Un Método para la Identificación Automática del Lenguaje Hablado Basado en Características Suprasegmentales

por

Ana Lilia Reyes Herrera

Tesis sometida como requisito parcial para obtener
el grado de

Doctor en Ciencias en el área de Ciencias Computacionales

en el
Instituto Nacional de Astrofísica,
Óptica y Electrónica
Noviembre 2007
Tonantzintla, Puebla

Supervisada por:

Dr. Luis Villaseñor Pineda
Investigador Titular del INAOE

RESUMEN

La identificación automática del lenguaje hablado consiste en determinar por medios computacionales el idioma de quien habla basándose sólo en una muestra de voz sin considerar al hablante o lo que está diciendo. La identificación automática del lenguaje hablado tiene muy diversas aplicaciones. Por ejemplo, las compañías de teléfono quisieran tener un identificador de idiomas eficiente para los hablantes extranjeros y así poder reenviar sus llamadas a un operador capaz de comprender dicho idioma; un sistema de traducción multilingüe con más de dos o tres idiomas necesita un sistema de identificación de lenguaje hablado como primer paso para seleccionar el sistema de traducción apropiado; además, los gobiernos alrededor del mundo han estado interesados por mucho tiempo en un sistema identificador de idiomas para propósitos de monitoreo y seguridad.

En la actualidad, los mejores sistemas automáticos de identificación de lenguaje hablado utilizan información lingüística para la tipificación del idioma, es decir, dependen de la representación fonética de la señal de voz. A pesar de los buenos resultados de estos métodos, se depende de un estudio lingüístico previo para cada uno de los lenguajes a identificar; y en base a este estudio, se pueden establecer los valores de los parámetros de identificación. La presente investigación doctoral propone un método para la identificación del lenguaje hablado basado en información extraída directamente de la señal acústica sin requerir de un módulo de tratamiento fonético. Dos nuevos métodos son propuestos para la extracción de características acústicas directamente de la señal de voz aplicada a la identificación de los idiomas. Uno basado en el uso de la transformada de Fourier, específicamente el uso de los coeficientes cepstrales de frecuencia Mel (MFCC por sus siglas en inglés) y un segundo método basado en la transformada Wavelet. Los resultados de aplicar estos dos nuevos métodos superaron lo reportado en el estado del arte. Estos resultados son muy alentadores, creando la posibilidad de obtener sistemas automáticos de identificación de lenguas con escasos recursos lingüísticos, como es el caso de la mayoría de las lenguas indígenas de México.

ABSTRACT

The Automatic Language Identification problem consists on recognizing a language based on a sample of speech from an unknown speaker by computational means. The Language Identification has several applications. For example, the telephone companies want to have an efficient identifier of languages for foreign speakers and then to be able to send calls to a capable user to understand the language; a system of multilingual translation with more than two or three languages needs an automatic language identification system as first step to select the system of appropriate translation; And of course, the governments around the world have been interested for a long time in a language identification system for monitoring and security purposes.

At present, the best automatic language identification systems use the phonotactic content of each language, that is, they depend on the phonetic representation of the speech signal. In spite of good results of these methods, they depend on a prior linguistic study for each one of the languages to be identified. Those methods are based on the segmentation of the speech signal in phonemes, and on the use of language models – which capture all possible combinations of phonemes from a particular language– to determine the language. The present doctoral investigation proposes the creation of a method for language identification based on information extracted directly of the speech signal without requiring a phonetic processing module. The results of the investigation are two new methods for the extraction of characteristic acoustics directly of the speech signal applied to the language identification. The first one is based on the use of the Fourier transform, specifically on the use of the Mel Frequency Cepstral Coefficients (MFCC) and the second method is based on the Wavelet transform. Results of these two new methods achieve better results than the state of the art methods. And finally, these new methods were applied to the identification of Mexican native languages, since these native languages do not have phonetics transcription, with very good results. With these results we create the possibility to obtain an automatic language identification of México's native languages.

AGRADECIMIENTOS

Mi total agradecimiento a mi asesor el Dr. Luis Villaseñor Pineda, porque siempre creyó en mi y mis ideas. Me enseño a tener confianza en mi misma y a ser autosuficiente, a cómo y dónde encontrar las respuestas. Sobre todo por su apoyo incondicional durante todo el proyecto de investigación. Por los conocimientos trasmitidos y por su paciencia en muchos aspectos. ¡¡MIL GRACIAS!!

