcripción de palabras y la recuperación es realizada usando ambos índices, de palabra y de fono.
- Jones et al. [Jones96] describen un sistema que combina un sistema LVCSR y un Word-spotter de lattices de fonos para recuperar mensajes de voz y de mail video.

- Wechsler et al. [Wech98] detectan ocurrencias de características de consulta, un nuevo método que estima la probabilidad de ocurrencia, una técnica de amplia colección de reestimación probabilidades y ponderación de características largas.
- Srinivasan and Petkovic [Srin00] introducen un método para la recuperación fonética basada en la formulación probabilística de una ponderación de términos usando fonos de confusión de datos.
- Amir et al. [Amir01] usan indexado basado en grupos de fonos confundibles y una distancia de edición fonética Bayesiana para recuperación fonética de habla.
- Logan et al. [Logan02] comparan tres métodos de indexado basados en palabras, partículas sílabas, y fonemas para estudiar el problema de consultas OOV en sistemas de indexado de audio.
- Logan and Van Thong [LogVan02] dan una alternativa al problema de OOV expandiendo las palabras a consultar en frases dentro de vocabulario, mientras toman confusibilidad acústica y teniendo en cuenta las puntuaciones del modelo lingüístico.

2.3. Algoritmos de búsqueda en reconocimiento de voz

El reconocimiento de habla es en definitiva un problema de búsqueda. La acústica y los modelos de lenguaje se basan en un marco de reconocimiento de patrones estadísticos. El reconocimiento de habla, hace una búsqueda de decisión que también se conoce como decodificación. El proceso de descifrado de un reconocedor de habla consiste en encontrar una secuencia de palabras cuyos correspondientes modelos acústicos y de lenguaje concuerdan lo mejor posible con la señal de entrada [Huang01].

La búsqueda de reconocimiento de voz se realiza habitualmente con el método Viterbi o pila de descodificadores A *. Ambos algoritmos se han aplicado con éxito a varios sistemas de reconocimiento de voz. Viterbi ha sido el método preferido para casi todas las tareas de reconocimiento de voz. La pila de descodificación, por otra parte, sigue siendo una estrategia importante para descubrir los n-best y estructuras de lattices.

2.3.1. Viterbi

El algoritmo de Viterbi proporciona una solución óptima para manejo de tiempos no lineales entre modelos HMM y observaciones acústicas, el límite de detección de palabra, y la identificación de la misma en reconocimiento de habla continuo.

La búsqueda Viterbi es un algoritmo de tiempo síncrono que procesa completamente el tiempo t antes de ir al t+1. Para el tiempo t, cada estado es actualizado por la mejor puntuación (en lugar de la suma de todos sus caminos de entrada) desde todos los estados en el tiempo t-1.

Inicialización: Para todos los estados de palabra w en los que puede comenzar una sentencia

$$D(0;I(w);w) = 0$$
$$h(0;I(w);w) = null$$

Inducción: Para un tiempo t=1 a T hacer
Para todos los estados activos hacer
Transiciones entre palabra de acuerdo a:

$$D(t;s_t;w) = \min_{s_{t-1}} \left\{ d(\mathbf{x}_t, s_t \mid s_{t-1};w) + D(t-1;s_{t-1};w) \right\}$$
$$h(t;s_t;w) = h(t-1, b_{\min}(t;s_t;w);w)$$

Para todos los estados activos de palabras finales hacer
Transiciones entre palabras de acuerdo a:

$$D(t;\eta;w) = \min_{v} \left\{ \log P(w \mid v) + D(t;F(v);v) \right\}$$
$$h(t;\eta;w) = \langle v_{\min}, t \rangle :: h(t, F(v_{\min});v_{\min})$$

si $D(t;\eta;w) < D(t;I(w);w)$

$D(t;I(w);w) = D(t;\eta;w)$ y $h(t;I(w);w) = h(t;\eta;w)$

Podado: Encontrar el coste del mejor path y decidir el límite
Podar las hipótesis menos prometedoras

Terminación: Elige el mejor path a través de todos los posibles estados finales en un tiempo T
Obtiene la secuencia de palabra óptima de acuerdo a $h(t;\eta;w)$

Figura 12. Algoritmo de Viterbi

2.3.2. Decodificación de pila (A* SEARCH)

La decodificación de pila representa la mejor opción para usar búsquedas A* en lugar de búsquedas de tiempo síncrono en reconocimiento de habla continuo.

La decodificación de pila como árbol de búsqueda trata de encontrar un camino en un árbol entero que corresponde a palabras en vocabulario, nodos no terminales corresponden a sentencias incompletas, y los nodos terminales corresponden a completas sentencias. La búsqueda en árbol tiene un factor de ramificación constante, $|V|$, si se permite a cada palabra ser seguida por otra.

A diferencia de las búsquedas Viterbi, la decodificación por pila es asíncrona, se necesita un mecanismo efectivo para determinar cuando finaliza la evaluación de un fono o palabra para moverse al siguiente.

Las claves de la decodificación por pila son:

1. Encontrar una función heurística efectiva y eficiente para estimar las futuras características restantes entradas en cadena.
2. Determinar cuando extiendes la búsqueda para el siguiente fono/palabra.

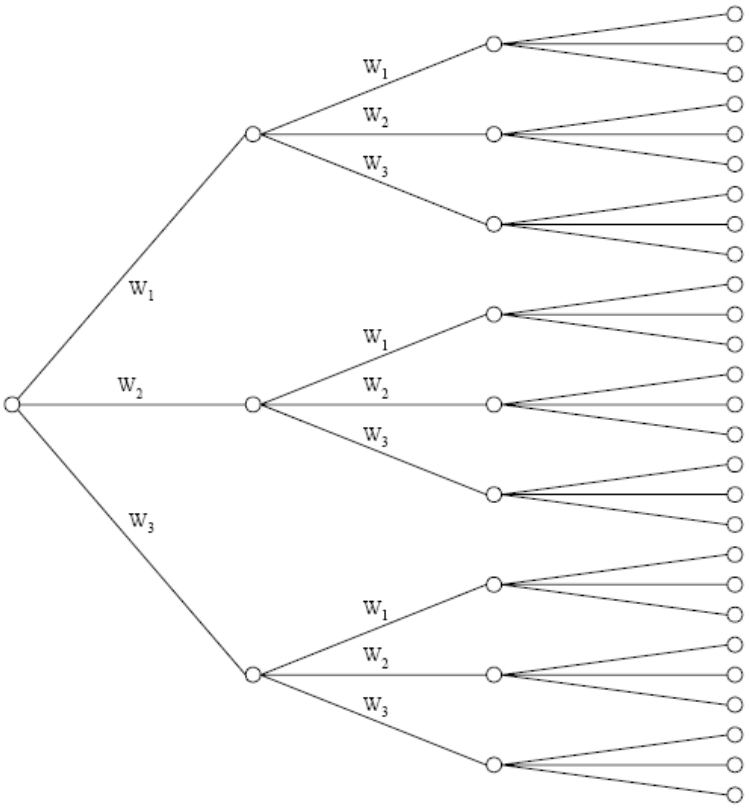


Figura 13. Árbol de búsqueda por decodificación de pila para un vocabulario de tamaño 3 [Huang01].

2.3.3. N-best y estrategias de búsqueda multi-paso

Idealmente, un algoritmo de búsqueda debe considerar todas las posibles hipótesis basadas en un marco de trabajo unificado que integra todas las fuentes de conocimiento. Estas fuentes pueden ser modelos de lenguaje, modelos acústicos y modelos de pronunciación léxica.

Con técnicas de desarrollo más potentes, la complejidad de modelos tiende a crecer drásticamente. Una posible solución a este inconveniente es realizar una búsqueda multi pasos progresiva. En el paso inicial, se usan fuentes de conocimiento más discriminativas y asequibles computacionalmente, para reducir el número de hipótesis. En los pasos siguientes, se examinan conjuntos de hipótesis progresivamente reducidas y se usan fuentes de conocimiento más potentes y caras, hasta que la solución óptima es encontrada.

La estrategia más directa es la llamada búsqueda n-best. La idea es usar las fuentes de conocimiento para producir una lista de las n mejores secuencias de palabra en un tiempo razonable. Entonces, estas n hipótesis son puntuadas usando modelos más detallados para obtener una secuencia de palabra más probable. La idea de la lista de n-best puede ser extendida para crear una representación de las hipótesis más compacta – llamada lattice de palabra o grafo. Un lattice de palabra es un camino más eficiente para representar hipótesis alternativas. N-best o búsqueda de lattices es usada para muchos de los sistemas de reconocimiento de habla, será el usado en el desarrollo de la implementación.

2.3.4. Listas *N*-Best y Lattices de palabra

El marco de trabajo en búsquedas N-best es efectivo sólo para n de los órdenes de decenas o centenas. Si la lista corta de n-best que es generada mediante el uso de modelos poco óptimos no incluye la secuencia de palabra correcta, las fases de puntuación sucesivas no tienen oportunidad para generar una respuesta correcta.

Los lattices de palabra están compuestos por hipótesis palabra. Cada una de estas hipótesis esta asociada con una puntuación y un intervalo de tiempo explícito.

Los grafos de palabra, por otro lado, se parecen a máquinas de estado finitas, en las cuales los arcos son etiquetados con palabras.

La Figura 14. ilustra el marco de trabajo de la búsqueda general n-best/lattice. Aquellas fuentes de conocimiento que proporcionan más restricciones, con el menor coste, se usan primero para generar las listas n-best o lattices de palabra, a continuación pasan al módulo de puntuación, que usa fuentes de conocimiento para seleccionar el camino óptimo [Gauv95].

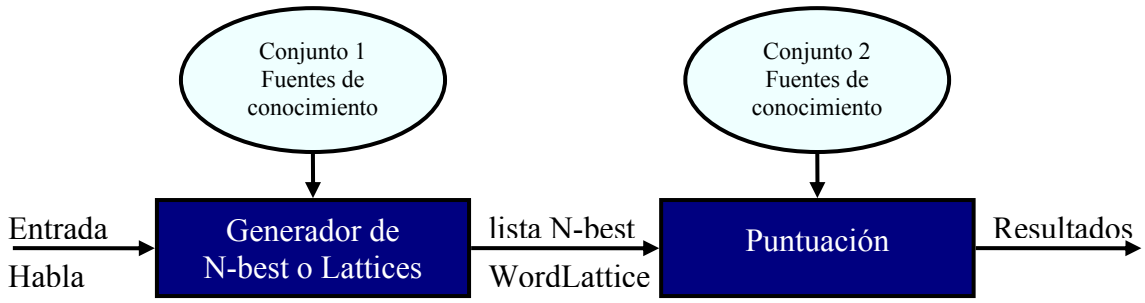


Figura 14. Marco de trabajo de una búsqueda N-best/lattice

2.3.5. Búsqueda Forward-Backward

La habilidad para predecir cómo de buena es una búsqueda en el futuro para las porciones restantes de habla ayudan a reducir el esfuerzo de búsqueda significablemente. La estrategia de búsqueda en un paso que en general tiene muy pocas oportunidades para predecir el coste de una porción que no ha sido vista. Esta dificultad puede ser aliviada por estrategias de búsqueda multipaso. En sucesivas fases la búsqueda debe ser capaz de proporcionar buenas estimaciones de los caminos restantes, ya que el sonido completo ha sido examinado por pasos más tempranos.

En la búsqueda forward-backward la idea es primero realizar una búsqueda forward, durante la cual puntuaciones parciales forward α para cada estado son guardadas.

Entonces se realiza el segundo paso la búsqueda backward, este segundo paso comienza tomando el marco final de voz y buscando su camino hacia atrás hasta conseguir encontrar el comienzo del habla. Durante la búsqueda backward, las puntuaciones parciales forward α pueden ser usadas como una estimación exacta de la función heurística o la puntuación rápida de la combinación para los caminos restantes.

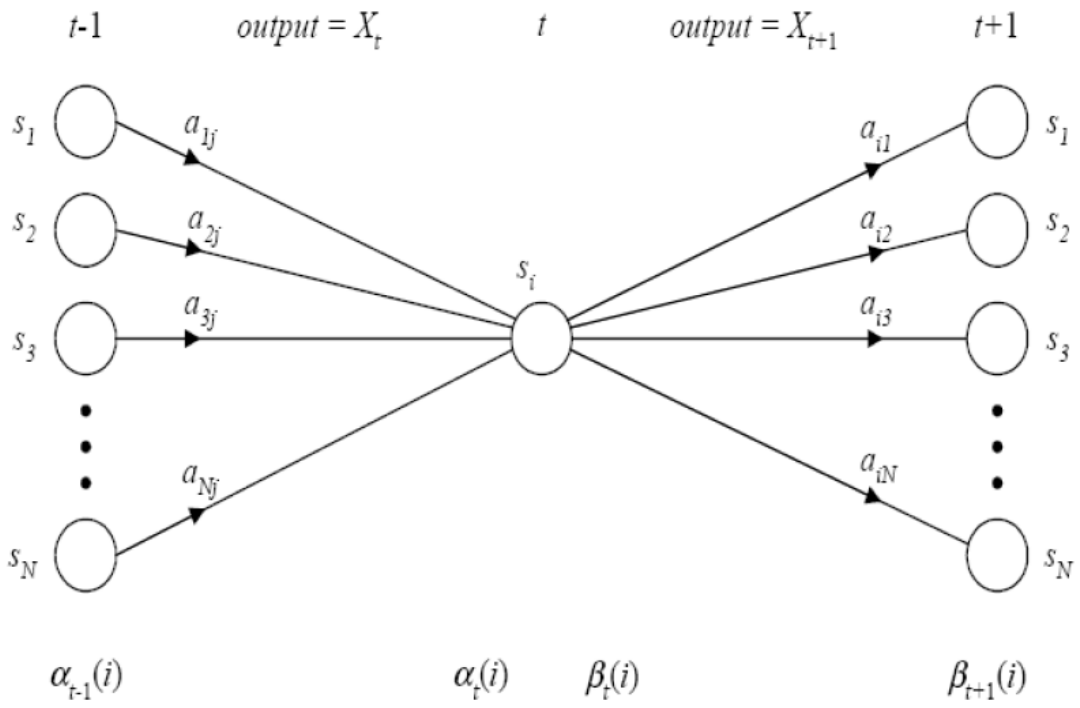


Figura 15. Algoritmo Forward-Backward

La búsqueda forward debe ser muy rápida y es generalmente una búsqueda Viterbi en tiempo síncrono. Como en las estrategias de búsqueda multi paso, modelos de lenguaje y acústica simplificados son con frecuencia usados en la búsqueda forward. Para búsquedas backward, sea en tiempo síncrono o asíncrono, pueden ser empleadas búsquedas A* para encontrar las n mejores secuencias de palabra.

3. Diseño

3.1. Medios disponibles

3.1.1. Base de datos

La base de datos usada para llevar a cabo el conjunto de experimentos del presente proyecto ha sido el corpus multi-lengua C-ORAL-ROM que comprende cuatro lenguas románicas: Italiano, Francés, Portugués y Español. Sólo se usa el sub-corpus de Español, que contiene un total de 300.000 palabras. Desde un punto de vista socio lingüístico, los hablantes son caracterizados por su edad, lugar de nacimiento, nivel de educación y profesión. Desde un punto de vista textual el corpus es divido en las partes mostradas en la siguiente tabla:

<table>
<tr><th colspan="4">Informal
150.000 palabras</th><th>Formal
150.000 palabras</th></tr>
<tr><td colspan="2">Familiar
113.000</td><td colspan="2">Pública
37.000</td><td>Formal en contexto natural
65.000</td></tr>
<tr><td rowspan="2">Monólgos
33.000</td><td rowspan="2">Diálogos/
Conversaciones
80.000</td><td rowspan="2">Monólogos
6.000</td><td rowspan="2">Diálogos/
Conversaciones
31.000</td><td>Formal en medios
60.000</td></tr>
<tr><td>Conversaciones telefónicas
25.000</td></tr>
</table>

Tabla 4. División del cuerpo de datos en los diferentes tipos de habla

Se muestra que la principal división es balanceada entre habla formal e informal. Para habla informal una división es considerada entre habla en contexto familiar y en contexto público. El primer grupo es clasificado en monólogos, diálogos y conversaciones con tres o más personas. El segundo grupo es clasificado similarmente. Dentro de habla formal, existe una división entre contexto natural y habla en los medios. Se incluyen discursos políticos, debates, enseñanza, conferencias, habla en contexto de negocios y en contextos legales, etc.

Dentro de la división de habla en los medios (también referida como emisión de noticias) se incluyen noticias, deportes, entrevistas, meteorología, ciencia y reportajes. Las conversaciones telefónicas, aunque inicialmente son consideradas bajo habla formal en C-ORAL-ROM, tienen características peculiares y son más similares a habla informal que a formal. Por esta razón se consideran bajo la categoría de habla informal en una subdivisión de sí misma.

Estas divisiones y subdivisiones de C-ORAL-ROM permitieran comparar la bondad del sistema mediante el uso de diferentes tipos de habla espontánea.

C-ORAL-ROM contiene en total 184 grabaciones sobre 40 horas de habla. Hay básicamente tres tipos de grabaciones dependiendo de la duración, 7-10 minutos, 15 minutos, y 30 minutos. Estas grabaciones son demasiado grandes para ser procesadas automáticamente. Por esta razón, se extrae cada sonido de habla (entre pausas) en un archivo separado usando un manual de segmentación existente.

Se selecciona una parte de estos segmentos extraídos a partir de técnicas de segmentación, para la realización de las pruebas del sistema desarrollado.

3.1.1.1. Transcripción fonética

C-ORAL-ROM no incluye transcripciones fonéticas, sólo ortográficas. Por esta razón, la trascripción fonológica fue generada desde una ortográfica, haciendo uso de una trascripción simple fonológica basada en reglas. Éste transcriptor usa un conjunto mínimo de fonemas para Español (23 fonemas). Obviamente, este simple transcriptor no permite obtener una trascripción correcta en todos los casos, pero se considera que la precisión lograda es bastante buena.

3.1.1.2. Entrenamiento de HMMs para decodificación acústico-fonética

Los Hidden Harkov Models usados para los resultados de descodificación de C-ORAL-ROM fueron entrenados en el corpus de ALBAYZIN usando el Hidden Harkov Model Toolkit (HTK) software.

El conjunto de fonemas usado en todos los experimentos es el mínimo conjunto de 23 fonemas en Español. Se consideran modelos para silencios iniciales, intermedios y finales. Se usan modelos independientes del contexto.

3.1.2. Hardware

El hardware empleado en el desarrollo de este proyecto es un ordenador con procesador Intel Pentium IV. Se dispone además de una red interna que incluye todos los ordenadores del grupo de trabajo, tanto los de uso personal, como los de pruebas. Estos medios fueron suministrados por el grupo ATVS de la Universidad Autónoma de Madrid (UAM).

Se usaron varios nodos de cómputo para la fase de búsqueda dentro de la red interna del grupo. Todos estos nodos han sido ejecutados sobre el mismo sistema operativo, la principal distribución Linux del proyecto Debian. La plataforma sobre la que hace soporte es `i386` – x86-32.

Los equipos son Intel Xeon CPU's (cada proceso corre en una sola CPU). Con procesadores de 32-bits corriendo a 3.0GHz.

3.1.3. Software

3.1.3.1. Software fase de indexado

HTK (*Hidden Markov Model Toolkit*)

El Hidden Markov Model Toolkit (HTK) es una herramienta que consiste en un toolkit portable usado para la construcción y manipulación de modelos de Markov. Aunque en un momento inicial fue diseñada principalmente para construir modelos basados en el procesamiento de señales de habla HTK, posteriormente se han encontrado muchas otras aplicaciones como la síntesis de voz o secuencias de ADN.

HTK está compuesto por un conjunto de librerías y herramientas desarrolladas en C. Las sofisticadas herramientas facilitan el análisis de la voz, el entrenamiento de los HMM, el test y la extracción de resultados. El software soporta la creación de HMM con distribuciones continuas de mezclas de Gaussinas o por medio de distribuciones discretas pudiendo crear de esta forma complejos sistemas de HMM.

HTK fue desarrollado originalmente por el departamento de Ingeniería de la Universidad de Cambridge, conocido como el grupo "the Speech Vision and Robotics". Aquí se ha utilizado para construir grandes sistemas de reconocimiento de habla.

Contiene diferentes algoritmos de estimación de parámetros como el algoritmo Baum-Welch o el algoritmo Viterbi.

La principal herramienta usada en el proceso de indexado, para la decodificación fonética es Hvite un reconocedor de palabras Viterbi de propósito general, que pone un archivo de habla en una red HMM y da como salida una trascripción para cada uno. Cuando se realizan reconocimientos N-best, un lattice a nivel de palabra contiene múltiples hipótesis que pueden también ser producidas [Young02].

Cada lattice de nivel de palabra o archivo de nivel es leído, y cuando es expandido usa el diccionario suministrado para crear un modelo basado en la red. Esto permite una red arbitraria de estado de palabras finitas y un alineamiento simple que ha de ser especificado.