Además quiero agradecer al Dr. Manuel Montes y Gómez, por sus valiosas discusiones y comentarios durante el desarrollo del proyecto.

Al Dr. Guillaume Gravier por su gran interés en mi trabajo, por su apoyo y conocimientos en la revisión de mi tesis. Sobre todo por sus valiosos comentarios.

Al Dr. Jesús Ariel Carrasco Ochoa por su apoyo incondicional desde que ingrese al doctorado y por sus críticas constructivas acerca de mi trabajo de investigación.

Estoy en deuda con el Dr. Alfonso Fernández Vázquez, por su gran apoyo y conocimientos hacia mi trabajo en la parte de wavelet. Siempre con una vista hacia lo electrónico que completo el panorama de mi trabajo.

Quiero agradecer también a los doctores René Cumplido Parra, Miguel Arias Estrada, Carlos Reyes García por su valiosa participación como miembros del jurado y sus acertadas observaciones durante todo el proyecto de investigación.

Al Consejo Nacional de Ciencia y Tecnología por la ayuda económica otorgada a través de la beca de postgrado 184660. De igual manera al Instituto Nacional de Astrofísica, Óptica y Electrónica por los apoyos y facilidades que me brindaron durante mi estancia en el postgrado. En especial a la dirección de Docencia. Y Sobre todo a la academia de Ciencias computacionales por darme la oportunidad de realizar mis estudios.

También quiero agradecer a mis compañeros de generación, Rita, Carmen, Rafa y Jorge; así como a mis compañeros del cubículo 8310, Irene, Griselda, Toño, Rene y Milton por ser una buena influencia y hacer agradable mi estancia en el instituto.

DEDICATORIA

A mi esposo Alberto: Por el apoyo y compresión que siempre me brindas y con eso me haces CRECER y cumplir mis sueños.

A mi pequeño Diego: El motor de mi vida.

A mis padres Alicia† y Emilio†: Esto es un logro más de lo que me enseñaron. Gracias por darme las alas y el valor para volar.

CONTENIDO

Página

LISTA DE FIGURAS

Página

LISTA DE TABLAS

Página

CAPÍTULO 1

INTRODUCCIÓN

La identificación del lenguaje hablado consiste en determinar el idioma de quien habla basándose sólo en una muestra de voz sin considerar al hablante y lo que está diciendo. De acuerdo a esto, la identificación automática del lenguaje hablado es el proceso por el cual el lenguaje (idioma) de una muestra de señal de voz digitalizada es reconocido por una computadora.

Las aplicaciones de la identificación del lenguaje se pueden agrupar en dos grandes categorías: pre-procesamiento para sistemas y pre-procesamiento para humanos. Para el primer tipo de aplicación, considere el lobby de un hotel o un aeropuerto internacional donde se emplean sistemas para la recuperación de información de los viajes con portavoces multilingües. Si el modo de entrada al sistema es por medio de la voz, entonces el sistema debe ser capaz de determinar el lenguaje (idioma) de la señal de voz mientras la persona está hablando, antes de reconocer los comandos en sí. Determinar el lenguaje durante el reconocimiento debería requerir muchos reconocedores del habla (uno por cada lenguaje) ejecutándose en paralelo. Entonces cientos de personas hablando diferentes lenguas, necesitarían estar disponibles al mismo tiempo, por lo que el costo de los requerimientos de hardware para trabajar en tiempo real puede ser prohibitivo. Alternativamente, los sistemas de identificación del lenguaje hablado (LID por sus siglas en inglés, Language IDentification) podrían ser ejecutados antes de los sistemas de

reconocimiento del habla. En este caso, los sistemas LID deberían ejecutarse rápidamente. Después de identificar el idioma se podría cargar y correr el reconocedor del habla apropiado para el idioma. Otro ejemplo, es el de un sistema de traducción multilingüe por medio de la voz con dos o más idiomas, el cual necesita un sistema de identificación de lenguaje hablado como primer paso de su sistema para poder reenviar la voz al sistema de traducción apropiado.