Esta expansión puede ser usada para crear:

- un contexto independiente
- una palabra interna en contexto dependiente
- y un cruce de palabras en redes de contexto dependiente

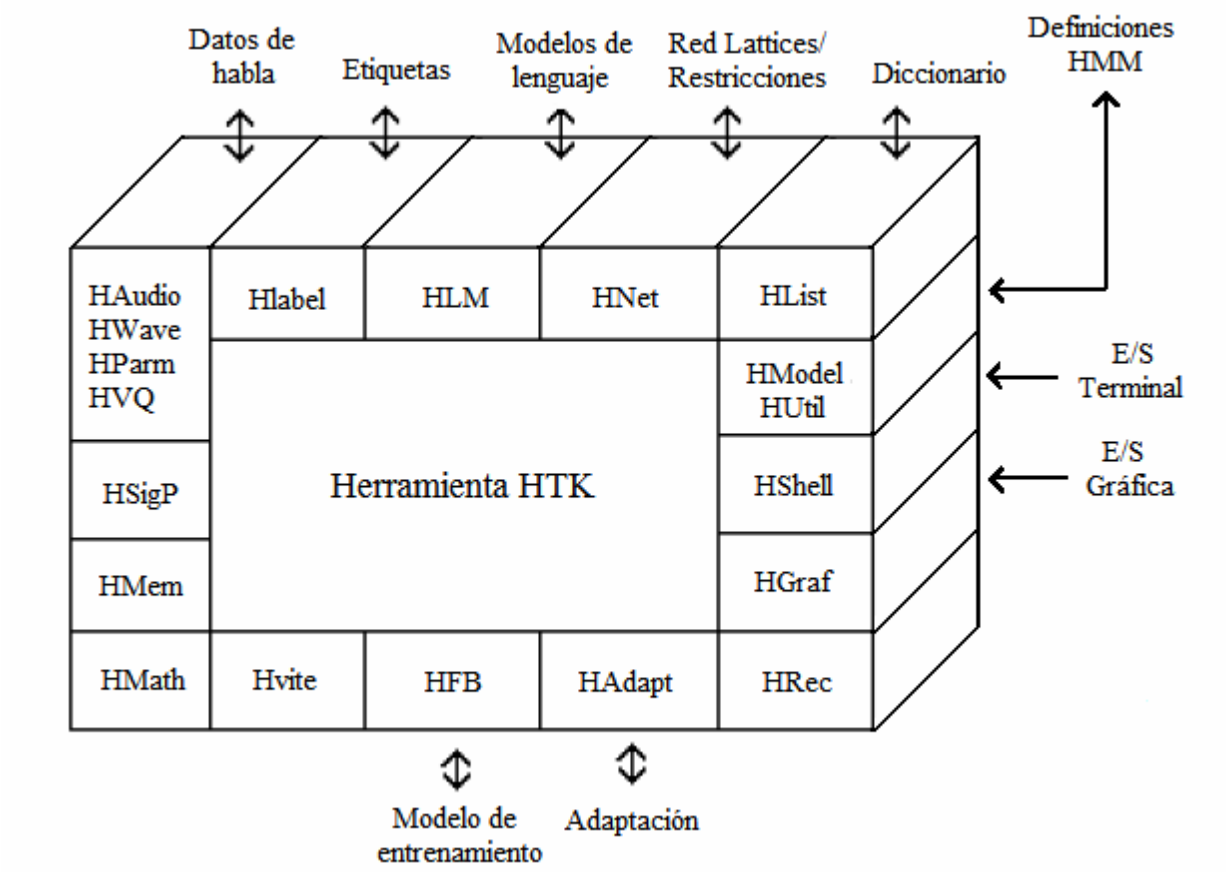


Figura 16. Arquitectura del software HTK

El camino en el que la expansión es formada es determinado automáticamente desde el diccionario HMMList. Cuando todos los niveles aparecidos en el diccionario están definidos en el HMMList ninguna expansión de nombres de modelos es formada. De otro modo si todos los niveles en el diccionario pueden ser satisfechos por modelos dependientes sólo con contextos de palabras internas, estos serán usados también como contexto de expansión de cruce de palabras. Esas omisiones pueden ser sobre leídas por parámetros de configuración Hnet.

Hvite soporta parámetros compartidos y apropiadamente precomputa salidas de probabilidades.

SPHINX

SPHINX es considerado uno de los mejores sistemas de reconocimiento existentes en la actualidad. Fue desarrollado en la Universidad de Carnegie Mellon y al igual que HTK, se basa en la construcción de Modelos Ocultos de Markov. Los componentes de SPHINX son el *Sphinxtrain*, para entrenamiento de los modelos, y el SPHINX *decoder* para reconocimiento.

El *Sphinxtrain* esta compuesto por un conjunto de programas que han sido compilados para dos tipos de sistemas: Linux y alpha. Genera modelos acústicos discretos, semicontinuos o continuos (HMM con topología left to right) que pueden tener desde 1 hasta 24 gaussianas en cada estado.
El SPHINX *decoder* contiene algoritmos de programación dinámica como Baum-welch o Viterbi.

DARPA (Defense Advanced Research Projects Agency) financia extensamente el proyecto para estimular la creación de herramientas de discurso y el uso de las mismas, en el reconocimiento de voz, así como en áreas relacionadas incluyendo sistemas de diálogo y síntesis de discurso.

Sphinx ha sido apoyado durante muchos años por la financiación del DARPA y los motores del reconocimiento que se lanzan son los que el grupo utilizó para varios de los proyectos de DARPA y sus evaluaciones respectivas.

Procesamiento de las señales de audio con Sphinx

Una señal de audio a una frecuencia determinada es segmentada en frames. Cada frame con un tiempo en milisegundos y unas muestras de habla dadas, es multiplicado por una función de ventana de *Hamming* que se aplica sucesivamente cada 'x' milisegundos. Consecutivos fotogramas solapan cada 'x' milisegundos o muestras de habla. De éstas muestras, se calculan los coeficientes LPC para finalmente obtener los coeficientes cepstrales de derivada LPC.

Las principales necesidades de SPHINX a la hora de realizar un reconocimiento son:

1. El diccionario de pronunciación.
2. El diccionario con sonidos de relleno.
3. Los modelos acústicos.
4. El modelo de lenguaje.
5. Los datos de prueba.

3.1.3.2. Software de manipulación de lattices

SRILM es un toolkit para construcción y aplicación de modelos de lenguaje estadísticos principalmente para uso en reconocimiento de habla, etiquetado estadístico y segmentación. Comenzó a ser desarrollado por el SRI Speech Technology and Research Laboratory en 1995. El toolkit tiene también grandes beneficios de su uso y mejoras con los workshops de la universidad Johns Hopkins University/CLSP en 1996, 1997, y 2002.

El principal tipo de modelo de lenguaje soportado por SRILM es el modelo de N-gramas. SRILM consiste en los siguientes componentes:

- Un conjunto de librerías de clases C++ implementando modelos de lenguaje, soportando estructuras de datos y funciones de utilidad miscelánea.
- Un conjunto de programas ejecutables construidos sobre estas librerías para realizar tareas de entrenamiento de modelos y testing de los mismos en bases de datos, etiquetado o segmentación de texto, etc.
- Una colección de scripts misceláneos facilitando las tareas menos relacionadas. SRILM se ejecuta sobre plataformas UNIX o Windows, en nuestro caso UNIX ya que es el sistema operativo sobre el que se está desarrollando el proyecto.

3.1.3.3. Software para el desarrollo del sistema detector

El desarrollo software del sistema detector ha sido implementado con lenguaje de programación C sobre el entorno de desarrollo KDevelop, el mismo nombre alude a su perfil, KDE Development Environment (Entorno de Desarrollo para KDE), está integrado para sistemas Linux y otros sistemas Unix, y orientado al uso bajo el entorno gráfico KDE.

3.1.3.4. Software para el desarrollo del sistema evaluador

Los lenguajes usados para el desarrollo del sistema evaluador han sido Bash y Perl debido a la necesidad de una continua gestión de ficheros.

Bash es un shell de Unix (interprete de órdenes) escrito para el proyecto GNU. Su nombre es un acrónimo de bourne-again shell, siendo Bourne uno de los primeros intérpretes importantes de Unix, escrito por Stephen Bourne. La sintaxis de órdenes de bash es un superconjunto de la sintaxis del intérprete Bourne, incluye ideas tomadas desde el Korn Shell (ksh) y el C Shell (csh), como la edición de la línea de órdenes, la pila de directorios, etc.

Perl es un lenguaje de programación diseñado por Larry Wall que toma características de C, del lenguaje interpretado shell (sh), AWK, sed, Lisp y, en un grado inferior, de muchos otros lenguajes de programación. Es un lenguaje imperativo, con variables, expresiones, asignaciones, bloques de código delimitados por llaves, estructuras de control y subrutinas.

Estructuralmente, Perl está basado en un estilo de bloques como los del C o AWK, y fue ampliamente adoptado por su destreza en el procesado de texto y no tener ninguna de las limitaciones de los otros lenguajes de script. Se previó que fuera práctico (facilidad de uso, eficiente, completo) en lugar de estético (pequeño, elegante, mínimo). Sus principales características son su facilidad de uso, su soporte tanto para la programación estructurada como para la orientada a objetos o la funcional, tiene incorporado un poderoso sistema de procesamiento de texto y una enorme colección de módulos disponibles.

3.2. Diseño

En un sistema de detección de términos de voz se pueden diferenciar dos fases clave. En primer lugar, el indexado de audio, una fase offline, que convierte la señal de entrada de habla en un índice de búsqueda. Y en segundo lugar, la recuperación de términos, posiblemente una fase online, cuyo resultado será una lista completa de todas las ocurrencias detectadas por cada término de búsqueda consultado.

En general un sistema reconocedor está formado por cuatro componentes: el sistema voz-a-texto, el indexador, el detector y el decisor. El sistema voz-a-texto procesa ficheros de audio dando como salida lattices de palabra y transcripciones de fonemas. El indexador toma éstas como entradas, y crea un índice que contiene una lista precomputada de grabaciones de detección, para cada palabra en la léxica voz-a-texto. El detector por su parte carga dicho índice y procesa una lista de términos de búsqueda, generando una lista ordenada de puntuaciones y detección de términos. Finalmente el decisor coge la lista de detecciones candidatas y un parámetro de coste β_s y fija un límite de puntuación por término para hacer las decisiones sí/no.

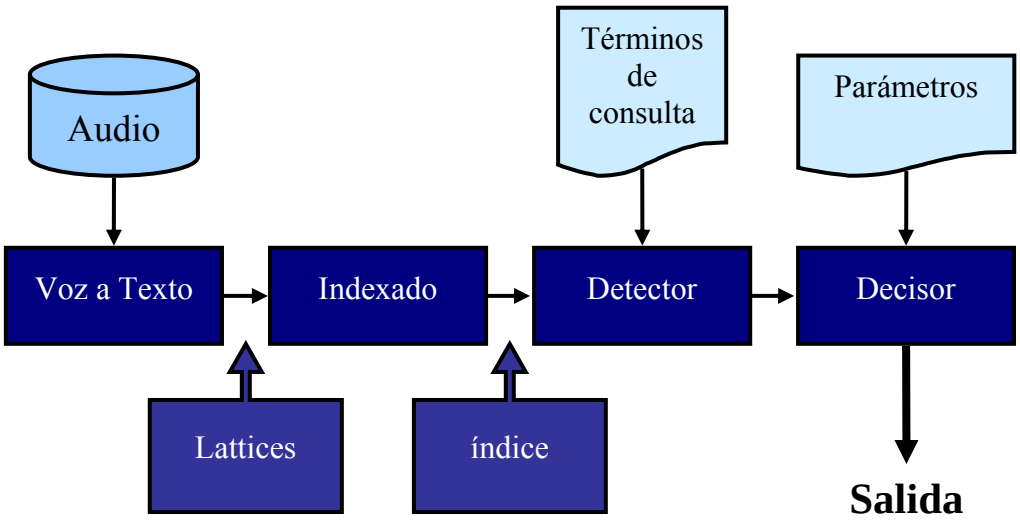


Figura 17. Esquema general detector de términos de voz

3.2.1. Descripción del sistema

En esta sección se describe la estructura del sistema global y se da un detalle de las técnicas usadas en el desarrollo. El sistema general se divide en 3 principales fases.

La primera, consiste en un sistema voz-a-texto que convierte el audio de voz en representación de lattices junto con su información de tiempo. La segunda, indexa la representación para una eficiente recuperación. Y la tercera, fase de búsqueda, que tras la consulta de un determinado término, produce una salida con las coincidencias para éste y sus probabilidades. Dicha salida será evaluada y verificada llegando así a la obtención de los resultados.

Un esquema representativo de la arquitectura es el siguiente:

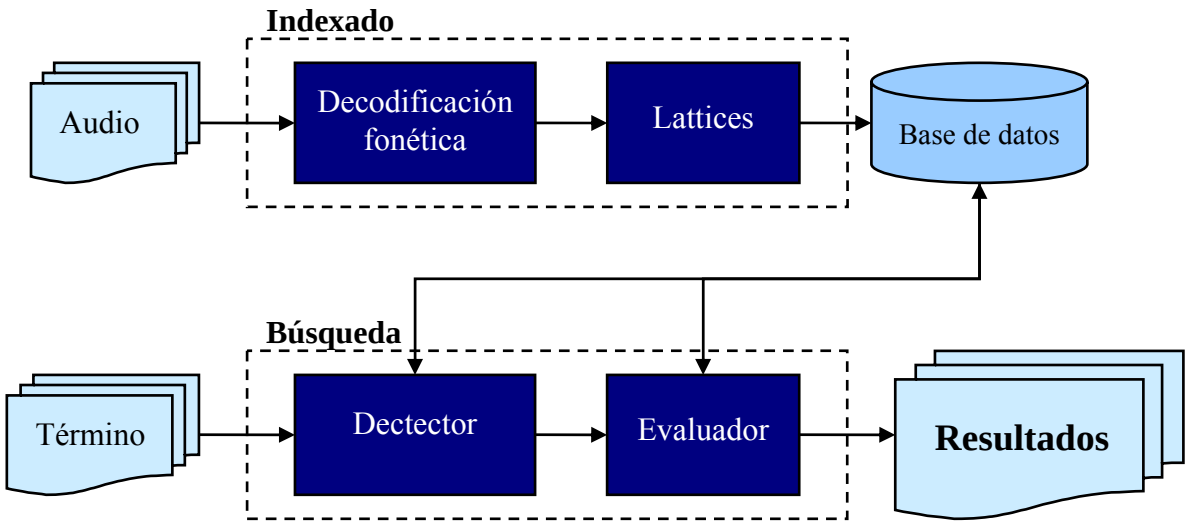


Figura 18. Arquitectura del sistema reconocedor

El sistema "voz-a-texto" procesa ficheros de audio, grabaciones de voz. Automáticamente segmenta el cuerpo de audio en fragmentos y produce para cada término que aparece en la grabación su correspondiente transcripción, hallada mediante la selección del mejor camino a través de los lattices que se han obtenido tras el proceso de voz a texto.

Los arcos de los lattices son anotados sus probabilidades del modelo de lenguaje desde el paso final de la decodificación. Se entrenan distintos sistemas, así como modelos acústicos y de lenguaje.

El componente indexador toma éstos como entrada y precomputa una lista de candidatos de detección, grabados para cada palabra individual en el sistema "voz-a-texto". Esto asigna a cada grabación una estimación de la probabilidad a posteriori de que la palabra objetivo aparezca en el audio cerca del tiempo específico. Los candidatos a detección pueden proceder de cualquier lattice, no sólo de los mejores caminos. Las listas son indexadas por palabra y guardadas junto con las transcripciones fonéticas sin indexar.

De este modo, se tienen finalmente un conjunto de lattices correspondientes a cada fragmento del audio que ha sido segmentado, junto con su respectivo fichero de vocabulario, en el que aparece cada término de la grabación con su correspondiente transcripción fonética; y el fichero de tiempos con el inicio y la duración de cada palabra pronunciada.

El detector es básicamente, un algoritmo de búsqueda que carga los índices y procesa el término de la lista a ser consultado. Tras su ejecución por cada término de búsqueda, genera una lista ordenada con las puntuaciones de las mejores detecciones obtenidas.

El decisor es el encargado de filtrar la lista de detecciones para cada término de búsqueda. Recibe como argumento, un parámetro *N* cuyo valor fija un límite de puntuación a partir del cual se toma la decisión de aceptación o rechazo del candidato, decisión SI/NO.

El parámetro *N* define las N-best probabilidades, las más altas de la lista completa de detecciones candidatas. A mayor *N*, mayor probabilidad de que se de la ocurrencia exacta entre el término consultado y la detección considerada verdadera, pero sin embargo aumenta en gran medida el número de falsas alarmas, que son las detecciones consideradas verdaderas en la fase de decisión, que realmente no lo son.

Por último el evaluador, procesa los resultados obtenidos del decisor, y comprueba el porcentaje de aciertos, para ello se basa en los ficheros de referencia que contienen la transcripción de las grabaciones incluyendo los tiempos de inicio y duración de cada uno de los términos que aparecen en el audio.

3.2.2. Sistema Voz a Texto

La funcionalidad básica del sistema Voz-a-Texto consiste en la lectura de un fichero de audio y la producción de un lattice a partir del mismo. Los lattices a continuación, son modificados y guardados para la posterior búsqueda en los subsistemas de detección.

El sistema primero segmenta el audio en distintos fragmentos, usando un algoritmo de segmentación. Después se procede a la formación de las características de análisis. Posteriormente el audio es dividido en frames solapados, cada 25 mseg de duración, con una tasa de 100 frames por segundo. Cada frame es enventanado con una función Hamming y analizado en la banda de frecuencias. Coeficientes cepstrales PLP (Perceptual Linear Prediction) son computados para cada frame, 14 coeficientes por cada uno, con la energía normalizada.

El paso final de adaptación a la decodificación produce un lattice de fonema para cada fragmento de grabación, conteniendo las mejores hipótesis candidatas, anotadas con sus probabilidades acústicas y de modelo de lenguaje. En el post procesado, el sistema computa la probabilidad a posteriori de cada arco en el lattice desde las probabilidades acústicas y de modelo de lenguaje, para usarlo más tarde en la detección de las palabras.

La mejor salida de cada palabra es usada junto con la léxica de decodificación para encontrar la mejor transcripción fonética, que más tarde será usada en el sistema detector.

3.2.3. Fase de Indexado

Tras el proceso de decodificación, se persigue el indexado de la información que facilite la búsqueda de términos en el sistema detector, la opción elegida para dicha tarea es el uso de lattices hipotéticos con los que se obtienen índices aptos para búsqueda.
Para la fase de generación de lattices, existen 5 opciones: palabra, carácter, sílaba tonal, sílaba atonal y fonema. En nuestro caso son generados lattices de fonema.

Éstos lattices contienen toda la información fonológica, así como la información de tiempos de los archivos de audio procesados.

Tras el reconocimiento, el sistema procesa toda la colección de lattices y precomputa un conjunto de candidatos de detección para cada fonema f_1, f_2, …f_L dentro del sistema voz-a-texto.

Para cada fonema en un determinado lattice, el sistema estima su probabilidad a posteriori, que será la fracción de la probabilidad total de lattice que pasa a través del mismo.

Son entonces grupos de fonemas que ocupan aproximadamente el mismo intervalo de tiempo. Con la suma de sus probabilidades a posteriori se consigue una representación simple para el fonema.

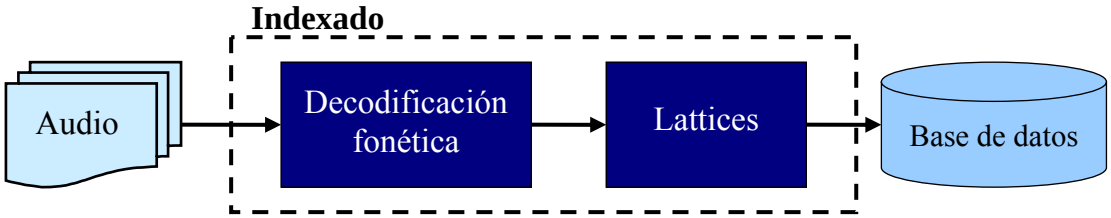


Figura 19. Fase de indexado del sistema reconocedor

3.2.3.1. Búsqueda basada en Lattices

Trabajos recientes de investigación de documentos de habla sugieren que es adecuado tomar las mejores salidas simples de RAV, y formar resultados de texto sobre esas salidas. Esto es bastante razonable para la tarea de recuperación de emisión de noticias, donde las tasas de error son relativamente bajas, y las historias son amplias como para contener redundancias. Pero esto no es del todo razonable, si una de las tareas a recuperar se perfila en otro dominio de habla, donde la tasa de error puede más baja, tal sería la situación con teleconferencias, donde una de las tareas es encontrar sí y cuándo un sonido pertenece a una determinada frase.

Un procedimiento de indexado para sonidos de habla, recuperando estas palabras en lattices, es preferible al simple 'mejor texto exacto'.

Se ha demostrado que este procedimiento puede incrementar las mejores puntuaciones obtenidas en recuperaciones sobre tareas con tasa de error (Word Error Rate) pobres y bajas redundancias. La representación es flexible de manera que podrían representarse ambos, lattices de palabra, y lattices de fonema.

La elección de lattices de fonema para llevar a cabo la fase de búsqueda se toma por las ventajas que ofrece este tipo de representación, por un lado se deja a un lado el problema de los términos out-of-vocabulary (OOV) que provocarían los lattices de palabra, por otro se reduce la complejidad de los diccionarios, en definitiva la disminución exponencial del tamaño de los diccionarios con la representación de lattices de fonema en lugar de lattices de palabra, un orden de 10^2 , pasando de 23 fonemas a 2300 palabras que como mínimo ha de tener un diccionario medianamente aceptable.

3.2.4. Fase de Búsqueda

La segunda parte del sistema está formada por el detector que es el algoritmo de búsqueda en lattices, y el sistema evaluador.

Después de la fase de transición de voz a texto, y el indexado, el sistema esta listo para procesar los términos de búsqueda. Los términos serán especificados sólo por su representación ortográfica. Idealmente, un término tendría una interpretación o significado específico y simple.
Los términos serán presentados con la ortografía de la lengua nativa: Castellano.

El módulo de detección produce una lista de candidatos puntuados en respuesta a la búsqueda de las consultas introducidas en él.

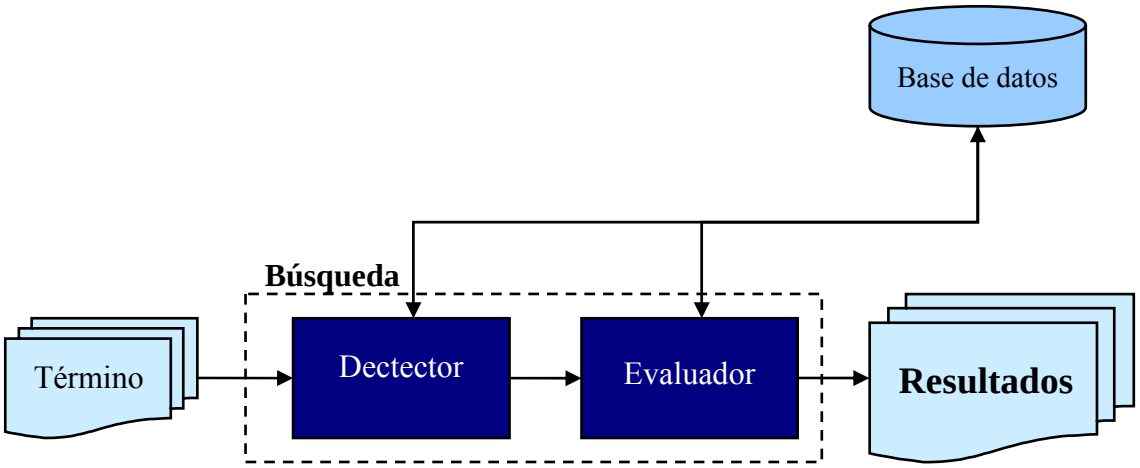


Figura 20. Fase de búsqueda del sistema detector

Es necesaria una manipulación previa de los lattices adquiridos mediante la fase de indexado, para facilitar el procesado de los mismos en la fase de búsqueda.

3.2.4.1. Manipulación de lattices

Una vez obtenida la transcripción de los archivos de audio en formato lattice, se procede a la manipulación de los mismos con la ayuda de la herramienta Lattice-Tool de SRILM [Srilm]. Esta herramienta realiza operaciones sobre lattices palabra o fonema en formato pfsg o HTK Standar Lattices Format (SLF).

Lo que se consigue con el uso de dicha herramienta es la producción de unos índices aptos para búsqueda.
Uno de los formatos que maneja el Lattice-Tool es ‘pfsg’, los lattices de partida no están en éste formato, sino ‘sphinx’, por lo que mediante un software ya implementado, proporcionado por el grupo de investigación ATVS, son convertidos a formato ‘pfsg’.

Entre las opciones del Lattice-Tool se incluyen reducción de tamaño, eliminación de nodos nulos, asignación de costes mediante modelos de lenguaje, computación de error de lattices de palabra y decodificación de las mejores hipótesis.
En un conjunto de lattices, cada uno de ellos es procesado por turnos, pudiendo realizar múltiples operaciones sobre los mismos.

Tras un estudio de las distintas opciones de manipulación que ofrece esta herramienta, se eligió la más conveniente teniendo en cuenta la actividad que se iba a llevar a cabo sobre la salida.
Ésta opción es **–write-posteriors** cuya tarea consiste en el cómputo de los nodos, así como las probabilidades a posteriori de los lattices, mediante el uso del algoritmo forward-backward, y cuya salida es el lattice posterior en formato *wlat*.

3.2.4.2. Sistema detector

El detector, es básicamente un algoritmo de búsqueda en lattices. Como se ha explicado anteriormente, cuando un término de búsqueda es presentado en el sistema, el término es primero transcrito a su representación fonética usando un diccionario de pronunciación. Si cualquiera de los términos no está en el diccionario, reglas "letra-a-sonido" son usadas para estimar las pronunciaciones fonéticas correspondientes.

Una vez hallada la transcripción del término a buscar, ésta se introducirá como parámetro de entrada en el sistema, junto con el lattice dónde ha de ser encontrada.

La ventaja de audio indexado basado en fonemas, comparado con aproximaciones basadas en palabras, es que las palabras no incluidas en el vocabulario de la transcripción voz-a-texto pueden también ser detectadas.

La tarea de la detección de términos de voz definida anteriormente puntúa sistemas basados en una decisión binaria de cuales de las detecciones candidatas son ciertas y cuales no.

3.2.4.2.1. Salida del detector

Por cada término de búsqueda introducido en el sistema son encontradas todas las ocurrencias de este término en el test, se devuelven como salida en un fichero, en el que por cada ocurrencia encontrada del término, se incluyen:

- la localización del término en el audio grabado en segundos
- la secuencias de índices fonéticos que coinciden con los del término consultado, junto con la probabilidad de ocurrencia de dicha secuencia

La salida del sistema pasará a ser evaluada, siendo considerada correcta si el término y su tiempo de ocurrencia aparece en la trascripción con una correspondencia exacta.
El propio sistema detector, incorpora el decidor, teniendo comos parámetro de entrada el valor N, que determina el número de las N mejores secuencias coincidentes que se quieren obtener como resultado. La salida de estas N mejores, corresponderán a las N probabilidades más altas.

3.2.4.3. Sistema evaluador

Dado el resultado de la secuencia de fonemas buscada, y la colección de secuencias de fonemas indexados que están guardados en la base de datos, la tarea es comparar el objetivo y las secuencias indexadas y emitir ocurrencias donde una coincidencia sea detectada.

El evaluador recorre cada uno de los ficheros de salida, pertenecientes a los distintos términos de búsqueda obtenidos en la etapa previa, y por comparación con el fichero de tiempos real de los audios grabados, toma la decisión binaria si/no determinando la detección como válida o inválida.

Una detección del sistema es considerada correcta si una coincidencia exacta ortográfica del término aparece en la transcripción de referencia con un margen de 0.5 segundos.

La salida de ésta etapa devuelve un fichero en el que se incluyen:

- cada uno de los términos de búsqueda encontrados
- una decisión binaria sobre si la detección es correcta
- porcentaje total de aciertos en el número total de términos a buscar

4. Desarrollo

La implementación comprende la fase de búsqueda del sistema global, compuesta a su vez por la fase de manipulación de lattices, la de detección y por último la de evaluación. El siguiente esquema ofrece una visión general:

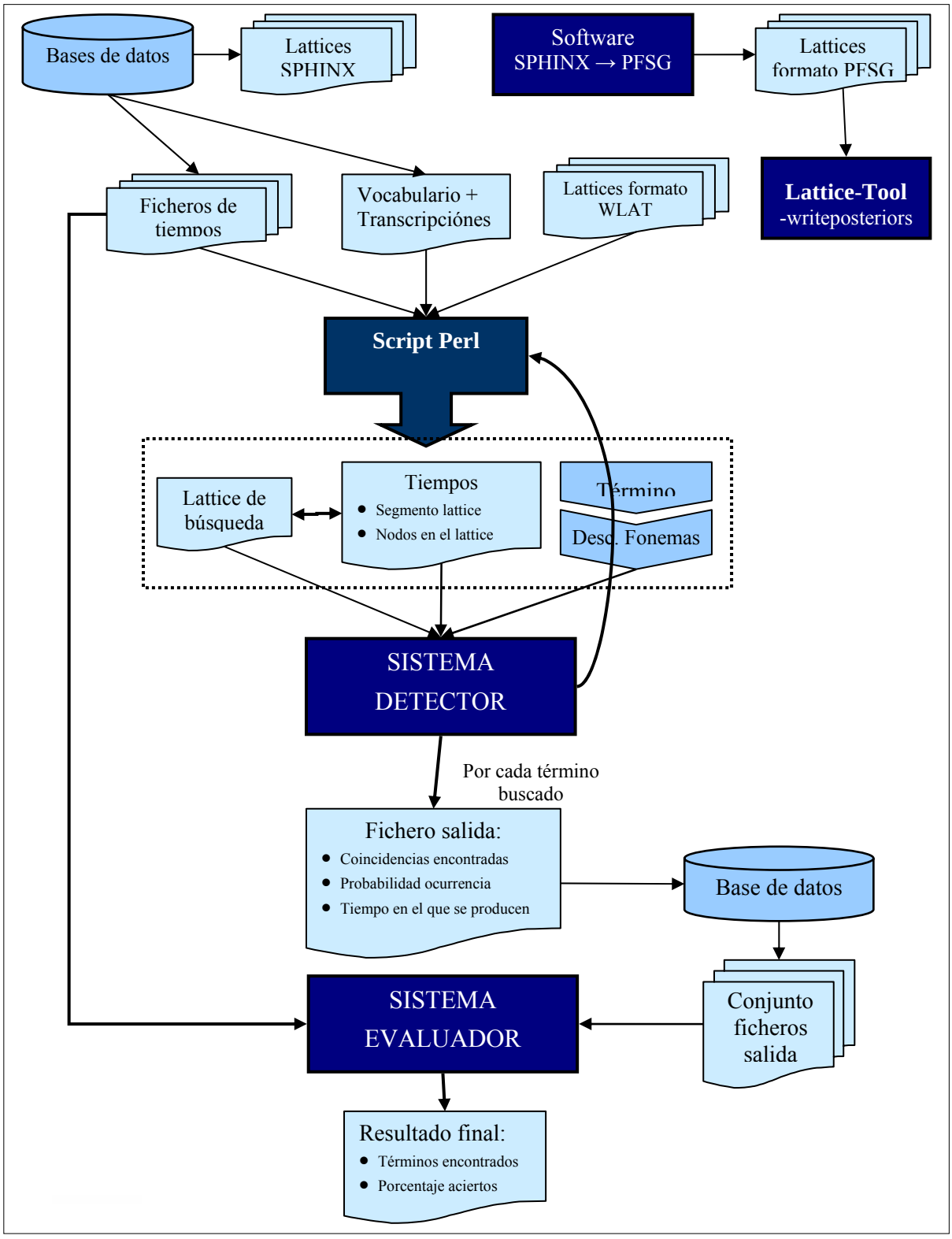


Figura 21. Diagrama de bloques de la fase de búsqueda

Ya que el punto de partida es un reconocedor de voz implementado que devuelve un archivo de audio en el formato de lattice de SPHINX '.sphx.lat', la parte de desarrollo comienza tras la fase de indexado.

El primer paso consiste en el procesado de la información que se tiene en las distintas bases de datos para convertirla en un formato que permita un manejo óptimo, mediante éste proceso de manipulación ya descrito en la sección anterior, se obtiene el formato de lattices WLAT con el uso de la herramienta Lattice-Tool de SRILM. Éste ha sido transformado previamente desde SPHX a PFSG, mediante un software ya implementado, facilitado por el grupo ATVS.

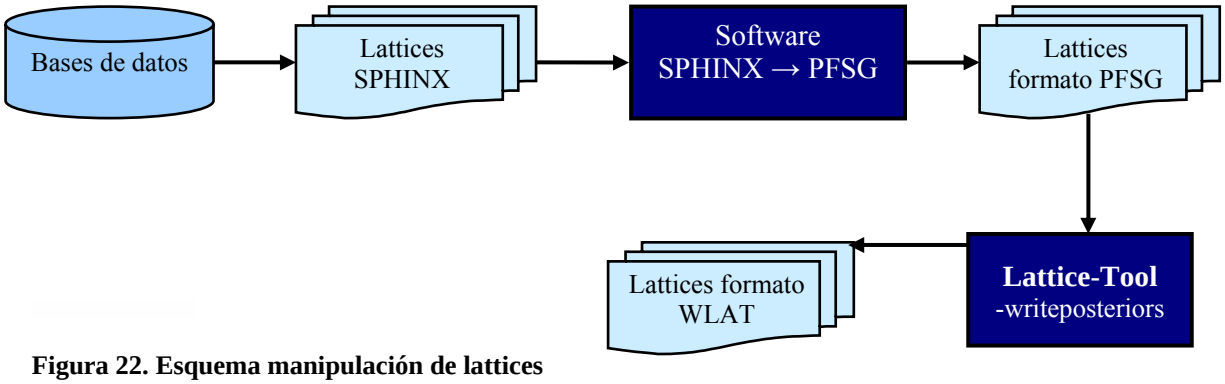


Figura 22. Esquema manipulación de lattices

Además de los lattices WLAT, en las bases de datos se han generado los ficheros .vocab que contienen el vocabulario completo de los archivos de audio, y las correspondientes transcripciones fonéticas de cada uno de éstos términos.

También el archivo de tiempos, 'segment.time' en el que se guarda información sobre el momento de inicio y la duración de cada uno de los segmentos en los que se divide el archivo de audio; y los asociados a cada segmento concreto, 'segmentX.sphx.lat' dando el tiempo de inicio de cada uno de los nodos (fonemas) contenidos en el segmento.

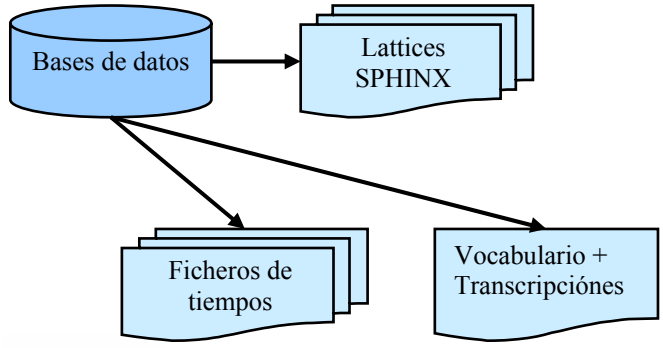


Figura 23. Tipos de ficheros extraídos

En esta imagen se observan los ejemplos de los tres formatos de lattice usados en el desarrollo, y su orden de procesado:

.sphx → .pfsg + .pfsg → .wlat

Figura 24. Formatos de lattice

El siguiente paso fue la creación de un script en lenguaje Perl que encargado del tratamiento de ficheros que entran secuencialmente en el sistema detector. Contiene variables con los caminos de ficheros entrada y salida, así como el ejecutable del detector; mediante un primer bucle recorre todos los términos a consultar que irán pasando al ejecutable, dentro de éste existe un segundo bucle que recorre todos los fragmentos de audio en que está dividido el original, ahora convertidos en segmentos con formato lattice.

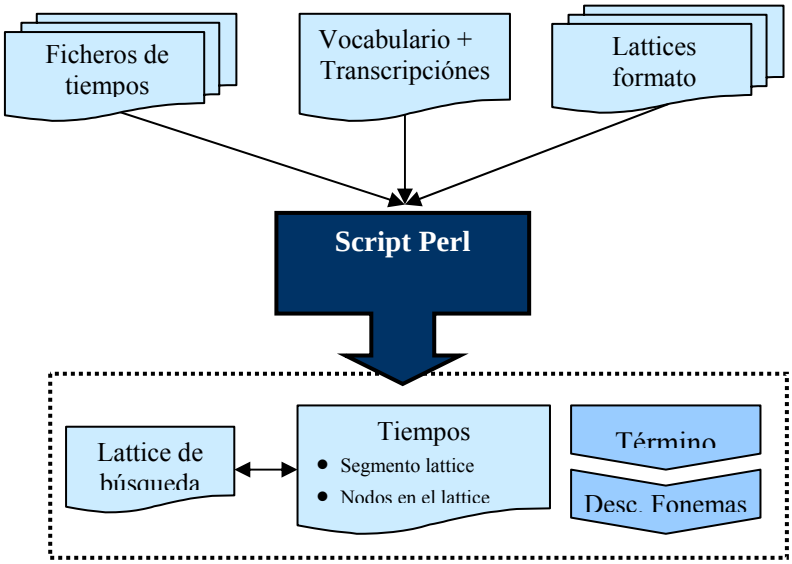


Figura 25. Entrada y salida del sistema perl

Así, se llama al ejecutable del sistema detector tantas veces como consultas de términos se hagan, pasándole los argumentos requeridos:

- Segmento lattice en el que se llevará a cabo el proceso de búsqueda.
- Transcripción fonética del término a buscar.
- Fichero de salida para la escritura de resultados.
- Fichero de tiempos para la posterior localización del término buscado en su tiempo real dentro del audio.

Dichos parámetros serán pasados por lo tanto al sistema detector. Si el archivo de audio está fragmentado en 20 segmentos, el número de llamadas al detector, será de 20 por la cantidad total de términos a consultar.

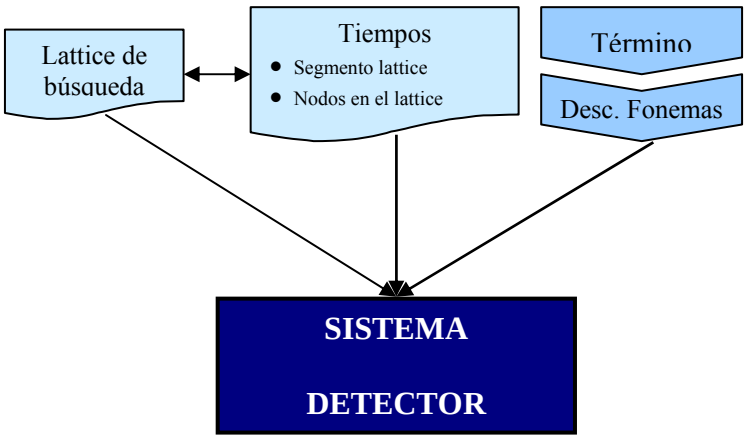


Figura 26. Parámetros entrada del sistema detector

4.1. Detección

Como se describe en la sección de Medios disponibles, la plataforma utilizada para llevar a cabo el desarrollo software de este algoritmo ha sido la aplicación KDevelop sobre el sistema operativo Unix. El lenguaje de programación usado ha sido C, la elección del mismo se debe a la necesidad de un lenguaje que ofrezca una velocidad de cómputo considerable, debido a la necesidad de un tiempo de ejecución que no exceda unos límites proporcionales a la duración total del archivo de audio en el que se están realizando las búsquedas.

4.1.1. Descripción del sistema detector

La funcionalidad del algoritmo se puede resumir como el proceso de búsqueda de unos términos determinados en un fichero de texto previamente convertido, y cuyo resultado será la localización de los términos detectados, el tiempo relativo en el que se produjeron dentro del fragmento de audio, y además la probabilidad de que se dé dicha ocurrencia.

4.1.2. Estructura

El software desarrollado para la implementación del detector está dividida en dos módulos, principal y funcional.
El primero de ellos contiene el main.c de nuestro programa, la función principal, ella recibe todos los parámetros necesarios para la ejecución del mismo.

El código de la función comienza con la declaración e inicialización de todas las variables usadas. A continuación se pasa a la lectura de la cabecera del archivo perteneciente al lattice de búsqueda, de este modo se reserva la memoria necesaria para la creación de la estructura ‘grafo’, que consiste en un array compuesto por nodos, tantos como fonemas contiene el lattice de entrada.

Dichos nodos tienen una estructura definida. Cada uno de ellos guardado en una de las posiciones del array estará formado por:

- un entero que guarda el índice del nodo,
- una cadena con el fonema que representa dicho nodo,
- un valor que guarda la probabilidad correspondiente al fonema,
- un array de nodos vecinos que contiene los nodos vecinos, es decir aquellos a los que se puede producir una transición desde el nodo índice,
- un array con las probabilidades de que se produzcan las transiciones a los nodos vecinos.

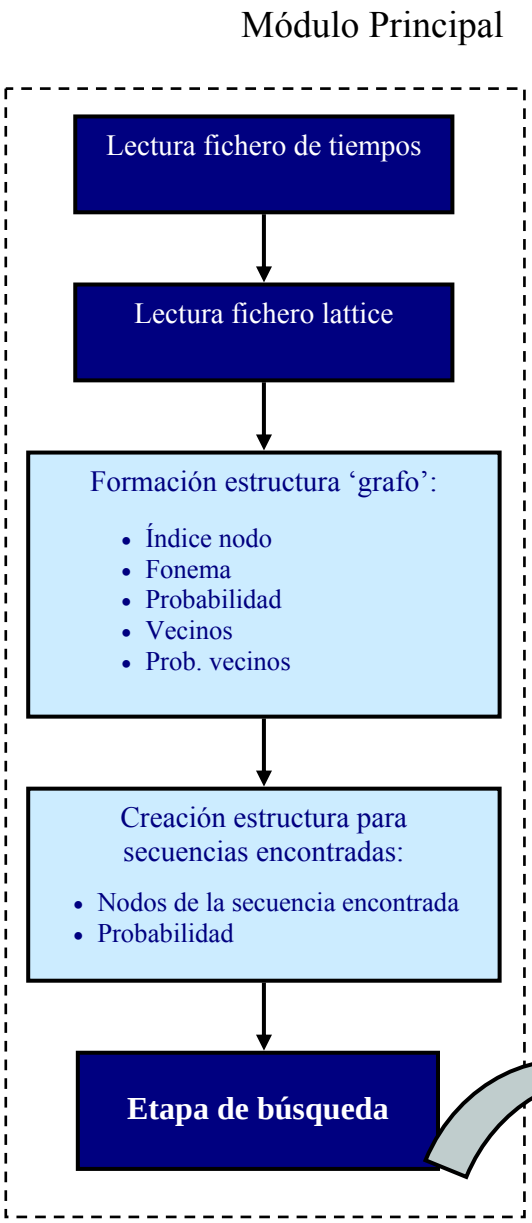


Figura 27. Módulo principal del detector

Llegados a este punto, comienza el tiempo del segundo módulo, la fase de búsqueda del detector, que está dividida en dos tipos:

- búsqueda lineal, y
- búsqueda en vecinos.

La primera de ellas, implementada en la función *busqueda_lineal*, recibe como argumentos la estructura grafo, la secuencia a buscar y la estructura que guardará las secuencias encontradas.

Para seguir un orden coherente el índice del array 'grafo' corresponde con el valor índice de cada nodo. También se crea una estructura dónde se irá guardando cada secuencia encontrada coincidente con el término de búsqueda.

Dicha estructura contendrá:

- secuencia de nodos del término encontrado
- probabilidad

Tras formar la estructura 'grafo', se procede a leer el término de búsqueda, que ya es recibido en el programa como transcripción fonética, por lo tanto, se tiene una secuencia de fonemas cuyo tamaño equivale al número de nodos que han de ser encontrados en el grafo.

A continuación se crea la estructura en la que serán almacenadas las secuencias que se vayan encontrando durante la ejecución. Serán guardados los nodos que componen la secuencia, así como la probabilidad de que se dé la misma.

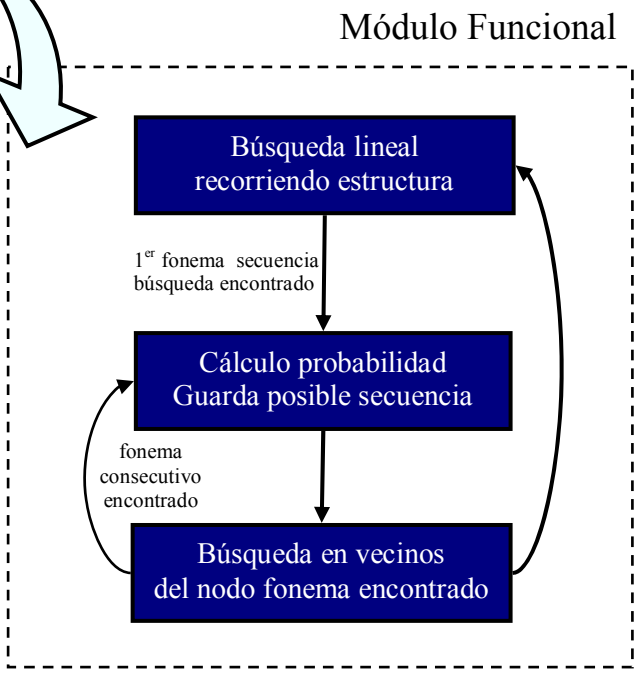


Figura 28. Módulo funcional del detector

Se comienza a realizar la búsqueda recorriendo la estructura grafo en orden descendente, desde el último nodo hasta el inicial, ya que el formato WLAT desde el que se ha construido la estructura, contiene los fonemas del archivo de voz en dicho orden, es decir, los nodos finales corresponden a los primeros fonemas del audio.