En el segundo tipo de aplicación; las compañías de teléfono quisieran tener un eficiente identificador de idiomas para los hablantes extranjeros y poder enviar sus llamadas a los operadores que pueden hablar su idioma. Lo que es más, un identificador de lenguaje hablado puede salvar vidas. En Estados Unidos de Norteamérica, existen muchos casos reportados al 911 (número telefónico de emergencias) los cuales no fueron resueltos rápidamente porque el operador no pudo reconocer el idioma de los hablantes extranjeros. Ya que sin importar que se trate de un hablante bilingüe, la lengua materna es utilizada inconscientemente en condiciones de alto estrés. Por otro lado, en el caso legal en México y EE.UU. se han cometido grandes errores al enjuiciar a personas que no hablan el idioma español. Por ejemplo, durante los años ochentas, se encarceló a dos indígenas en el estado de Oregon, USA., por que no sabían hablar español e inglés. El primer caso fue el de Santiago Ventura Morales un inmigrante mixteco quien fue arrestado en 1986, acusándolo de un asesinato que nunca cometió. Fue condenado a cadena perpetua, sin haber tomado en cuenta su idioma indígena, ya que nunca entendió que es lo que estaba pasando. La acusación se basó en los testimonios de los indígenas que supuestamente hablaban la misma lengua. Fue encarcelado por más de cuatro años bajo el cargo de homicidio antes de que su sentencia fuera anulada. Véase el propio testimonio de Ventura Morales [1], así como el análisis de Lourdes De León [2] sobre el papel de las incongruencias lingüísticas durante el juicio. El segundo caso fue el de Adolfo Ruiz Álvarez [3], indígena *triqui*, a quien lo detuvo la policía y fue interrogado en inglés, después en español, llegando a la conclusión de que él estaba loco por hablar otra lengua. Lo llevaron a un hospital psiquiátrico en 1990 y fue drogado por dos años antes de ser liberado en Junio de 1992. Si existiera un identificador de lenguas indígenas de México este tipo de problemas no existirían. Estos son sólo algunos de los ejemplos que muestran el interés en la identificación automática del lenguaje.

Actualmente los sistemas automáticos de identificación de lenguaje varían en sus enfoques. Los sistemas con mejores resultados son los que basan la identificación del lenguaje hablado en las características lingüísticas propias de cada lenguaje. Este tipo de sistemas trabajan básicamente en dos pasos. Primero, se segmenta la señal acústica en sus correspondientes fonemas, de igual forma que un reconocedor del habla [4]. Posteriormente, la secuencia de fonemas se mide contra uno o varios modelos de lenguaje [5][6]. Un modelo de lenguaje captura las combinaciones posibles de los fonemas de un lenguaje específico. Así, el modelo de lenguaje que mejor cubra la muestra en cuestión será considerado el lenguaje de la muestra. Desafortunadamente, este enfoque requiere de recursos lingüísticos para realizar estas dos tareas. Por un lado, para reconocer los fonemas se necesita de grandes cantidades de datos (grabaciones) previamente etiquetados. Cada grabación debe etiquetarse manualmente a nivel fonético, para después calcular el modelo acústico de cada fonema del lenguaje en cuestión. Por otro lado, para construir el modelo de lenguaje es necesario recopilar grandes cantidades de texto y voz. Con estos datos se calculan las probabilidades de las diversas secuencias de los fonemas. Así, este enfoque depende de una gran cantidad de trabajo previo orientado a caracterizar el lenguaje ó lenguajes a identificar. Como es de imaginar, bajo este esquema, agregar un nuevo lenguaje al sistema de identificación es muy costoso.

Desafortunadamente, para las lenguas que no tienen trascripción a texto (ni trascripción fonética), como muchas de las lenguas indígenas de México; este enfoque no es viable. Además, existe la necesidad real de crear un sistema de asistencia lingüística a emigrantes indígenas monolingües de México [7]; el cual permita la asistencia a los hablantes indígenas monolingües que se encuentran en la necesidad de interactuar con las autoridades, en México o en EE.UU., cuando no se conoce su lengua o procedencia, como en un caso de detención o tratamiento médico de emergencia [www.cdi.gob.mx].