Si se realiza la búsqueda lineal completa desde el último al primer nodo, y no ocurre ninguna coincidencia, el fonema y por lo tanto la secuencia de fonemas, no se encuentra en el segmento lattice de búsqueda. No se darán resultados en la salida.

Si por el contrario se da una coincidencia entre el primer fonema de la secuencia de búsqueda y uno de los nodos (fonemas) de la estructura grafo, se está ante el primer posible fonema de la secuencia de búsqueda, se procede a guardar dicho nodo con su probabilidad en la estructura destinada a ello, y se continúa la búsqueda con el segundo fonema.

Es aquí cuando se recurre al segundo tipo de búsqueda, cuya función es la llamada *busqueda_vecinos.*
Puesto que en la estructura grafo se dispone de toda la información de cada uno de los nodos, simplemente se necesita identificar mediante su índice para obtener sus nodos vecinos, que son pasados como argumento a la función *busqueda_vecinos*. Una vez en ésta, se procede a la búsqueda del siguiente fonema coincidente.

En este punto existen dos posibilidades, la primera es hallar entre los vecinos el segundo fonema de búsqueda, con lo que tras guardarlo y calcular la probabilidad de estar en dicho nodo, se llamaría recursivamente a la función *busqueda_vecinos* que buscaría entre los vecinos de éste segundo nodo, el tercero, volviendo al principio de la función al ser la misma recursiva, y repitiendo por lo tanto éste proceso.

Sin embargo, la segunda posibilidad es que el segundo fonema no se encuentre entre los vecinos del primero, esto será indicativo de que el primer fonema es erróneo y no pertenece a la secuencia de búsqueda, por lo tanto, se volverá a la función *busqueda_lineal* continuando la búsqueda en orden descendente a partir de la falsa alarma (primer fonema coincidente no válido).

Esta situación se puede producir en cualquiera de los fonemas que se van buscando secuencialmente, de modo que si la búsqueda consiste por ejemplo, en una secuencia de 6 fonemas, y se da una situación en la que 5 de ellos ya han sido hallados, si entre los vecinos de este quinto nodo no se encuentra el último de los fonemas, se descartará la secuencia completa, continuando la búsqueda con otros posibles caminos que aún no se han descartado.

Una vez recorrido todo el lattice de búsqueda y con la estructura de secuencias encontradas rellena se vuelve al módulo principal y se procede al cálculo de los tiempos de cada una de las secuencias. Estando calculados se realiza la escritura de un fichero de texto, que contendrá:

- Todas las secuencias encontradas representadas mediante sus nodos (número)
- Tiempo en el que se produce cada una de las secuencias
- Probabilidad de ocurrencia de dichas coincidencias

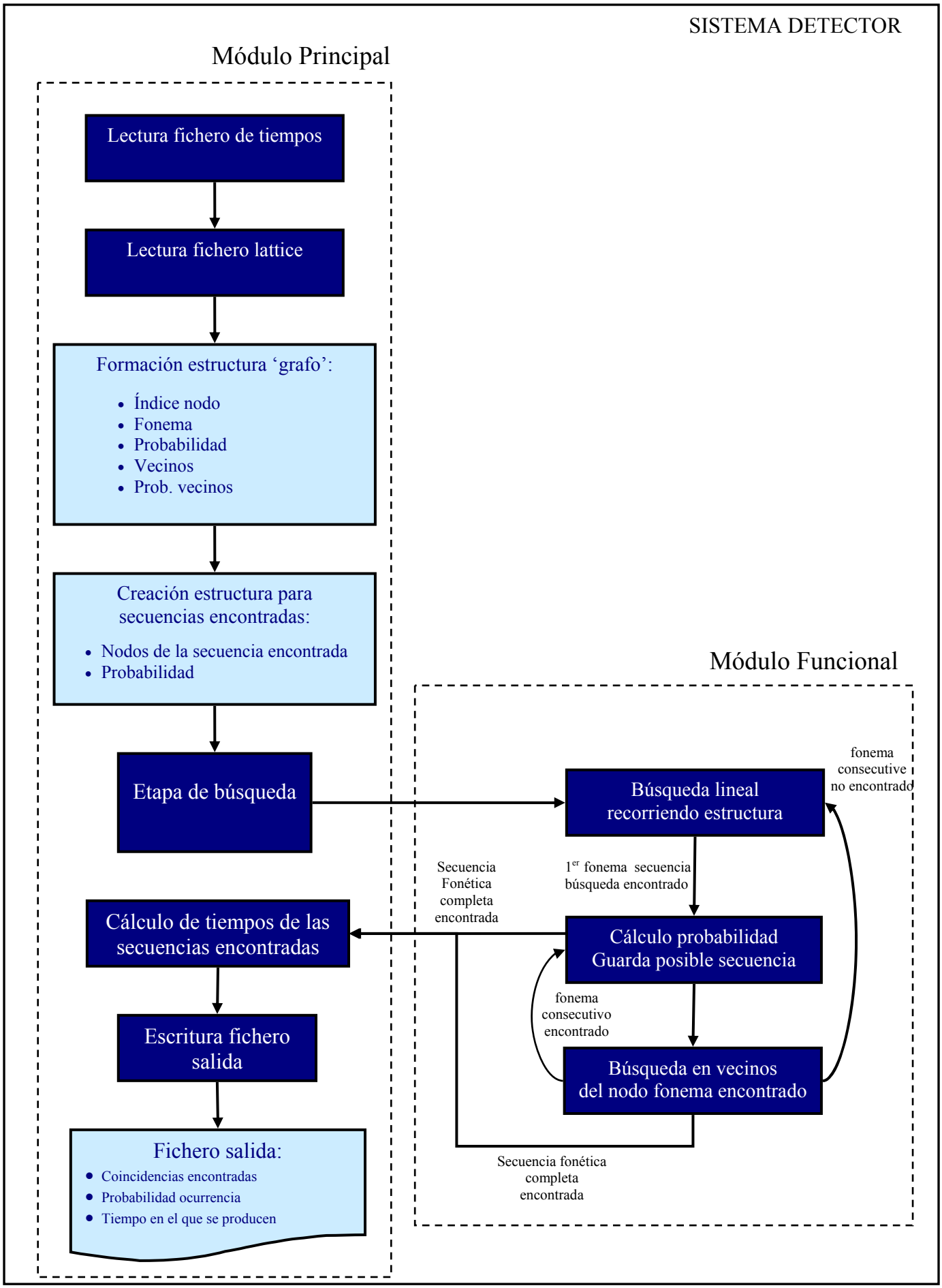


Figura 29. Sistema detector

4.2. Evaluación

Para el desarrollo del sistema evaluador se implementa un script en lenguaje Perl cuya funcionalidad, descrita en la sección de diseño, es la verificación de los resultados obtenidos por el detector.

La verificación de los resultados hallados se realiza mediante un proceso de comparación con los archivos de referencia, que comprobarán la salida del sistema detector dando lugar a unos resultados identificadores de la tasa de aciertos en las coincidencias encontradas.

4.2.1. Estructura

La clave del sistema evaluador es el acceso al fichero de tiempo real del audio que está siendo procesado, por cada segmento de conversación existe un fichero .rttm (Real-Time Trade Matching). En dicho archivo se encuentra la conversación completa con los tiempos de inicio y duración de cada una de las palabras.
Están incluidos los propios silencios etiquetados como NON-LEX para distinguirlos de los términos válidos.

Un ejemplo del fichero .rttm de un segmento de audio del tipo de habla espontánea de noticias, es el siguiente:

Tiempo de inicio **Duración**

```
LEXEME emednw01 1 0.05  0.18 buenas lex PRE <NA>
LEXEME emednw01 1 0.23  0.32 tardes lex PRE <NA>
LEXEME emednw01 1 0.69  0.1 la lex PRE <NA>
LEXEME emednw01 1 0.79  0.39 economía lex PRE <NA>
LEXEME emednw01 1 1.18  0.43 española lex PRE <NA>
NON-LEX emednw01 1 1.61  0.03 _ other PRE <NA>
LEXEME emednw01 1 1.64  0.19 con lex PRE <NA>
LEXEME emednw01 1 1.83  0.06 el lex PRE <NA>
LEXEME emednw01 1 1.89  0.34 cuadro lex PRE <NA>
LEXEME emednw01 1 2.23  0.1 de lex PRE <NA>
LEXEME emednw01 1 2.33  0.55 previsiones lex PRE <NA>
NON-LEX emednw01 1 2.88 0.03 _ other PRE <NA>
LEXEME emednw01 1 2.91  0.11 del lex PRE <NA>
LEXEME emednw01 1 3.02  0.34 gobierno lex PRE <NA>
NON-LEX emednw01 1 3.36  0.03 _ other PRE <NA>
LEXEME emednw01 1 3.39  0.15 para lex PRE <NA>
LEXEME emednw01 1 3.54  0.21 este lex PRE <NA>
LEXEME emednw01 1 3.75  0.21 año lex PRE <NA>
NON-LEX emednw01 1 3.96  0.14 _ other PRE <NA>
LEXEME emednw01 1 4.1  0.08 y lex PRE <NA>
LEXEME emednw01 1 4.18  0.09 la lex PRE <NA>
LEXEME emednw01 1 4.27  0.51 política lex PRE <NA>
```

Por cada una de las palabras de la grabación, se tiene tanto el momento de inicio como el tiempo que el hablante tarda en pronunciar la palabra, ambos en segundos.

El evaluador comienza leyendo cada una de éstas líneas, omitiendo aquellas cuya etiqueta es NON-LEX. De ellas extrae los tiempos, guardándolos en variables:

```
LEXEME emednw01 1  0.05  0.18  buenas lex PRE <NA>
```

Tiempo de inicio = 0,05 seg.

```
LEXEME emednw01 1  0.05  0.18  buenas lex PRE <NA>
```

Tiempo de fin = 0,23 seg.

Y el término a evaluar:

```
LEXEME emednw01 1 0.05  0.18 buenas lex PRE <NA>
```

Término a evaluar = "buenas"

Dicho término se busca en el conjunto de ficheros de salida que se han generado tras el proceso de búsqueda con el sistema detector. Dentro de este fichero aparecen todas las coincidencias encontradas junto con su tiempo de ocurrencia, por lo que el programa comienza a leer el fichero y por cada coincidencia realiza un cálculo para determinar si realmente corresponde con los tiempos del fichero .rttm.

La coincidencia tendrá que darse entre los tiempos de inicio y fin, para que se considere válida la detección. Puesto que no es real hablar de un sistema perfecto, hay que considerar un margen de error y aplicarlo en el cálculo, de manera que se fija este valor en 0,5 segundos, dando por válidas aquellas detecciones que correspondan con los tiempos del fichero .rttm ± 0,5 segundos. Se determina como acierto o fallo, y se procede a la evaluación del siguiente término.

Una vez recorrido en su totalidad, se tienen las variables que guardan el número de aciertos y fallos dados, así como el número total de términos en la grabación. Con estos valores se calcula el porcentaje de aciertos, que se guarda en un fichero de evaluación.

En el siguiente esquema se puede observar un diagrama de bloques del proceso que se acaba de describir.

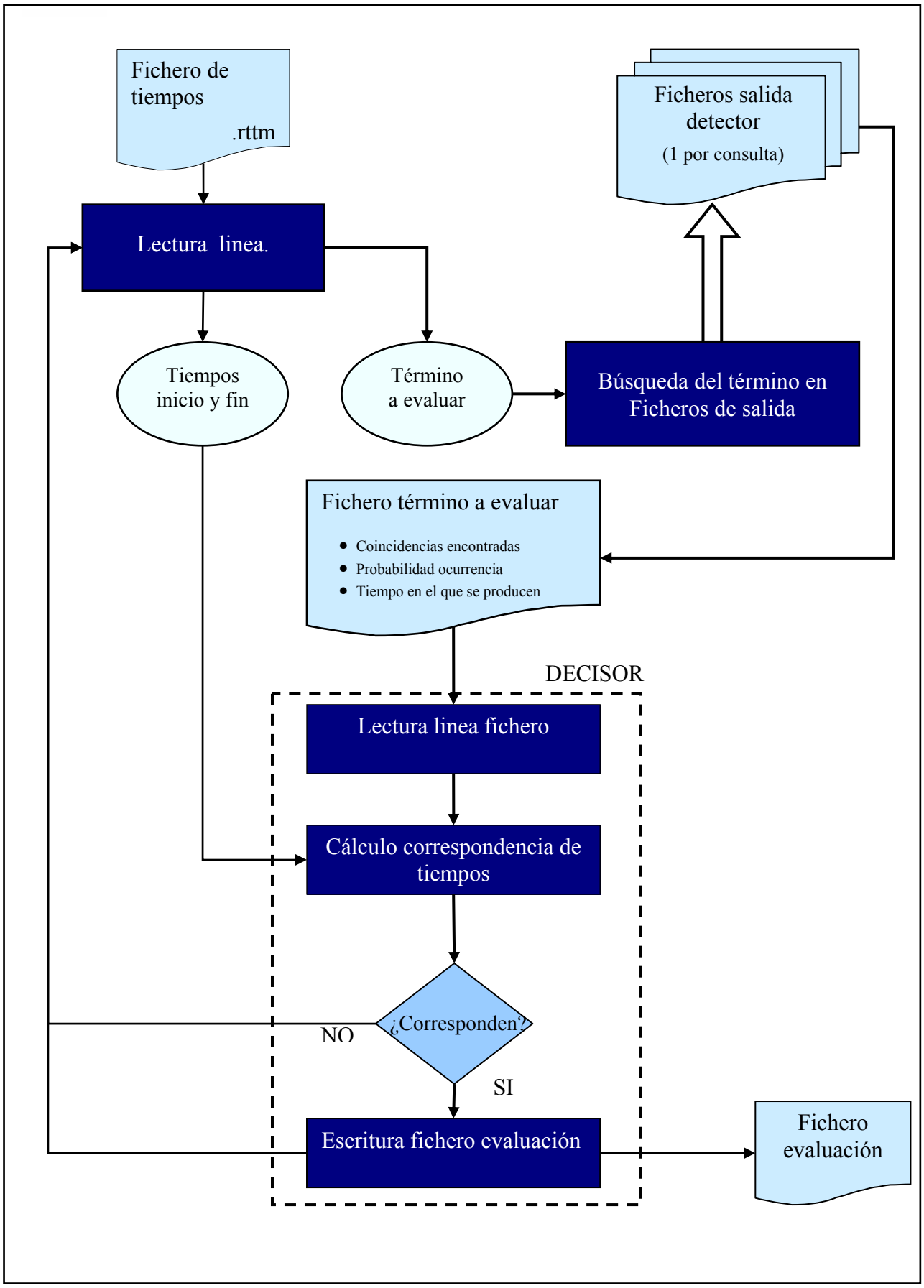


Figura 30. Sistema evaluador

5. Pruebas y resultados

Una vez finalizada la etapa de desarrollo se procede a la integración del sistema, para llevar a cabo la fase de pruebas, encadenando todos los elementos que lo componen.

5.1. Pruebas

Como se describe en la sección anterior, las pruebas realizadas se basan en el corpus CORAL-ROM desarrollado por el laboratorio de Lingüística Informática de la UAM. El conjunto de subtipos de habla espontánea que se ha sometido a evaluación, es:

1. Habla Formal	
1.1. Emisión de noticias	
1.1.1. Entrevistas	emedin
1.1.2. Meteorología	emedmt
1.1.3. Noticias	emednw
1.1.4. Informes	emedrp
1.1.5. Ciencia	emedsc
1.1.6. Deportes	emedsp
1.1.7. 'Talk shows'	emedts
1.2. Habla en contexto natural	
1.2.1. Negocios	enatbu
1.2.2. Conferencias	enatco
1.2.3. Leyes	enatla
1.2.4. Debates políticos	enatpd
1.2.5. Profesional	enatpe
1.2.6. Discurso	enatpr
1.2.7. Enseñanza	enatte
2. Habla Informal	
2.1. Contexto público	
2.1.1. Conversaciones públicas	epubcv
2.1.2. Diálogos públicos	epubdl
2.1.3. Monólogos públicos	epubmn
2.2. Contexto Familiar	
2.2.1. Conversaciones familiares	efamcv
2.2.2. Diálogos familiares	efamdl
2.2.3. Monólogos familiares	efammn
2.3. Telefónico	
2.3.1. Conversaciones telefónicas	etelef

Tabla 5. Conjunto de subtipos de habla espontánea sometido a evaluación

Como primer paso, por cada tipo de habla espontánea, se ha seleccionado un fragmento de grabación de manera totalmente aleatoria, cuya duración oscila entre 5 y 30 minutos, dependiendo del subtipo, aproximadamente se evalúan un total de 4 horas y media.

Cada grabación se divide a su vez en segmentos, el número de los mismos difiere dependiendo de la base de datos, la duración promedio es de 12,17 segundos.

	Número Segmentos	Duración (seg)	Duración promedio(seg)	Número Términos
emedin	32	504,15	15,75	2063
emedmt	4	152,67	38,17	654
emednw	63	576,85	9,16	2061
emedrp	69	527,47	7,64	2105
emedsc	6	375,53	62,59	1902
emedsp	22	399,32	18,15	1957
emedts	32	504,15	15,75	1885
enatbu	310	949,36	3,06	4452
enatco	409	1290,83	3,16	4405
enatla	66	1006,91	15,26	4769
enatpd	46	997,36	21,68	4132
enatpe	231	1009,09	4,37	4354
enatpr	64	164,44	2,57	499
enatte	138	655,45	4,75	2167
epubcv	190	700	3,68	2642
epubdl	112	602,08	5,38	2372
epubmn	160	831,44	5,20	2333
efamcv	205	554,69	2,71	2630
efamdl	129	383,05	2,97	2264
efammn	723	2006,56	2,78	7464
etelef	90	368,46	4,09	1670

Tabla 6. Tipos de habla evaluados

Una vez obtenido el número total de términos por cada tipo de grabación, reflejado en la tabla superior, se desarrolla un software con el lenguaje Perl, mediante el cuál, los términos con distinto número de fonemas quedan divididos en grupos, calculando así la cantidad de palabras por cada grupo.

De este modo se obtiene la proporción de palabras según su número de fonemas, (desde 1 hasta 15), para cada uno de los entornos analizados.

Se comienza con este análisis de los datos, ya que uno de los efectos a tener en cuenta a la hora de evaluar, es la influencia de la longitud de los términos en la detección. Hay que tener en cuenta que se experimentará una de las dificultades comúnmente asociadas a la búsqueda fonética, el gran número de falsas alarmas generadas cuando se buscan términos cortos.

Habla formal. Emisión de Noticias

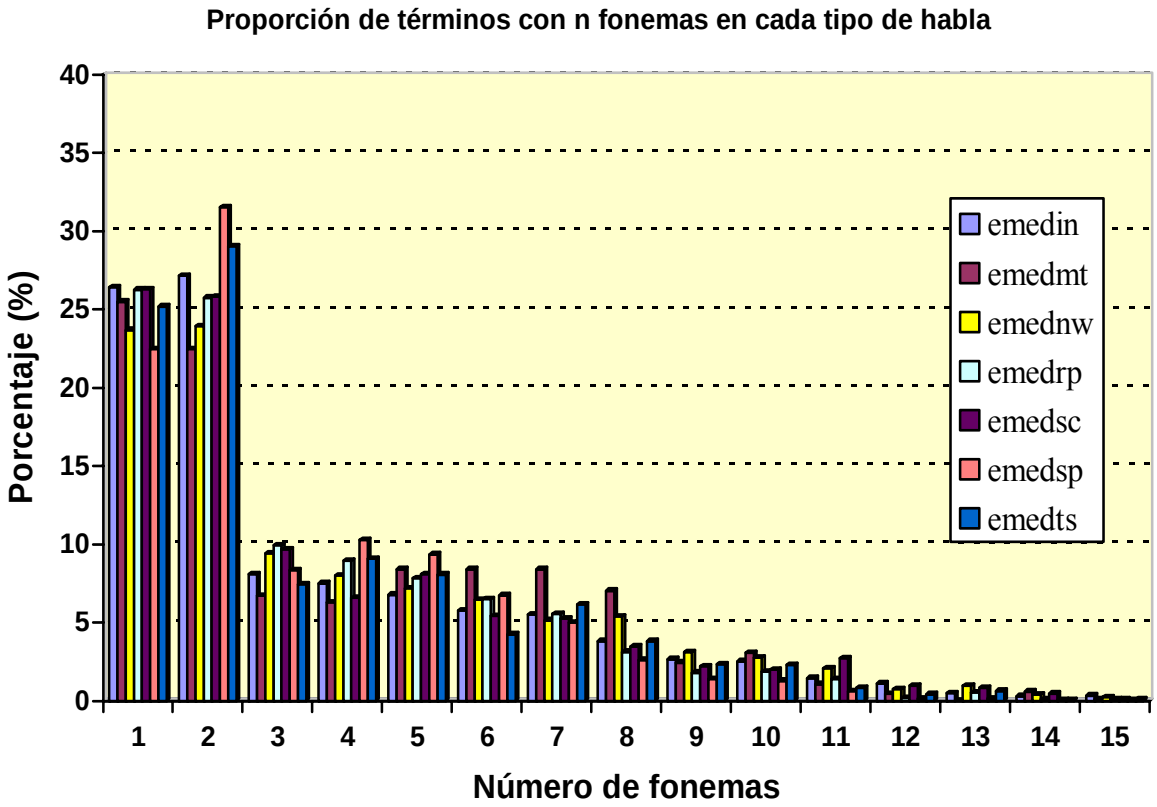


Figura 31. Proporción de términos con N fonemas

Se puede observar como decrece el número de términos a medida que hablamos de mayor número de fonemas en ellos, la mayor proporción de términos viene de aquellos con un número de fonemas bajo, (uno y dos fonemas en el término) debido a las preposiciones, artículos, conjunciones… que se dan con mucha frecuencia independientemente del tipo de grabación del que hablemos.

Dentro del subgrupo de habla formal en el contexto de medios, todos los tipos de habla siguen un patrón similar, la proporción de palabras de cada tamaño es parecida entre ellos. Destacan algunos como los de ámbito deportivo y 'talk shows', en los que es mayor la proporción de términos cortos, disminuyendo sin embargo los términos de mayor número de fonemas. Esta proporción es mínima a partir de los 11 fonemas en el término, esto se debe a que los términos que pueden darse en 'talk shows' y deportes, aun tratándose de habla formal, se encuentran dentro de un ámbito más simple, con lenguaje común, coloquial y no muy complejo, aumenta el uso de conectores entre

frases, onomatopeyas y coletillas en las sentencias, en definitiva un uso del lenguaje poco cuidado. Resulta extraño que se den términos excesivamente largos.

Ámbitos en los que más se dan términos de alto número de fonemas dentro del contexto de medios, son ciencia y meteorología, obviamente por el uso de un lenguaje más culto, con términos científicos y técnicos.

Habla formal. Contexto natural

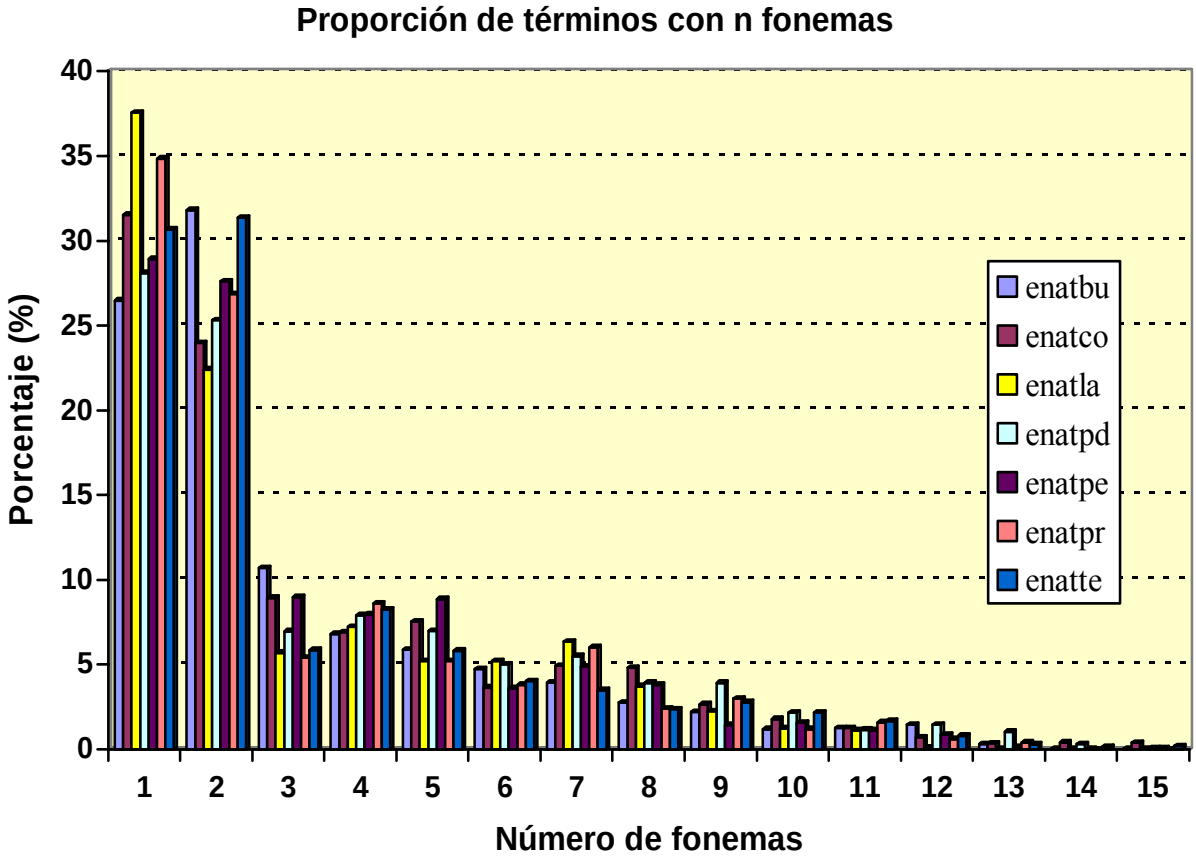


Figura 32. Proporción de términos con N fonemas

Como diferencia principal frente a los tipos de habla en el contexto de medios, se aprecia que en contexto natural, la proporción de términos con bajo número de fonemas (uno y dos) crece, llegando muchos de ellos a un 30% e incluso acercándose alguno a un 40%, frente al 25% en el que se encuentran en el contexto de medios, exceptuando los grupos de 'talk shows' y deportes.

Mientras que esta proporción de términos con bajo número de fonemas crece, la proporción de términos con un número de fonemas medio (entre 4 y 9) decrece, hay que tener en cuenta que la riqueza léxica en el contexto natural baja con respecto al contexto de medios.

De modo similar al caso anterior, a partir de 12 fonemas, el número de términos es escaso, destaca entre todos los demás, el habla en conversaciones de tipo profesional, con las puntuaciones más altas en términos de 12 y 13 fonemas, ya que estamos hablando de un vocabulario más sofisticado.

Habla informal

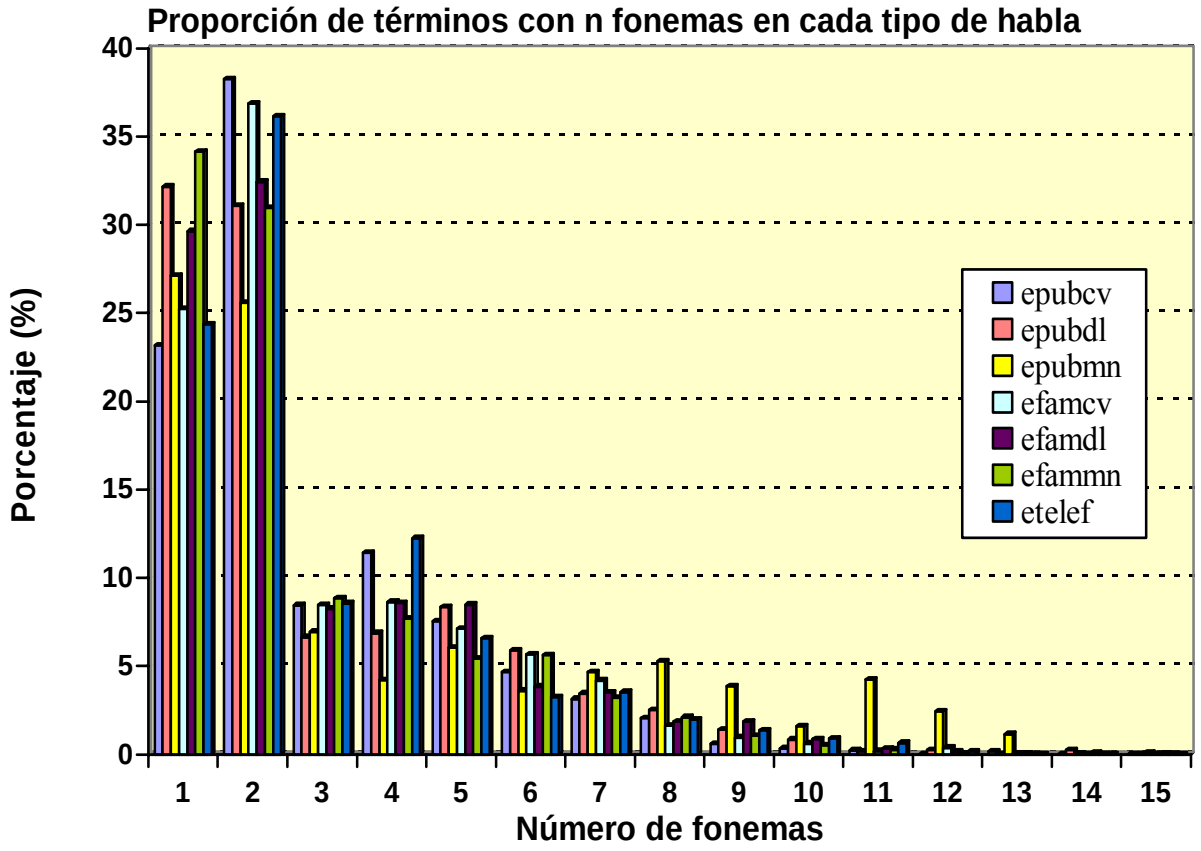


Figura 33. Proporción de términos con N fonemas

En la gráfica superior se puede comprobar que el porcentaje de palabras con bajo número de fonemas es especialmente alto en el contexto de habla informal, superando las proporciones anteriores en tipo de habla formal.

Siendo el vocabulario más pobre en habla informal, la proporción de términos con más de 9 fonemas es mínima. Destaca un tipo de habla en contexto público, la de los monólogos, con proporciones altas dentro de la media, en términos de gran número de fonemas (11, 12 o 13) se considera que estos monólogos están normalmente preparados, y tratan exposiciones sobre un tema que ha sido estudiado y del que se tiene mucha información.

Pese a las variaciones dependiendo del tipo de habla, los términos con mayor ocurrencia en las grabaciones, son aquellos con un bajo número de fonemas, cuando se habla de secuencias fonéticas cortas, la detección se convierte en una tarea de alta complejidad, ya que dichas secuencias pueden aparecer con mucha probabilidad como partes de otras palabras, y las detecciones deben hacerse en base a una información limitada.

En un experimento posterior dichos términos quedarán excluidos de la evaluación, para probar el sistema con términos que no influyan negativamente en los resultados, debido a su bajo número de fonemas.

5.1.1. Procedimiento

El criterio de evaluación para detecciones correctas es la exacta coincidencia fonética entre el término de consulta y la transcripción, con un margen de 0,5 segundos entre la correspondencia del término en la referencia y el detectado.

Se realizarán experimentos y evaluaciones sobre cada una de las 21 grabaciones en función del tipo de habla espontánea, en los primeros experimentos los términos a consultar serán todos los existentes en la grabación, más tarde se reducirán.

Por un lado se generan 21 directorios, uno por cada tipo de habla espontánea (emedin, emedmt, emednw, emedrp, emedsc, emedsp, emedts, enatbu, enatco, enatla, enatpd, enatpe, enatpr, enatte, epubcv, epubdl, epubmn, efamcv, efamdl, efammn, etelef) en los que se encuentra:

	DISPONIBLE
• directorio que contiene el conjunto de lattices que compone la grabación (del tipo específico) en formato .sphx	**Sí**
• directorio que contiene el conjunto de lattices que compone la grabación en formato .pfsg	No
• directorio que contiene el conjunto de lattices que compone la grabación en formato .wlat	No
• fichero de tiempos "segment.time" que define cuando comienza y termina cada uno de los lattices (segmentos en los que está dividido el audio)	**Sí**
• fichero de vocabulario ".vocab" con todos los términos que aparecen en la grabación y sus respectivas transcripciones	**Sí**
• fichero de tiempos ".rttm" con la transcripción de la grabación completa, incluyendo cada término con su tiempo de inicio y duración	**Sí**
• directorio que contiene los resultados de la detección, conteniendo un fichero distinto por cada término de búsqueda	No
• directorio donde se guardarán los resultados de la evaluación	No

Con el uso de estos componentes de forma combinada, se llevarán a cabo las pruebas y experimentos. Parte de ellos están disponibles al inicio, otros tendrán que ser generados a partir de los disponibles.

Como primer paso, partiendo de los lattices .shpx, se generan mediante transformación los lattices .pfsg, guardándose en su directorio respecivo, análogamente se generan y guardan los .wlat a partir de los .pfsg.

	DISPONIBLE
• directorio que contiene el conjunto de lattices que compone la grabación (del tipo específico) en formato .sphx	**Sí**
• directorio que contiene el conjunto de lattices que compone la grabación en formato .pfsg	**Sí**
• directorio que contiene el conjunto de lattices que compone la grabación en formato .wlat	**Sí**
• fichero de tiempos "segment.time" que define cuando comienza y termina cada uno de los lattices (segmentos en los que está dividido el audio)	**Sí**
• fichero de vocabulario ".vocab" con todos los términos que aparecen en la grabación y sus respectivas transcripciones	**Sí**
• fichero de tiempos ".rttm" con la transcripción de la grabación completa, incluyendo cada término con su tiempo de inicio y duración	**Sí**
• directorio que contiene los resultados de la detección, conteniendo un fichero distinto por cada término de búsqueda	No
• directorio donde se guardarán los resultados de la evaluación	No

Como segundo paso, se procederá a la generación de los resultados de detección, para ello, se ejecuta el script encargado de llamar al algoritmo de búsqueda desarrollado (detector) pasándole cada una de las consultas (términos de la grabación) junto con el lattice en el que ha de ser encontrada.

Como se describe en la sección 4., el detector tiene un componente de decisión a la hora de generar los resultados de búsqueda, obviamente son muchas las ocurrencias que se pueden dar si se valora simplemente la coincidencia exacta ortográfica entre el término de búsqueda y la posible secuencia encontrada, por ello entra en juego la probabilidad que tiene asignada cada posible ocurrencia.

Cuando se llama al detector, uno de los argumentos de entrada determina el umbral de decisión, éste será un número entero, que determina el valor N, este valor define las N-best probabilidades de la totalidad de ocurrencias halladas cuando el algoritmo recorre el lattice entero. Así, si para un determinado término de consulta, se encuentran 20 ocurrencias en el lattice, y el valor N=2, el detector se quedará con las 2 probabilidades más altas de las 20 existentes.

Se han realizado 3 experimentos distintos variando el umbral, realizando búsquedas en los 21 tipos de habla.

Ya que determinados términos, tienen gran frecuencia de ocurrencia, especialmente los de número de fonemas bajo como se ha visto en la sección anterior, el parámetro N se fijará en N=100 para que todos los términos con gran número de apariciones en la grabación tengan posibilidad de ser hallados, de esta manera será 100 el número de salidas por término consultado que dará el sistema detector.

Un segundo experimento se realizará con N=20, de esta manera habrá términos en las referencias que no llegarán a ser encontrados, pero con el caso anterior de N=100, se darán demasiadas falsas alarmas en términos con un número de fonemas medio que aparecen en la grabación en una modesta cantidad.

Por último, se lleva a cabo una búsqueda dando al parámetro N el valor de 5. A priori es conocido, que determinados términos aparecen más de 5 veces en la grabación, por lo que se tendrá un error inherente que asumir, como ventaja, el número de falsas alarmas será mínimo.

Así, en el directorio en el que se guardan los resultados de detección, existen 3 subdirectorios, uno por cada tipo de detección:

- 100 best
- 20 best
- 5 best

Éstos contienen un fichero por cada término de consulta realizado, dicho término da nombre al fichero, y facilita su posterior búsqueda a la hora de evaluar. El número de ficheros y su nombre será idéntico en estos 3 directorios, ya que corresponde a los términos de consulta. El contenido variará en función de las ocurrencias encontradas con cada experimento.

5.1.2. Evaluación

Llegados a este punto, están disponibles todos los componentes necesarios para desarrollar la evaluación.

	DISPONIBLE
• directorio que contiene el conjunto de lattices que compone la grabación (del tipo específico) en formato .sphx	**Sí**
• directorio que contiene el conjunto de lattices que compone la grabación en formato .pfsg	**Sí**
• directorio que contiene el conjunto de lattices que compone la grabación en formato .wlat	**Sí**
• fichero de tiempos "segment.time" que define cuando comienza y termina cada uno de los lattices (segmentos en los que está dividido el audio)	**Sí**
• fichero de vocabulario ".vocab" con todos los términos que aparecen en la grabación y sus respectivas transcripciones	**Sí**
• fichero de tiempos ".rttm" con la transcripción de la grabación completa, incluyendo cada término con su tiempo de inicio y duración	**Sí**
• directorio que contiene los resultados de la detección, conteniendo un fichero distinto por cada término de búsqueda	**Sí**
• directorio donde se guardarán los resultados de la evaluación	No

Por lo que se procede a la ejecución del evaluador, éste para cada tipo de habla, selecciona su fichero de tiempos ".rttm" y lo recorre quedándose con el término a consultar, el tiempo de inicio y la duración del mismo, con estos datos se procede a la búsqueda, entre los archivos de resultados, del fichero cuyo nombre coincida con el de la consulta, y recorriendo dicho fichero íntegramente calcula si las ocurrencias que se citan en él, entran dentro del tiempo que se ha obtenido del fichero ".rttm".

Dependiendo de éste cálculo, toma la decisión SI/NO determinando si la detección es correcta. Estos porcentajes quedan guardados en los ficheros de evaluación, cada uno en su directorio respectivo dependiendo del tipo de habla.

	DISPONIBLE
• directorio que contiene el conjunto de lattices que componen la grabación (del tipo específico) en formato .sphx	**Sí**
• directorio que contiene el conjunto de lattices que componen la grabación en formato .pfsg	**Sí**
• directorio que contiene el conjunto de lattices que componen la grabación en formato .wlat	**Sí**
• fichero de tiempos "segment.time" que define cuando comienza y termina cada uno de los lattices (segmentos en los que está dividido el audio)	**Sí**
• fichero de vocabulario ".vocab" con todos los términos que aparecen en la grabación y sus respectivas transcripciones	**Sí**
• fichero de tiempos ".rttm" con la transcripción de la grabación completa, incluyendo cada término con su tiempo de inicio y duración	**Sí**
• directorio que contiene los resultados de la detección, conteniendo un fichero por cada término de búsqueda distinto	**Sí**
• directorio que contiene los resultados de la evaluación	**Sí**

5.2. Resultados

Una vez obtenidos los resultados de la evaluación, se procede al análisis y resumen de los mismos.

5.2.1. Experimento A

Como se ha descrito anteriormente el primer experimento de búsqueda sobre el cuerpo de datos, se realiza fijando el parámetro N del detector a 100.

Esto implica que por cada consulta realizada, el detector devolverá las mejores 100 probabilidades de todas las ocurrencias encontradas, con estos resultados se ha evaluado cada tipo de habla independientemente, y el porcentaje de aciertos hallado ha sido el siguiente:

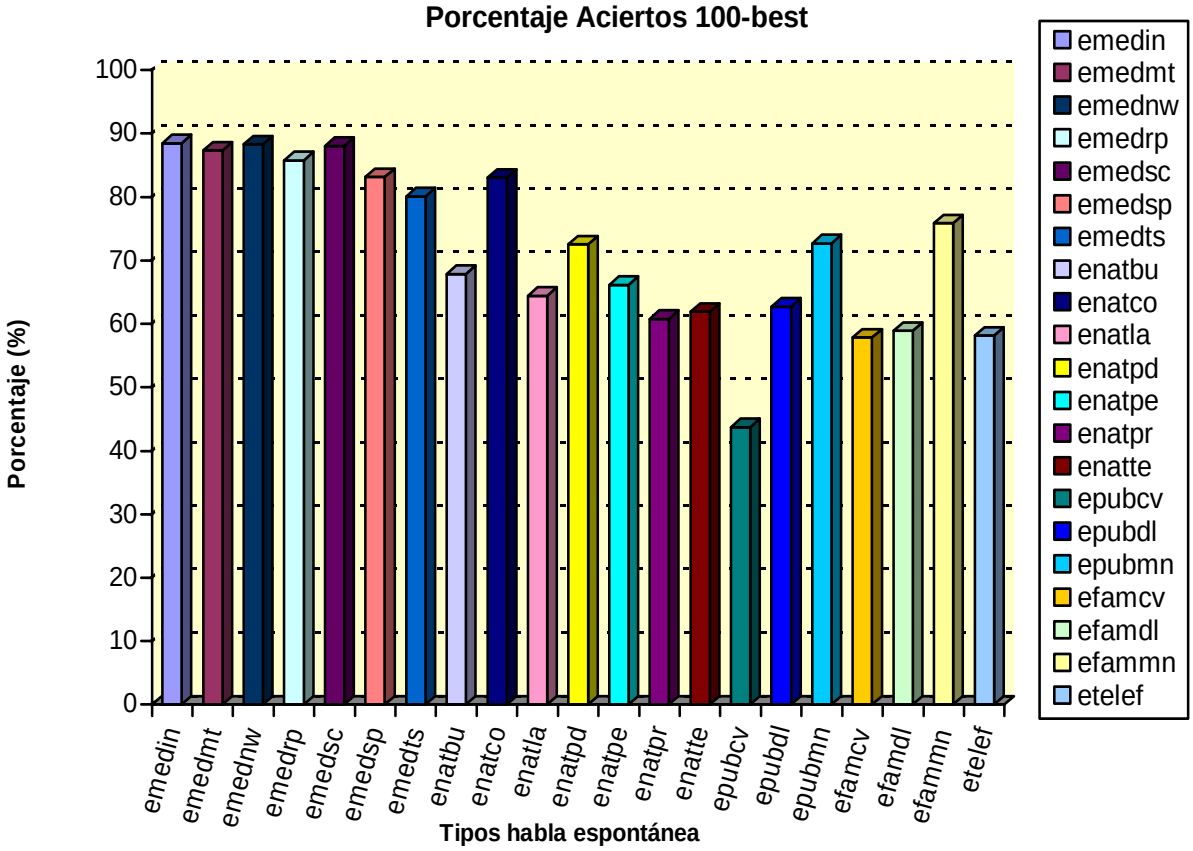


Figura 34. Porcentaje de aciertos 100-best

Los resultados obtenidos son muy buenos, ya que el umbral de decisión es poco restrictivo, los aciertos oscilan entre un 60% y 90% para los distintos tipos de habla espontánea, exceptuando el caso de conversaciones públicas, en el que el porcentaje de aciertos se reduce al 45% aproximadamente. Esto se debe a que el tipo de habla es informal, conversaciones públicas, y en este tipo se pueden dar muchos factores que provocan dificultad en la detección, como por ejemplo el número de hablantes que interactúan simultáneamente, provocando confusión a la hora de enlazar los fonemas de forma sucesiva para formar la palabra.

También se tiene en cuenta que el lenguaje y uso del mismo es más descuidado, tratándose de un ámbito informal, conversaciones coloquiales.

El mejor porcentaje se obtiene en habla en medios, exactamente entrevistas, en este caso no tenemos el inconveniente anterior, ya que los participantes hablan por turnos, sin que los fonemas se solapen en el tiempo. Además el lenguaje es mucho más cuidado, y los hablantes pronuncian de forma clara y metódica.

A pesar de los buenos resultados, hay que tener en cuenta el inconveniente de las falsas alarmas que se producen, considerando como detecciones correctas muchas que no lo son.