Otro enfoque, que trata de eliminar el etiquetado fonético es aquel que usa un sistema de segmentación automática sobre las grabaciones de entrenamiento [8][9][10]. En este caso, no reconocemos estrictamente fonemas sino “tokens” (posibles secuencias de fonemas). De esta manera, caracterizamos un lenguaje a partir de estos “tokens”, eliminando el etiquetado manual. Pero este enfoque se mantiene muy cercano al anterior

enfoque, pues la identificación del lenguaje hablado se basa, en un reconocedor de "tokens" junto con modelos de lenguaje basados en dichos "tokens".

Por último, un tercer enfoque trata de explotar directamente la señal acústica para identificar el lenguaje hablado. En este caso, se trata de explotar las características de la señal acústica, como la prosodia, la entonación, el ritmo, etc. Hasta este momento, este enfoque no obtiene resultados comparables a los obtenidos con los enfoques anteriores sin embargo no depende de ningún estudio lingüístico, ya sea para la construcción del reconocedor como para la creación de los modelos de lenguaje.

1.1 NATURALEZA DEL PROBLEMA

De acuerdo a los lingüistas las diferencias entre los idiomas son múltiples y enormes. A pesar de que esas diferencias son evidentes a diferentes niveles (léxico, sintáctico, de articulación, ritmo, prosodia, etc.) la identificación del lenguaje hablado es aún un reto.

De acuerdo al estado del arte, podemos decir, que existen dos grandes áreas para la identificación automática del lenguaje hablado, una que se basa en la representación fonética de la señal de voz[1], es decir en la segmentación de fonemas y sus subsecuentes procesos, y otra en donde sólo se utilizan las características acústicas de la señal de voz para la identificación de los idiomas. Este último, hasta nuestros días, no ha tenido resultados comparables a los del primer enfoque.

Los sonidos que se generan cuando hablamos pueden ser descritos en términos de un conjunto de unidades lingüísticas abstractas llamadas "fonemas". Cada fonema corresponde a una única configuración del tracto vocal. Diferentes combinaciones de fonemas constituyen diferentes palabras. Por lo que, diferentes palabras están formadas

[1] En esta tesis se utiliza en forma equivalente el término "señal de voz" a "señal del habla".

de diferentes secuencias de fonemas que corresponden a diferentes movimientos del tracto vocal. Y mas aún, diferentes combinaciones de palabras producen un gran número de oraciones que contienen toda la información que uno quiere transmitir.

La fonética analiza los fonemas en términos de las características lingüísticas de esos sonidos y los relaciona con la posición y movimientos de las articulaciones. Los fonemas pueden ser clasificados por:

- Modo de articulación, el cual describe diferentes fonemas de acuerdo a la forma que el tracto vocal restringe el aire que sale de los pulmones. Los idiomas tienen diferentes categorías de fonemas: nasal, vocal, fricativa, entre otros.
- Características de los formantes, las consonantes pueden ser formadas dependiendo si las cuerdas vocales están o no están involucradas en su producción.
- Lugar donde se hace la articulación, es decir, el lugar donde se estrecha el tracto vocal durante la pronunciación.

Diferentes combinaciones de modo de articulación, formantes y lugar de articulación resultan en diferentes fonemas.

Generalmente un lenguaje no usa todas las posibles combinaciones de formantes, de articulación y de lugar de articulación. Es decir, un lenguaje usa sólo un subconjunto de todos los posibles fonemas que el ser humano puede producir. Así diferentes lenguajes tienen diferentes fonemas, por ejemplo el francés tiene 15 vocales mientras que el español sólo tiene 5, el alemán tiene vocales unidas mientras que en el inglés no están permitidas, etc. Por otro lado, la manera en que dichos fonemas se unen para formar una palabra debe respetar ciertas reglas propias de cada lenguaje, de la misma forma que cada lenguaje tiene su propia gramática. Los métodos tradicionales de identificación automática del lenguaje aprovechan estas características (los fonemas de un lenguaje y sus combinaciones permitidas) para reconocer un lenguaje.