En esta tabla aparecen ordenados en orden decreciente los porcentajes de acierto por cada tipo de habla espontánea evaluada:

	Habla	Contexto	Tipo	Aciertos
emedin	Formal	Medios	Entrevistas	88,48%
emednw	Formal	Medios	Noticias	88,29%
emedsc	Formal	Medios	Ciencia	88,03%
emedmt	Formal	Medios	Meteorología	87,33%
emedrp	Formal	Medios	Informes	85,81%
emedsp	Formal	Medios	Deportes	83,15%
enatco	Formal	Natural	Conferencias	83,03%
emedts	Formal	Medios	'Talk shows'	80,05%
efammn	Informal	Familiar	Monólogos	75,85%
epubmn	Informal	Público	Monólogos	72,71%
enatpd	Formal	Natural	Debates Políticos	72,52%
enatbu	Formal	Natural	Negocios	67,83%
enatpe	Formal	Natural	Discursos	66,14%
enatla	Formal	Natural	Leyes	64,41%
epubdl	Informal	Público	Diálogos	62,7%
enatte	Formal	Natural	Enseñanza	61,9%
enatpr	Formal	Natural	Conversaciones profesionales	60,8%
efamdl	Informal	Familiar	Diálogos	58,91%
etelef	Informal	Telefónico	Conversaciones telefónicas	58,16%
efamcv	Informal	Familiar	Conversaciones	57,82%
epubcv	Informal	Público	Conversaciones	43,7%

Tabla 7. Porcentaje de aciertos con detección 100-best

Entre las mejores puntuaciones se puede observar algún tipo de habla en contexto natural, ya que tiene características parecidas a los tipos de habla en contexto de medios, como las conferencias que siguen un patrón parecido en cuanto a cuidado del lenguaje, número de hablantes simultáneamente, etc, a pesar de pertenecer al subgrupo de contexto natural.

5.2.2. Experimento B

El segundo experimento de búsqueda sobre el cuerpo de datos, se realiza fijando el parámetro N del detector a 20.

Por cada consulta realizada, el detector devolverá las mejores 20 probabilidades de todas las ocurrencias encontradas, con estos resultados se ha evaluado cada tipo de habla, y el porcentaje de aciertos hallado ha sido el siguiente:

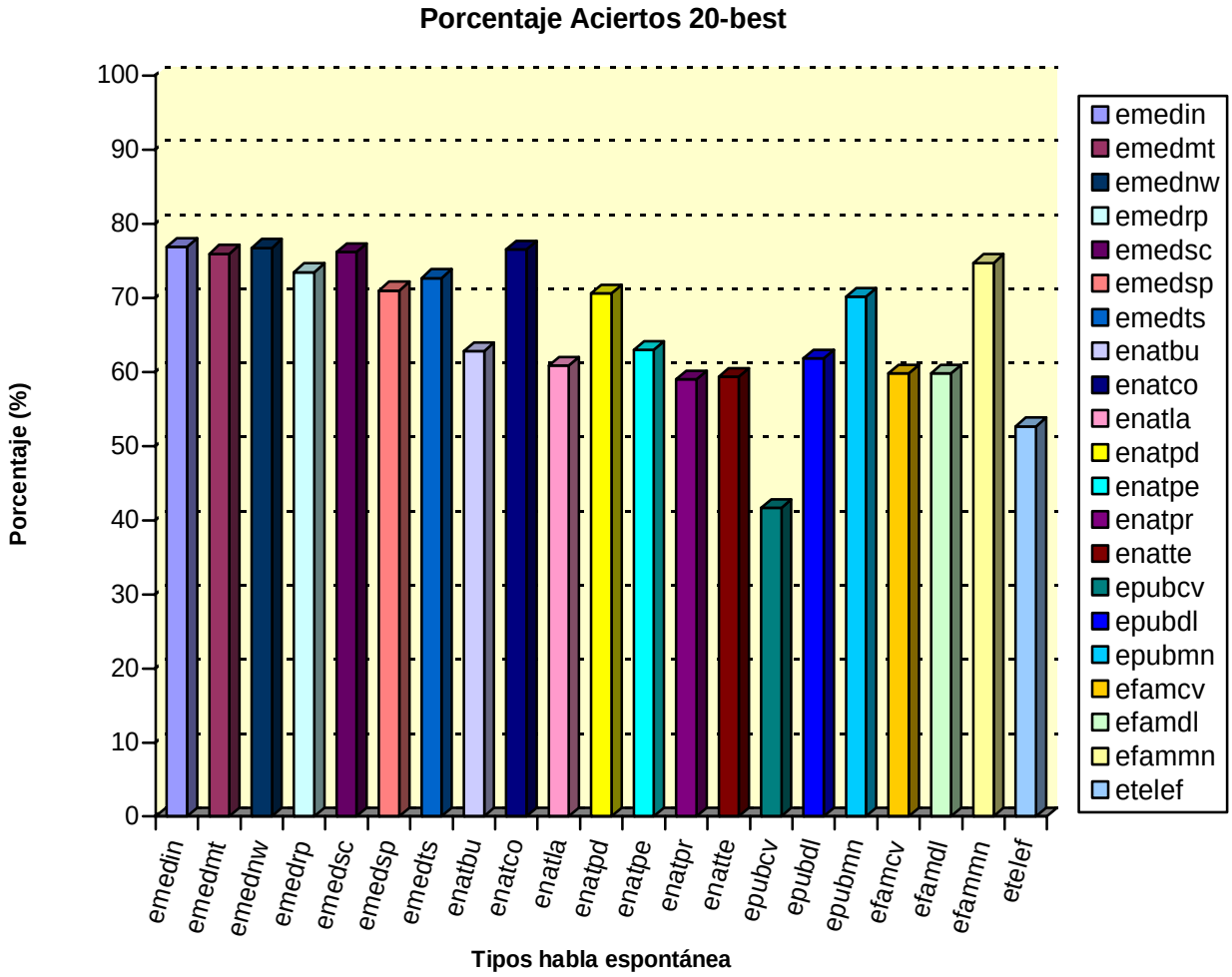


Figura 35. Porcentaje de aciertos 20-best

En esta segunda evaluación se comprueba que el porcentaje de aciertos baja al reducir el número de salidas del detector, pero por otro lado favorece la reducción de falsas alarmas.

Los resultados oscilan ahora entre un 50% y un 80% de aciertos en los distintos tipos de habla, excluyendo de nuevo las conversaciones públicas en las que el porcentaje es cercano al 40%.

La mejor puntuación se sigue dando en tipo de habla en medios, esta vez en entrevistas.

Los porcentajes de acierto de los distintos tipos de habla se ven reducidos debido al nuevo umbral de decisión, que es más restrictivo en este caso, siguen siendo los tipos de habla formal los que se mantienen con las mejores puntuaciones, aunque junto a ellos se incluye también el tipo de habla en contexto natural de conferencias. Sus resultados no han variado mucho de un experimento a otro, lo que implica que no existen demasiados términos en la grabación que han quedado fuera de la detección tras el cambio de umbral.

En la siguiente tabla se observan los valores en orden decreciente de los porcentajes de acierto en los distintos tipos de habla espontánea:

	Habla	Contexto	Tipo	Aciertos
emedin	Formal	Medios	Entrevistas	76,85%
emednw	Formal	Medios	Noticias	76,73%
enatco	Formal	Natural	Conferencias	76,52%
emedsc	Formal	Medios	Ciencia	76,18%
emedmt	Formal	Medios	Meteorología	75,93%
efammn	Informal	Familiar	Monólogos	74,69%
emedrp	Formal	Medios	Informes	73,45%
emedts	Formal	Medios	'Talk shows'	72,67%
emedsp	Formal	Medios	Deportes	70,98%
enatpd	Formal	Natural	Debates Políticos	70,6%
epubmn	Informal	Público	Monólogos	70,14%
enatpe	Formal	Natural	Discursos	62,95%
enatbu	Formal	Natural	Negocios	62,78%
epubdl	Informal	Público	Diálogos	61,83%
enatla	Formal	Natural	Leyes	60,84%
efamcv	Informal	Familiar	Conversaciones	59,78%
efamdl	Informal	Familiar	Diálogos	59,78%
enatte	Formal	Natural	Enseñanza	59,34%
enatpr	Formal	Natural	Conversaciones profesionales	58,96%
etelef	Informal	Telefónico	Conversaciones telefónicas	52,67%
epubcv	Informal	Público	Conversaciones	41,61%

Tabla 8. Porcentaje de aciertos en los distintos tipos de habla espontánea para 20-best

El orden no está muy alterado en comparación a los resultados 100-best, los mejores y peores se mantienen. Para los distintos tipos de habla el porcentaje de aciertos baja alrededor de un 20%.

5.2.3. Experimento *C*

En el siguiente experimento, el parámetro de detección N se bajará el valor de 5. A priori es conocido que determinados términos aparecen más veces en la grabación, que número de ocurrencias encontrará el detector, por lo que se tendrá un error inherente que asumir.

Tras la evaluación los resultados obtenidos se hallan reflejados en la siguiente gráfica:

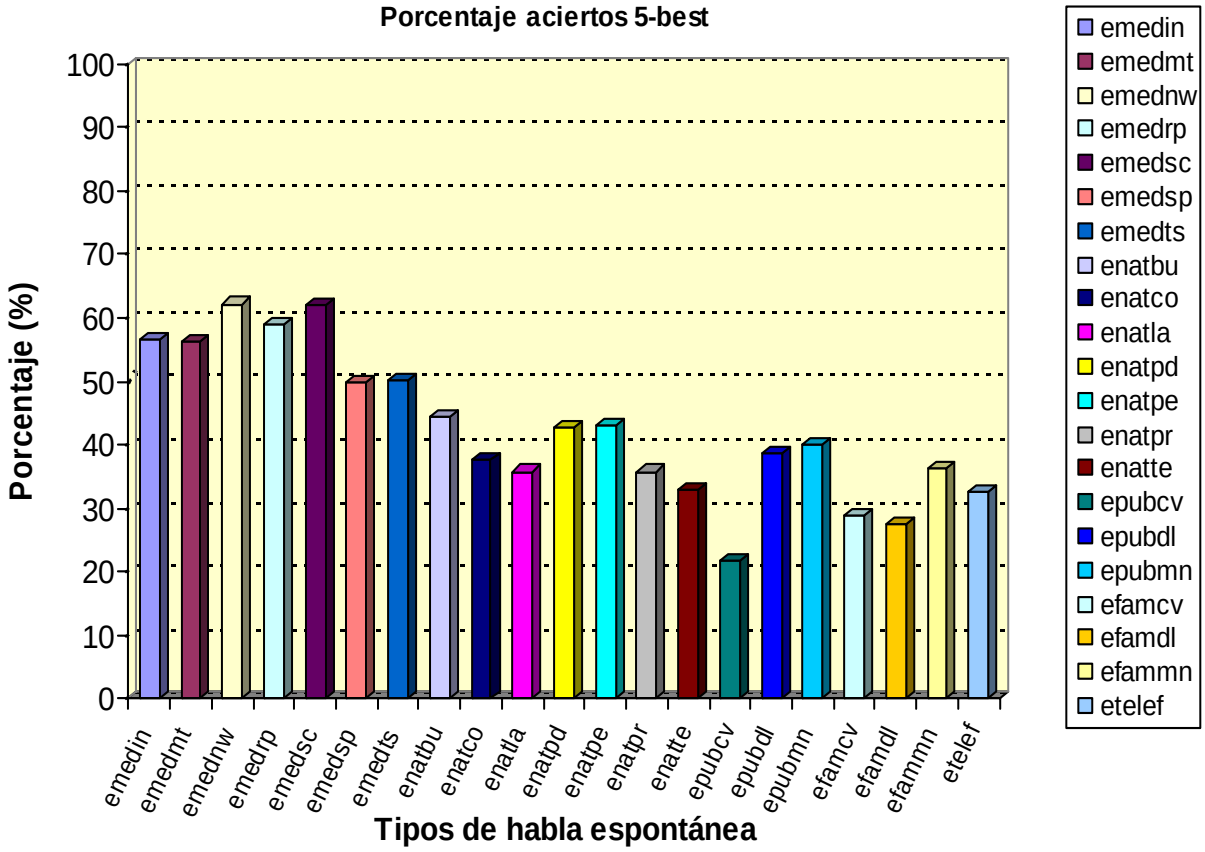


Figura 36. Porcentaje de aciertos 5-best

Los resultados han empeorado en torno a un 30% respecto al experimento A, y un 20% respecto al B. Los aciertos oscilan ahora entre un 30% y un 60% para los distintos tipos de habla, quedando como siempre por debajo de este rango las conversaciones públicas con un valor muy bajo, cercano al 20%.

Las mejores puntuaciones siguen siendo las de tipo de habla formal y contexto en medios, en este caso las grabaciones de noticias y ciencias.

Con la variación del valor de N a 5, se consigue eliminar el problema de las falsas alarmas, pero a cambio muchos de los términos quedan fuera de la detección.

El conjunto de resultados ordenados de mayor a menor según el porcentaje de aciertos en la detección, para los distintos tipos de habla espontánea, viene dado en la siguiente gráfica:

	Habla	Contexto	Tipo	Aciertos
emednw	Formal	Medios	Noticias	62,17%
emedsc	Formal	Medios	Ciencia	61,9%
emedrp	Formal	Medios	Informes	59,09%
emedin	Formal	Medios	Entrevistas	56,47%
emedmt	Formal	Medios	Meteorología	56,3%
emedts	Formal	Medios	'Talk Shows'	50,18%
emedsp	Formal	Medios	Deportes	49,74%
enatbu	Formal	Natural	Negocios	44,4%
enatpe	Formal	Natural	Discurso	43,1%
enatpd	Formal	Natural	Debates Políticos	42,6%
epubmn	Informal	Público	Monólogos	40,12%
epubdl	Informal	Público	Diálogos	38,7%
enatco	Formal	Natural	Conferencias	37,6%
efammn	Informal	Familiar	Monólogos	36,23%
enatla	Formal	Natural	Leyes	35,8%
enatpr	Formal	Natural	Profesional	35,8%
enatte	Formal	Natural	Enseñanza	32,97%
etelef	Informal	Telefónico	Conversaciones telefónicas	32,65%
efamcv	Informal	Familiar	Conversaciones	28,75%
efamdl	Informal	Familiar	Diálogos	27,54%
epubcv	Informal	Público	Conversaciones	21,8%

Tabla 9. Porcentaje de aciertos en los distintos tipos de habla espontánea para 5-best

En el siguiente gráfico se puede apreciar una comparativa entre los tres primeros experimentos. En todos los tipos de habla el porcentaje de aciertos disminuye a medida que decrece el valor de N. Se aprecia como en ciertos tipos de habla natural, el porcentaje se mantiene bastante constante entre N= 20 y N=100.

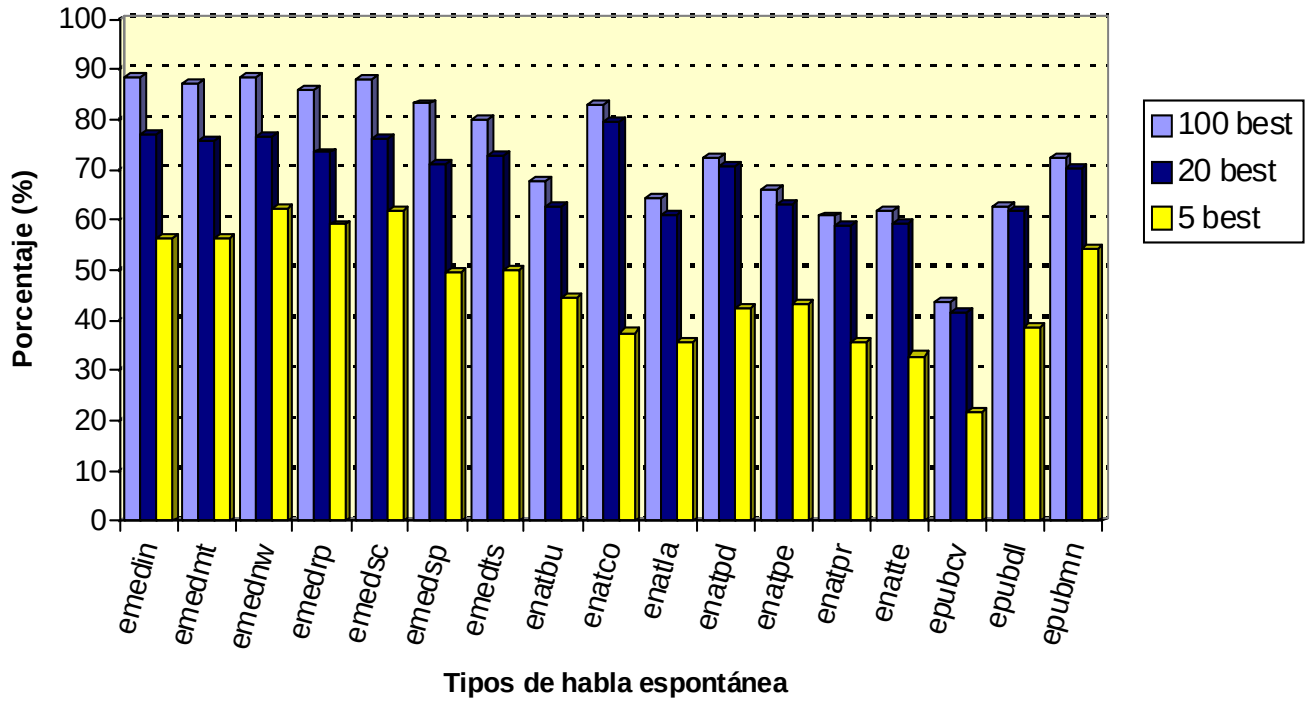


Figura 37. Comparativa aciertos entre las evaluaciones 5, 20 y 100-best consultado la totalidad de términos en la grabación

Para evitar los dos inconvenientes que se tienen en los experimentos anteriores: el excesivo número de falsas alarmas, en los casos (A y B), y la posibilidad de no encontrar determinados términos de la referencia entre las N mejores ocurrencias, en los casos (B y C), se lleva a cabo un nuevo experimento en el que varían dos factores con respecto a los anteriores: términos a consultar y valor N que define el umbral.

5.2.4. Experimento D

Los términos a consultar en este experimento se reducirán, en lugar de buscar todas las palabras que contiene la grabación se hará un filtrado de los términos en función de dos consideraciones:

- aquellos cuyo número de fonemas sea bajo, por ser los más propensos a crear falsas alarmas, ya que pueden ser sílabas o sub-palabras de términos largos. Se excluyen así las consultas, a términos con más de 4 fonemas

- aquellos cuyo número de fonemas sea muy alto. Los términos muy largos poseen una tasa de pérdidas mínima más alta, su probabilidad de ocurrencia correspondiente es baja, ya que para el cálculo de la misma, se van multiplicando sucesivamente la probabilidad de estar en un determinado fonema por la de transición al siguiente, al estar hablando de valores por debajo de 1, los resultados de dicho cálculo tienden a 0 cuando el término es muy largo. Así quedarán excluidos los términos con más de 9 fonemas.

Con estas dos consideraciones, se llevará a cabo el experimento, reduciendo los términos de consulta, dejando sólo aquellos cuyo número de fonemas se encuentre entre 5 y 9.

El porcentaje de términos a evaluar por cada grabación se reduce a menos de la mitad. Como ejemplo las grabaciones en el contexto de medios: entrevistas y meteorología.

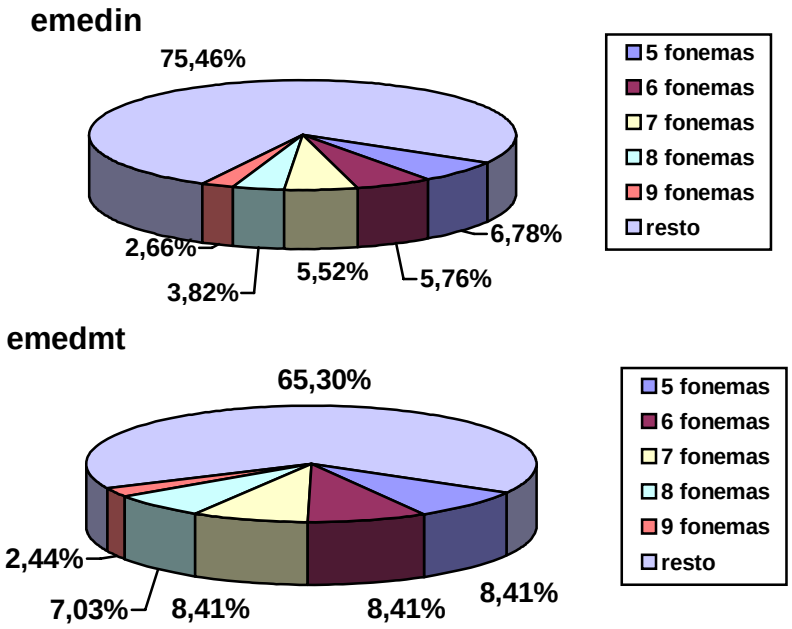
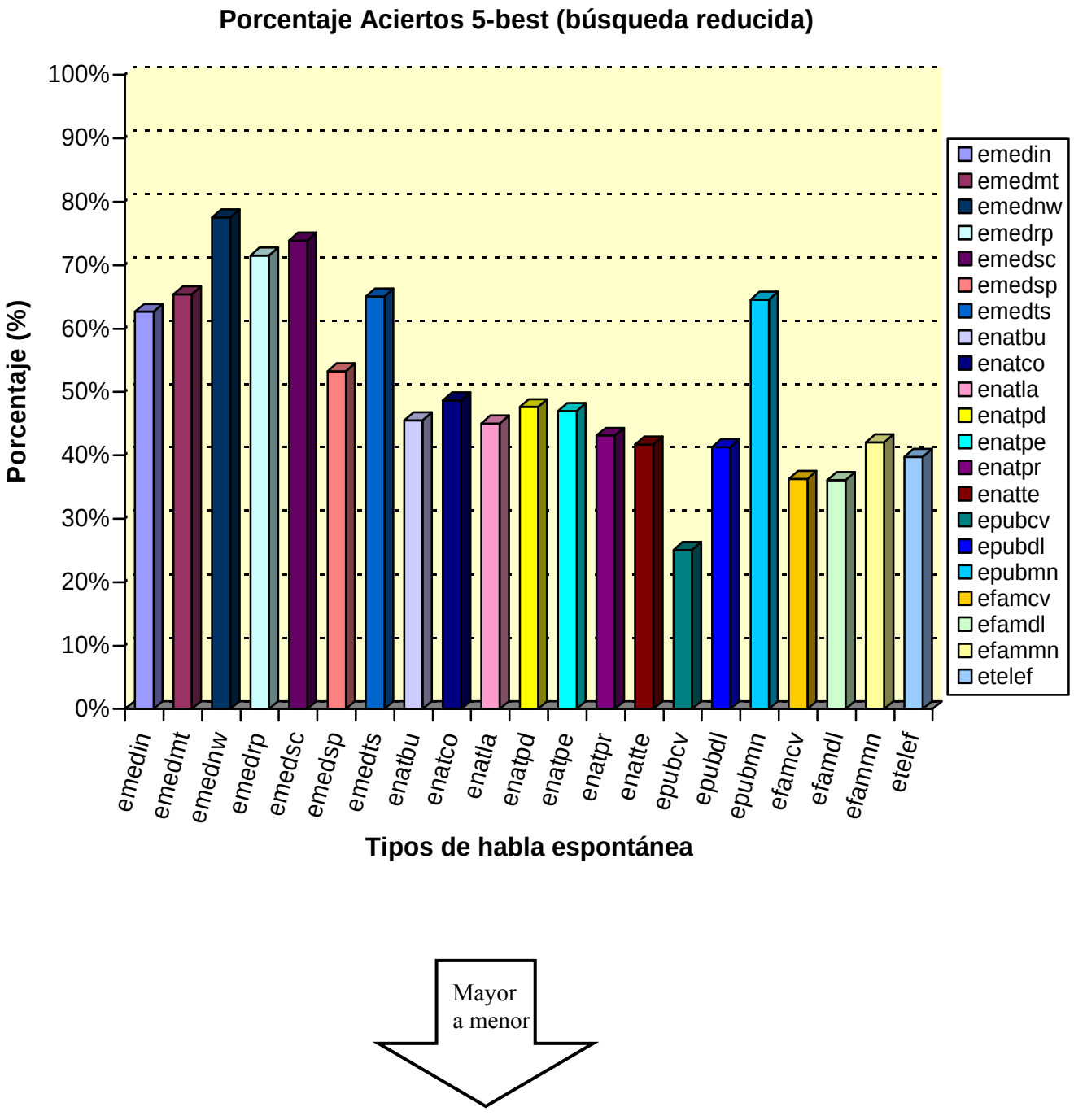


Figura 38. Porcentaje de palabras que quedan excluidas de la evaluación

Alrededor de un porcentaje de 65% en el caso de meteorología, y 75% en entrevistas pertenece a términos de 1, 2, 3, 4, o más de 9 fonemas, por lo que quedan excluidos en la evaluación, reduciendo considerablemente el número de consultas.

En el resto de tipos de habla se darán unos valores parecidos, el número de términos filtrados estará en torno al 70% de la totalidad. Pasando a ser evaluados, alrededor de un 30% del total, en cada uno de los tipos de habla espontánea.

Tras realizar la evaluación sobre los términos que quedan dentro del rango definido. Los resultados obtenidos son:



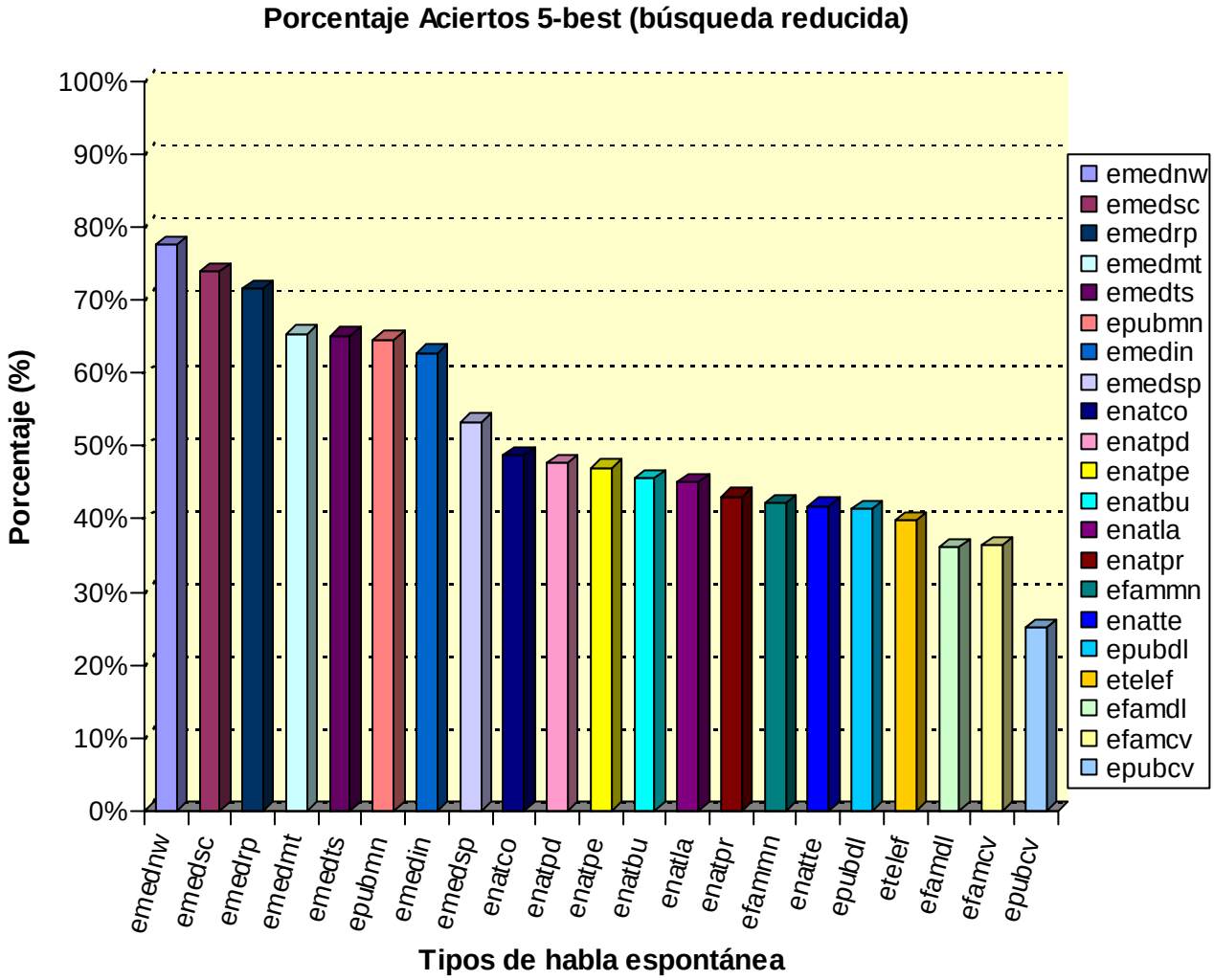


Figura 39. Porcentaje aciertos 5-best con términos cuyo número de fonemas se encuentra entre 5 y 9

Manteniendo el valor de umbral, N=5 (caso del experimento C) pero limitando las consultas exclusivamente a términos cuyo número de fonemas se encuentra entre 5 y 9, el número de detecciones correctas aumenta con respecto al experimento C, aunque estaríamos hablando de una búsqueda de menos de la mitad de los términos que contiene la grabación.

Antes de pasar a hacer un análisis exhaustivo de los resultados obtenidos con éste último experimento, se representa la nueva comparativa entre los experimentos con N=100, 20 y 5 (reducido), A, B y D respectivamente.

En la Figura 41. las barras correspondientes a los experimentos A y B mantienen sus valores, pero pasan a ser comparadas con el experimento D. Se observa que la disminución en el número de aciertos es proporcional al valor N, como se reflejó en la anterior comparativa de los experimentos A, B y C. Pero esta vez, el experimento D con el número de términos a consultar reducido, aumenta el porcentaje de aciertos en todos los tipos de habla espontánea, consiguiendo niveles cercanos al caso de N = 20 (experimento B).

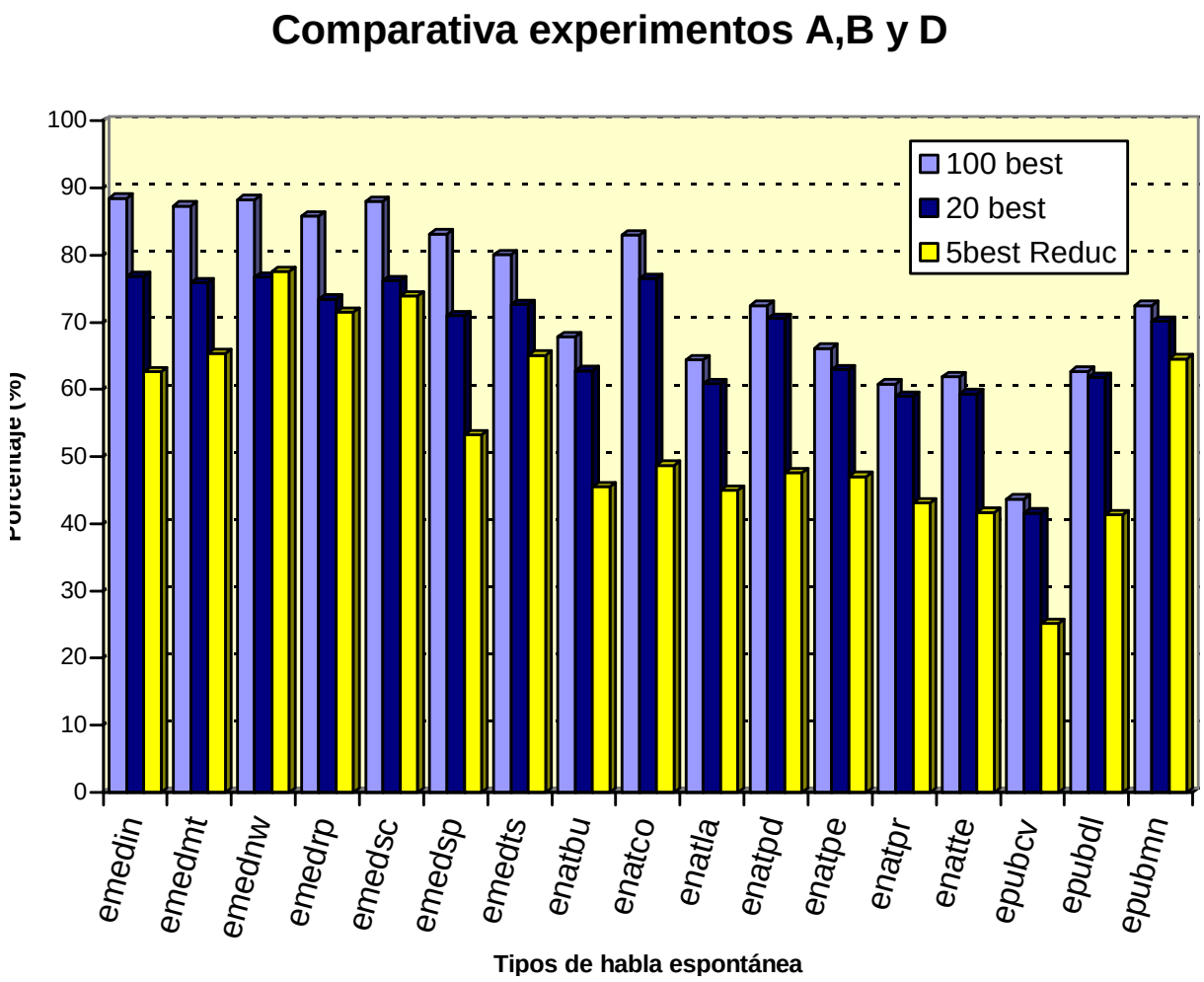


Figura 40. Comparativa de aciertos entre las evaluaciones 5, 20 y 100-best consultado la totalidad de términos en los casos de 20 y 100, pero reducida en el caso de 5.

Considerando el umbral de decisión de los experimentos A y B los resultados no son cercanos a la realidad, ya que el decisor está dando por válidas muchas detecciones que en realidad son falsas alarmas.

Si se considera el umbral N = 5 del experimento C, las falsas alarmas se reducen, pero todos los términos que aparecen con gran frecuencia en la grabación, no son detectados, ya que con este valor de N, el detector sólo dará como resultados las 5 ocurrencias del términos buscado, con mayor probabilidad.

Es por esto que la opción del valor N=5 como umbral, y el vocabulario reducido para evitar la inclusión de términos que provocarán errores de detección (experimento D) es la óptima, y la de resultados más reales.

Los resultados obtenidos ordenados descendentemente según la probabilidad de aciertos por cada tipo de habla, son los siguientes:

	Habla	Contexto	Tipo	Aciertos
emednw	Formal	Medios	Noticias	77,55%
emedsc	Formal	Medios	Ciencia	73,89%
emedrp	Formal	Medios	Informes	71,54%
emedmt	Formal	Medios	Meteorología	65,38%
emedts	Formal	Medios	'Talk shows'	65,10%
epubmn	Formal	Público	Monólogos	64,58%
emedin	Formal	Medios	Entrevistas	62,65%
emedsp	Formal	Medios	Deportes	53,27%
enatco	Formal	Natural	Conferencias	48,70%
enatpd	Formal	Natural	Debates Políticos	47,62%
enatpe	Formal	Natural	Discursos	47,01%
enatbu	Formal	Natural	Negocios	45,54%
enatla	Formal	Natural	Leyes	45,02%
enatpr	Formal	Natural	Conversaciones profesionales	43,12%
efammn	Informal	Familiar	Monólogos	42,09%
enatte	Formal	Natural	Enseñanza	41,69%
epubdl	Informal	Público	Diálogos	41,32%
etelef	Informal	Telefónico	Conversaciones telefónicas	39,76%
efamdl	Informal	Familiar	Diálogos	36,15%
efamcv	Informal	Familiar	Conversaciones	36,32%
epubcv	Informal	Público	Conversaciones	25,12%

Tabla 10. Porcentaje de aciertos en los distintos tipos de habla espontánea para 5-best

Resulta muy interesante ver que hay un amplio rango de variación entre los diferentes tipos de habla espontánea considerados: desde un 25% de aciertos en grabaciones de conversaciones en contexto público, hasta más de un 70% de aciertos en grabaciones de habla en medios como noticias o meteorología.

En general, se puede observar que para diálogos y conversaciones los resultados están entre los peores obtenidos (alrededor de un 40% de aciertos para el grupo entero). En todos estos casos parece obvio que la interacción (con frecuencia simultánea) entre los hablantes es la causa de la reducción del porcentaje de aciertos en la detección.

Otro subconjunto relacionado a ellos (que también contiene diálogos y conversaciones) es el subconjunto de conversaciones telefónicas, en este caso los resultados obtenidos son mejores, debido a la reducción de la interacción simultánea por parte de los hablantes.

En lo que se refiere a monólogos, se puede observar que en contexto familiar resultan ligeramente mejores que para diálogos y conversaciones (sobre un 42% aciertos), mientras que en contexto público resultan claramente superiores (cerca de 65% de aciertos).

Los subconjuntos mencionados en los anteriores párrafos corresponden a habla informal. Se puede observar que, con la única excepción de los monólogos en contexto público, los resultados son siempre peores que aquellos obtenidos con habla formal, ambos en contexto natural y en medios.

Comparando estos dos grandes grupos se puede observar que el habla en situaciones formales en contexto natural tiende a producir peores detecciones que aquellas provinentes de los medios, para los cuales los resultados están entre 60% y 70%.

Comparando los diferentes subconjuntos dentro de habla formal en los medios, se pueden observar diferencias interesantes. Los peores resultados obtenidos son programas deportivos, probablemente debido a la falta de cuidado del uso del lenguaje y las articulaciones exageradas así como la simultaneidad de diferentes hablantes. Ligeramente mejor son los resultados obtenidos en entrevistas, donde los solapamientos entre hablantes deben ser también muy frecuentes. Seguidamente, y con resultados intermedios, están los de programas meteorológicos y 'talk shows'. Finalmente, los mejores resultados se logran en programas de noticias, informes y programas científicos. Ha de tenerse en cuenta que en estos programas el número de solapamientos se reduce además de que el uso del lenguaje pasa a ser más cuidadoso, con supuestamente más fluidez.

Como experimento final, se han comparado los resultados de detección obtenidos con los problemas encontrados en la decodificación automática fonética en C-ORAL-ROM. Haciendo dicha comparación, se observa que existen coincidencias muy significativas. En particular, en la decodificación automática fonética se encuentran serias dificultades con características típicas de interacción en comunicación espontánea: solapamientos, número de palabras por turno, y tasa de habla.

En este caso se aplican las siguientes intuiciones:

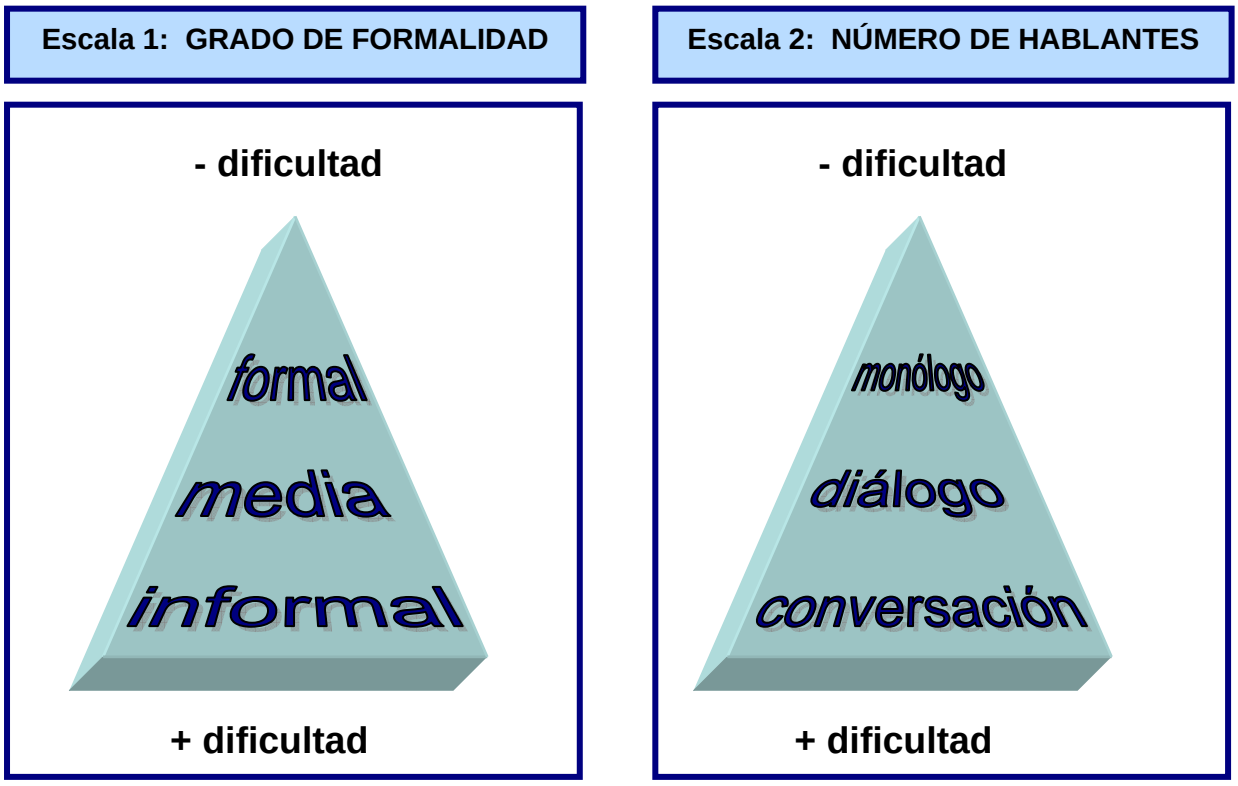


Figura 41. Escalas de dificultad en función de la formalidad y el número de hablantes

En la primera escala, el tipo más formal de habla, es más fácil de decodificar porque son seguidas convenciones más retóricas y discursivas. El habla es más predecible y la pronunciación más cuidadosa.

En la segunda escala, la dificultad se debe a la necesidad de distinguir entre los diferentes turnos y hablantes y de tener en cuenta los solapamientos. Con monólogos esta dificultad es reducida al mínimo.

Estas escalas de dificultades coinciden con los resultados obtenidos en las detecciones, ya que el punto de partida son las transcripciones fonéticas que han tenido que superar dichas dificultades. A mayor dificultad, peores resultados, se podría decir que son inversamente proporcionales.

Las decodificaciones de grabaciones que resultan más fáciles son aquellas pertenecientes a los medios, producidas por hablantes que son profesionales y combinan una buena dicción con experiencia en la elaboración dentro de las reglas lingüísticas. Así los mejores resultados en la detección se obtienen en dichas grabaciones. Según se avanza hacia habla informal con más hablantes, la complejidad aumentará, por lo que los resultados empeoran.

6. Conclusiones y trabajo futuro

6.1. Conclusiones

Cuando se habla de grandes volúmenes de datos es conveniente hablar de acceso instantáneo. Claramente, el volumen de material disponible presenta un reto para la búsqueda y recuperación tanto rápida como precisa de la información deseada.

Los datos de audio presentan un reto importante para la recuperación de información, la búsqueda basada en texto resulta insuficiente cuando se aplica a medios tales como conversiones grabadas de teléfono, conferencias y datos de video archivados.

Muchos sistemas como se ha visto anteriormente usan una representación de lattices de audio para modelar alternativas para las hipótesis de reconocimiento. Ha sido esta representación la usada para llevar a cabo el desarrollo del proyecto.

El objetivo de diseñar una arquitectura de reconocimiento de voz altamente flexible para adaptarse a cualquier vocabulario se ha conseguido gracias al modelado de unidades fonéticas. Estas unidades, ya sean dependientes o independientes de contexto, como era el caso, proporcionan una elevada flexibilidad y permiten modelar cualquier palabra o frase del idioma castellano.

Un sistema de búsqueda fonética puede detectar ocurrencias con éxitos en términos de distintas longitudes con velocidades de búsqueda inferiores al tiempo real. El sistema presentado demuestra que la búsqueda fonética puede ser muy útil en detección de términos de habla. Sin embargo, más mejoras del rendimiento en la verificación y la fiabilidad de puntuación han de ser hechas, en particular, para los términos de búsqueda de corta longitud son requeridos para competir con sistemas que incorporan un motor LVCSR.

El sistema desarrollado permite búsquedas para vocabulario completamente abierto, evitando el problema crítico de la búsqueda de términos fuera de vocabulario, problema asociado a los sistemas de búsqueda a nivel de palabra. La característica de la búsqueda fonética para lenguajes con datos entrenados limitados, o para datos de gran escala en aplicaciones de extracción, son también áreas prometedoras en la futura investigación.

Se han llevado a cabo experimentos con distintos documentos de habla, y es importante señalar que las diferencias entre estos, provoca un cambio en las propiedades y hace que la recuperación de habla cambie. Así se ha podido comprobar con las pruebas realizadas, en las que se observa la gran diferencia entre los distintos tipos de habla espontánea.

En definitiva es realmente interesante la comparación de los resultados de detección dependiendo del tipo, variando con un considerable rango, situándose con los mejores resultados el habla en medios, seguida por habla formal en contexto natural, y monólogos informales, y finalmente diálogos y conversaciones con los peores resultados.

6.2. Trabajo Futuro

El reconocimiento de voz es un campo de investigación que tiene todavía mucho por desarrollar.

En el desarrollo del presente proyecto, se ha trabajado exclusivamente con fonemas como unidades de búsqueda para la detección de términos de consulta. La consulta es convertida a cadenas de fonos mediante uso de sus pronunciaciones y así el índice del fonema puede ser buscado.

Una de las conclusiones obtenidas es que este enfoque generará muchas falsas alarmas, especialmente en consultas cortas, que podrán ser sílabas, o en general sub-palabras de otras. Por esta razón como futuro trabajo se buscan alternativas para solucionar el inconveniente.
La opción de desarrollar antes de la etapa de detección una de indexado que use índices palabra en lugar de índices fonema, sería factible en cuanto a adaptación al sistema actual, ya que el detector implementado se ha desarrollado de forma genérica, de manera que existe la posibilidad de realizar búsquedas de palabras completas, en lugar de fonemas simples. Así los términos de búsqueda podrían ser frases o sentencias. Pero con éste método perderíamos la gran ventaja que da el sistema fonético, la de ser capaces de procesar un vocabulario completamente abierto, y volveríamos al problema de los términos fuera de vocabulario OOV.

Ante esto una solución sería combinar las dos diferentes aproximaciones, indexar ambas transcripciones de palabras y transcripciones de fonemas; durante el procesado de consulta, la información sería recuperada desde el índice de palabra para términos que estuviesen contenidos en el vocabulario, y desde el índice fonético desde términos fuera de vocabulario. Sería capaz de procesar ambos consultas híbridas, es decir, las consultas que incluyeran ambos términos. Consecuentemente, se necesitaría mezclar piezas de información recuperadas desde palabras índice y fonemas índice.

Teniendo un índice de palabra y un índice de sub-palabra, sería posible mejorar el comportamiento de recuperación del sistema mediante el uso de ambos. Se podrían valorar varias estrategias para hacer esto:

1. Combinación: Buscar ambos el índice palabra y el índice sub-palabra y combinar los resultados.
2. Vocabulario en cascada: Buscar el índice palabra para consultas dentro de vocabulario, y buscar el índice de sub-palabras para las consultas OOV.
3. Búsqueda en cascada: Buscar el índice de palabra, y si no está buscar el índice de sub-palabra.

En el primer caso, si el índice es obtenido desde las mejores hipótesis de los RAV, entonces el resultado en combinación es una simple unión del conjunto separado de resultados. Sin embargo, si los índices son obtenidos desde lattices, entonces además de hablar de unión de resultados, la recuperación puede hacerse usando puntuaciones combinadas.

Referencias

[Amir01] *A. Amir, A. Efrat, and S. Srinivasan.* **Advances in phonetic word spotting**. In Proceedings of the Tenth International Conference on Information and Knowledge Management, pages 580–582, Atlanta, Georgia, USA. 2001.

[Baeck96] *Baecker, Ronald et al.;* **"Readings in Human-Computer Interaction."** RMorgan Kaufmann Publishers. San Francisco, CA., U.S.A. 1996 1.975.

[Bak91] *J.M. Baker;* **Large Vocabulary, Speaker Adaptive Continuous Speech Recognition Research Overview at Dragon Systems**. In Proceedings of Eurospeech 91 (Geneva, Italy, September 24-26). ESCA, 1991, pp. 29-32.

[Baker75a] *Baker, J. K.;* **"Stochastic Modeling for Automatic Speech Understanding"** in D. R. Reddy, ed., Speech Recognition, New York, Academic Press. 1.975.

[Baker75b] *Baker J., The DRAGON System - An overview IEEE Transactions on Acoustics, Speech and Signal Processing, vol 23, no 1, pp. 24-29. 1975.*

[Beal90] *Beale, R.; Jackson, T;* **"Neural Computing: an introduction",** Ed. Adam Hilger, 1.990.

[Bour96] *H. Bourlard, H. Hermansky y N. Morgan; "Towards increasing speech recognition error rates". Speech Communication, vol 18, nº 3, pp 205-232. 1996.*

[Bourl93] *H. Bourlard y N. Morgan*; **"Connectionist Speech Recognition - A hybrid approach."** Kluwer Academic, 1993

[Bourl96] *H. Bourlard, H. Hermansky y N. Morgan;* ***"Towards increasing speech recognition error rates"***. Speech Communication, vol 18, nº 3, pp 205-232. 1996.

[Cheng97] *R. Chengalvarayan y L. Deng;* **Use of Generalized Dynamic Feature Parameters for Speech Recognition.** IEEE Transactions on Acoustics, Speech and Signal Processing, vol 5, nº 3, pp 232-242, Mayo 1997

[Cox88] *Cox, S. J.;* **"Hidden Markov Models for Automatic Speech Recognition: Theory and Application"**, in Br Telecom Technol J, vol. 6, Nº 2, pp. 105-115, April 1.988.

[Dell93] *Deller Jr, J.R.; Proakis, J.G., Hausen J.H.L.;* **"Discrete-Time Processing of Speech Signals"**. Ed. MacMillan. 1.993.

[Drag90] Dragon Dictate User Manual, Dragon Systems, Inc. Newton, MA. 1990.

[Gauv95] *Gauvain, J.L., L. Lamel, and M. Adda-Decker*, "**Developments in Continuous Speech Dictation using the ARPA WSJ Task,**" *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1995, Detroit, MI pp. 65-68.

[Gray84] *R. Gray*; **Vector Quantization**. IEEE Acoustics, Speech and Signal Processing Magazine, vol. 1, nº 2, pp. 4-29, abril 1984.

[Hasan90] *H. Hasan*; **"Reconocimiento de 1000 Palabras Independiente del Locutor Mediante Modelos Ocultos de Markov"**. Tesis Doctoral. ETSITM-UPM, julio 1990.

[Hoch94] *M. Hochberg, T. Robinson y S. Renals*; **"Large Vocabulary Continuous Speech Recognition using a Hybrid Connectionist HMM System."** Proc. of the International Conference on Spoken Language Processing (ICSLP), 1994, pp. 1499-1502. 1994.

[Holt98] *T. Holter*; **Maximul Likelihood Modelling of Pronunciation in Automatic Speech Recognition** PhD. Thesis. Department of Telecommunications. Signal Processing Group. Norwegian University of Science and Technology, 1998.

[Huang01] *X. Huang, A. Acero, H. Hsiao-Wuen*; "**Spoken Language Processing**", in Dr. R. Reddy, ed., New Jersey, Prentice Hall PTR, pp. 377-664, 2.001.

[Huang89] *X.D. Huang, H.W. Hon y K.F. Lee*; **"Large-Vocabulary speaker independent CSR with semi-continuous HMMs"** Proc. of the European Conference on Speech Communication and Technology (Eurospeech),

[Huang90a] *B.H. Juang y L. R. Rabiner*; **"The segmental k-means algorithm for estimating parameters of hidden Markov model**s" IEEE Transactions on Speech and Audio Processing, Vol. 38, pp.1639--1641, 1990.

[Huang90b] *X.D. Huang, Y. Ariki y M.A. Jack;* **"Hidden Markov Models for speech recognition."** Edinburgh University Press, 1990

[Ilan07] *Ilana Bromberg2, Qiang Fu1 et. al* ; **Detection-Based ASR in the Automatic Speech Attribute Transcription Project** In Proc**.** InterSpeech 2007

[James94] *James, D. A. y Young, S. J.,* **"A Fast Lattice-based Approach to Vocabulary IndependentWordspotting,"** in Proc. ICASSP, 1994.

[Jelin76] *Jelinek, F.*; "**Continuous Speech Recognition by Statistical Methods"**, in Proc. of the IEEE, vol. 64, pp. 532-556, April 1.976.

[Jiny05] *Jinyu Li, Yu Tsao, y Chin-Hui Lee;* **A study on knowledge source integration for candidate rescoring in automatic speech recognition.** *In Proc. ICASSP-05.*

[Joel07] *Joel Pinto, Andrew Lovitt, Hynek Hermansky;* **"Exploiting Phoneme Similarities in Hybrid HMM-ANN Keyword Spotting"** In Proc. InterSpeech 2007.

[Jones96] *G. J. F. Jones, J. T. Foote, K. Sparck Jones, and S. J. Young.* **Retrieving spoken documents by combining multiple index sources.** In *Proc. SIGIR 96*, pages 30–38, Zürich, August. 1996

[Kirch98] *Kirchhoff, K.,* **"Combining Articulatory and Acoustic Information for Speech Recognition in Noisy and Reverberant Environments."** in Proc. ICSLP, 1998.

[Lang90] *K.J. Lang, A. Waibel y G.E. Hinton*; **"A time-delay Neural Network Architecture for Isolated Word Recognition."** Neural Networks, 3(1), pp 23-43. 1990.

[Lee03] *Lee, C.-H.;* **"On Automatic Speech Recognition at the Dawn of the 21st Century,"** *IEICE Trans. Inf. & Syst.*, pp. 377–396, 2003.

[Lee04] *Lee, C.-H.*; **"From Knowledge-ignorant to Knowledgerich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition,"** in *Proc. ICSLP*, 2004.

[Lee89] *Lee K.F. "**Hidden Markov Models: Past, Present and Future"**.* Proc. of the European Conference on Speech Communication and Technology (Eurospeech), 1989, pp. 148-155. 1989.

[Logan02] *B. Logan, P. Moreno, and O. Deshmukh.* **Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio**. *In Proc. HLT. 2002.*

[LogVan02] *B. Logan and JM Van Thong.* **Confusion-based query expansion for OOV words in spoken document retrieval.** In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Denver, Colorado, USA. 2002.

[Mark96] *J. MarkowitzBaker;* **"Using Speech Recognition."** Prentice Hall, 1996.

[Menén94] *X. Menéndez Pidal Sendrail; "***Arquitecturas Neuronales y su Integración con Algoritmos de Programación Dinámica en Tareas de Reconocimiento de Habla".** Tesis Doctoral. ETSIT-UPM, 1994

[Meng07] *Sha Meng, Peng Yu, Frank Seide, y Jia Liu;* **A study of lattice-based spoken term detection for Chinese spontaneous speech.** Department of Electronic Engineering, Tsinghua University. Beiging 2007.

[Morg91] *Morgan D. P. y Scofield C.L.* **"Neural Networks and Speech Processing"** Kluwer Academic Publishers, 1991.

[Morg95] *N. Morgan y H. Bourlard*; **"Continuous Speech REcognition: An introduction to the hybrid HMM/connectionist Approach"**. IEEE Signal Processing magazine, volume 12, number 3, pages 24-42, May 1995.

[Nadeu97] *Nadeu, P. Pachès-Leal y Biing-Hwang Juang*; "**Filtering the time sequences of spectral parameters for speech recognition"** Speech Communication , vol 22, nº 4, Septiembre 1997, pp 315-332. 1997.

[Rabin86] *L.R. Rabiner, B. H. Juang* ; **An introduction to Hidden Markov Models. IEEE Acoustics, Speech and Signal Processing Magazine**, enero 1986, pp 4-15, 1986.

[Rabin89] *L.R. Rabiner.* **A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition**. Proceedings of the IEEE 77,2 .February 1989 257-285.

[Rabin93] *Rabiner, L.R., Juang, Biing-Hwan; "***Fundamentals of Speech Recognition."** BPrentice Hall PTR. New Jerey, U.S.A. 1993.

[Renal94] *S. Renals, N. Morgan, H. Bourlard, M. Cohen y H. Franco*; "**Connectionist probability estimators in HMM speech Recognition."** IEEE Transactions on Speech and Audio Processing, 1994, vol 12 (1), pp 161-171. 1994.

[Robin91] *T. Robinson y F. Fallside;* A **Recurrent error propagation network speech recognition system.** Computer Speech and Language, 5, pp. 259-274, 1991.

[Robin95] *T. Robinson, M. Hochberg y S.Renals*; Th**e use of recurrent networks in continuous speech recognition, en Automatic Speech and Speaker Recognition** - Advanced Topics, Capítulo 19. Editores: C H Lee, K K Paliwal y F.K. Soong Kluwer Academic Publishers, 1995

[Rose90] *Rose, R. C. y Paul, D. B.*; **"A Hidden Markov Model Based Keyword Recognition System,"** in Proc. ICASSP, 1990.

[Sabat06] *Sabato Marco Siniscalchi, Jinyu Li, and Chin-Hui Lee*. **A study on lattice rescoring with knowledge scores for automatic speech recognition.** In *Proc. InterSpeech-06*

[Soong86] F. *K. Soong y A. E. Rosenberg;* ***"On the use of Instantaneous and Transitional Spectral Information in Speaker Recognition."*** Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1986, pp. 877-880, 1986.

[Soudo86] *S. Soudoplatoff;* **"Markov Modeling of Continuous Parameters in Speech Recognition."** Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1986, pp. 45-48. 1986.

[Sproat07] *M. Saraclar, R. Sproat;* **Lattice-Based Search for Spoken Utterance Retrieval.** AT&T Labs – Research 2006

[Srilm] *SRI International's Speech Technology and Research* (STAR) Laboratory. Speech technology organization. http://www.speech.sri.com

[Srin00] *S. Srinivasan and D. Petkovic*. **Phonetic confusion matrix based spoken document retrieval.** In *Proceedings of the 23rd Annual International ACM SIGIR. Conference on Research and Development in Information Retrieval*, pages 81–87. 2000.

[Szök05] *Szöke, I., Schwarz, P., Matejka, P., Burget, L., Karafiát, M., Fapso, M. y Cernocky, J.,* **"Comparison of Keyword Spotting Approaches for Informal Continuous Speech,"** in Proc. InterSpeech, 2005.

[Toled08] *D. T. Toledano;* **"Lattices y WordSpotting**". Área de Tratamiento de Voz y Señales. Escuela Politécnica Superior - UAM, Madrid. 2008.

[Vins68] *Vinstyuk, T. K.;* **"Speech Discrimination by Dynamic Programming"**; Kibernetika (Cybernetics), 4(2), pp. 81-88, January-February, 1.968.

[Wech98] *M. Wechsler, E. Munteanu, and P. Scäuble*. **New techniques for open-vocabulary spoken document retrieval.** In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 20–27, Melbourne, Australia. 1998.

[Witbr97] *M.Witbrock and A. Hauptmann*. **Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents**. In *2nd ACM International Conference on Digital Libraries (DL'97)*, pages 30–35, Philadelphia, PA, July. 1997.

[Young02] *S. Young et al.*, **The HTK Book** (for HTK version 3.2.1), 2002. Available on http://htk.eng.cam.ac.uk/.

[Zhou06] *Z. Y. Zhou, P. Yu, C. Chelba, F. Seide*, **Towards Spoken Document Retrieval for the Internet: Lattice Indexing For Large-Scale Web-search Architectures**, *Proc. HLT'2006*, New York, 2006.

Glosario

Curva DET (compensación por error de detección)
Trazo gráfico de las tasas de error medidas. Por lo general, las curvas DET trazan las tasas de error de decisión (tasa de falso rechazo vs. tasa de falsa aceptación).

DARPA (*Defense Advanced Research Projects Agency*)
Agencia de proyectos de investigación avanzada de defensa.

DTW (*Dynamic Time Warping*)
Alineamiento Temporal Dinámico.

Front-end
Parte del sistema que interactúa con el usuario. En un sintetizador de habla, el front-end se refiere a la parte del sistema que convierte la entrada del texto en una representación fonética.

HMM (*Hidden Markov Model*)
Modelo Oculto de Markov.

HTK (*Hidden Markov Model Toolkit*)
Conjunto de herramientas para la creación y tratamiento de HMMs, diseñado por la universidad de Cambridge.

KWS (*Key-Word Spotting*)
Sistema Wordspotting sobre palabras clave

Lattice
Representación compacta de una serie de hipotéticas posibles transcripciones para un archivo de audio concreto. Es muy utilizado en sistemas de reconocimiento de habla.

Lattice-Tool
lHerramienta que opera sobre lattices palabra o fonema en formato pfsg o HTK Standar Lattices Format (SLF).

LPC (*Linear Predictive Coding*)
Codificción Lineal Predictiva

LVCSR (*Large Vocabulary Continuos Speech Recognition*)
Vocabulario amplio usado en sistemas de Reconocimiento de Habla Continua.

MFCCs (*Mel-Frequency Cepstral Coefficients*)
Coeficientes cepstrales en escala de frecuencias Mel.

MLF (*Master Label Format*)
Formato de HTK para archivos que contienen transcripciones.

MLLR (*Maximum Likelihood Linear Regresión*)
Método de adaptación de modelos independientes de locutor a los distintos locutores mediante transformaciones lineales.

MLPs (*Multi-Layer Perceptrons*)
Red neuronal multi-capa.

NIST (*National Institute of Standars 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.

OOV (*Out of Vocabulary*)
Palabras perdidas en un sistema de vocabulario de un Reconocedor de Voz. Término fuera de vocabulario.

RAV
Reconocimiento Automática de Voz

Reconocimiento del habla
Tecnología que permite que una máquina reconozca las palabras pronunciadas. El reconocimiento del habla no es una tecnología biométrica.

ROC (Característica de funcionamiento del receptor)
Método para mostrar el rendimiento de precisión medida de un sistema biométrico. La característica ROC en una verificación compara la tasa de falsa aceptación con la tasa de verificación.

SLF (*Estándar Lattice Format*)
Formato de HTK para archivos que contienen *lattices.*
SphinxConjunto de herramientas para el tratamiento de voz diseñado por la universidad de Carnegie-Mellon

SRI (Speech Technology and Research Laboratory)
Laboratorio de investigación y tecnología del habla.

SRILM
Toolkit para construcción y aplicación de modelos de lenguaje estadísticos principalmente para uso en reconocimiento de habla, etiquetado estadístico y segmentación.

SVMs (Support Vector Machines)
Método de reconocimiento de patrones.

VSM (Vector-space modelling)
Modelado de espacio vectorial

Wordspotting
Nueva línea estratégica de I+D, que consiste en determinar si una palabra clave aparece en una grabación y en que instante de tiempo.

Anexos

A Formatos de lattice

1. Formato pfsg

Formato de programa para Decipher (TM) Probabilistic Finite-State Grammars

Sinopsis

```
name name
nodes N w1 ... wN
initial i
final f
transitions T
n1 n2 p
...
```

Dónde `n1` y `n2` son los nodos entre los que se produce la transición y `p` la probabilidad de que se produzca la misma.

Descripción

Probabilistic finite-state grammars (PFSG) es un método para automatizar estados finitos o traductores usados por el reconocedor Decipher (TM) de SRI (Stanford Research Institute). El formato PFSG emite palabras (salidas) de los nodos, no de los arcos. Ciertos tipos de lenguajes manipulados por SRILM pueden ser traducidos a PFSG por uso directo de un reconocedor.

Desde que es normal y relativamente fácil realizar conversiones entre diferentes tipos de representación de redes de estado finito, PFSG puede servir como un formato intermedio para la generación de otros de estado finito. Por ejemplo, puede ser convertido a formato AT&T.

2. Formato wlat

Formato de programa para lattices posteriors de SRILM

Sinopsis

```
Lattices de palabra:
version 2
name s
initial i
final f
node n w a p n1 p1 n2 p2 ...
```

Donde `n` es el número de nodo, `w` la palabra, `a` el estado, `p` la probabilidad, $n_1 \ldots n_n$ los nodos vecinos de `n`, y $p_1 \ldots p_n$ `la probabilidad de que se produzca la transición n->` n_n.

Descripción

Los lattices de palabra posterior son generados por alineamiento de las n mejores hipótesis con nbest-lattice, o por alineamiento de lattices PFSG o HTK mediante el lattice-tool. Ellos compactamente codifican las posibles secuencias de palabra hipotéticas y sus probabilidades posteriores.

Un lattice de palabra es un grafo dirigido y parcialmente ordenado con nodos representando las hipótesis de palabra. Los nodos son identificados por enteros no negativos. El formato del programa especifica el nodo inicial `i`, y el final `f`, y cualquier número adicional de nodos n. Para cada nodo n la información siguiente asociada es dada en la misma línea, la palabra identidad w (NULL para los nodos inicial y final), la posición de alineamiento a (valores idénticos en este campo identifican hipótesis que ocurren en la misma posición), y la probabilidad posterior p. Siguiendo a estos valores, se encuentran cero o más transiciones que suceden a los nodos especificados, cada una dada por el índice del nodo n_i y la probabilidad de transición posterior. En un lattice de palabra normalizado la posterior transición p_i se multiplica a la del nodo posterior p.

Puesto que los lattices existentes son lattices fonema, la herramienta Lattice-Tool es usada considerando cada uno de los fonemas como una palabra, a la hora de introducirlos en el sistema.

PRESUPUESTO

El cálculo del presupuesto está realizado en base al coste del material, aplicando a éste los gastos generales y beneficio industrial, más la mano de obra utilizada.

1) Ejecución Material

- Ordenador personal..... 1.400 €
- Sofware ..... 700 €
- Impresora láser..... 270 €
- Material de oficina ..... 130 €

Total de ejecución material ..... 2.500 €

2) Gastos generales

- 16 % sobre Ejecución Material ..... 400 €

3) Beneficio Industrial

- 6 % sobre Ejecución Material ..... 150 €

4) Honorarios Proyecto

800 horas a 15 € / hora ..... 12000 €

5) Material fungible

- Gastos de impresión..... 50 €
- Encuadernación..... 180 €

Total material fungible ..... 230 €

6) Subtotal del presupuesto

Subtotal Presupuesto ..... 15280 €

7) I.V.A. aplicable

- 16% Subtotal Presupuesto ..... 2444,8 €

8) Total presupuesto

Total Presupuesto ..... 17724,8 €

Madrid, Octubre de 2008
El Ingeniero Jefe de Proyecto

Fdo. Verónica Peña García
Ingeniero de Telecomunicación

PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un reconocedor de palabras clave en conversaciones espontáneas. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

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

Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la

misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contrata" y anteriormente llamado "Presupuesto de Ejecución Material" que hoy designa otro concepto.

Condiciones particulares

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

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

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.