Sin embargo, cuando hablamos no sólo generamos fonemas, también existen otros aportes de información dentro de la señal acústica tal como la entonación, la duración, el

acento y el ritmo. Estos elementos comúnmente son agrupados por los lingüistas bajo el concepto de prosodia. Este tipo de información también es distintiva en los lenguajes humanos –ver capítulo 3-. Por ejemplo, el ritmo, que es la pauta de tensión formada en el mismo por la combinación de sílabas tónicas y átonas, largas y breves 0; es un elemento distintivo entre los idiomas, porque como no todas las lenguas hacen el mismo uso de las sílabas largas y breves, y de las tónicas y átonas, habrá distintos tipos de ritmos; lo que nos lleva a tener un elemento distintivo entre los idiomas; ya que el ritmo es uno de los prosodemas o fonemas prosódicos (o suprasegmentales) más característicos de una lengua. Los ritmos más importantes son el acentual y el silábico. Para el oído inglés el "ritmo" español resulta marcial, porque le produce el efecto subjetivo de una ametralladora, ya que da timbre pleno a todas las vocales de las sílabas. En cambio, al oído español, el "ritmo" inglés le produce un efecto entrecortado y sujeto a tirones. El "ritmo" es probablemente el rasgo de la *base articulatoria* de una lengua cuya adquisición o dominio resulta más difícil al estudiante adulto de un idioma extranjero y, aunque la inteligibilidad depende en gran parte de su correcta emisión, a éste no se le presta la atención debida en la enseñanza de idiomas extranjeros [12].

El problema con este tipo de información (la entonación, la duración, el acento y el ritmo) es su extracción, ya que actualmente no existen métodos que extraigan la información suprasegmental del habla, simplemente se ha ligado la frecuencia fundamental F0 como un elemento de la prosodia (Itahashi [13], Thyme-Gobbel [14], Cummins [15], Muthusamy [39]). Así que realizar la identificación del lenguaje hablado sin utilizar la representación de los fonemas es un campo poco explorado; en el cual se desarrolla este trabajo de investigación.

1.2 LIMITANTES EN LA IDENTIFICACIÓN DEL LENGUAJE HABLADO

La identificación del lenguaje hablado por medios automáticos es una tarea difícil que inevitablemente debe limitarse en diversos aspectos. Por ejemplo, el tipo de locutores esperados (i.e. niños, adultos, hombres, mujeres); el tipo de conversación (palabra aislada,

frases claves, habla espontánea); el canal de transmisión de la señal de voz (micrófono, teléfono, entre otras); el nivel de ruido en la señal; el número de idiomas a identificar, etc. En particular, nuestro trabajo aborda la problemática de la identificación del lenguaje hablado cuando:

(i) El canal de transmisión es el teléfono. El canal del teléfono está limitado a anchos de banda bajos, aproximadamente de 3.2 Khz., con una frecuencia de muestreo de 8kHz, por lo que, la información en las altas frecuencias de la señal de voz se pierde, dando como resultado menos información para la discriminación;

(ii) el tipo de conversación es espontánea (introduce co-articulación y pausas);

(iii) se tiene una situación independiente del locutor (el tracto vocal en cada persona es diferente, entonces las variaciones de los hablantes en la realización de fonemas puede ser substancial).

Estas tres condiciones son los mínimos requeridos para alcanzar un sistema útil, lo cual hace el trabajo más difícil. Porque al trabajar con 8Khz de señal de voz perdemos frecuencias, las cuales podrían ser importantes en la discriminación de los lenguajes. Además, cuando el habla es espontánea existen silencios o risas que dificultan la extracción de características. Por último un sistema independiente del locutor es más difícil que uno dependiente del locutor, ya que se tendría que entrenar al sistema con un gran número de hablantes.

Los humanos no tenemos problemas en identificar un lenguaje cuando lo entendemos. Similarmente, no hay duda que un sistema de identificación de lenguaje parecido al humano debería conseguir resultados impecables, si pudiera tener un gran vocabulario almacenado con el cual es preciso reconocer y adquirir el conocimiento de las reglas sintácticas y semánticas para cada lenguaje. Con las recientes técnicas y recursos computacionales, el desarrollo de un sistema de este tipo es imposible. Las razones son las siguientes:

- La ejecución de sistemas de reconocimiento del habla está aún muy lejos de los niveles de ejecución humanos. Los actuales sistemas de reconocimiento del habla trabajan mejor con grandes restricciones en el tamaño del vocabulario. Pero para

los sistemas de identificación del lenguaje hablado, tales restricciones no pueden ser hechas.

- Recolectar y seleccionar el suficiente conocimiento de los múltiples lenguajes no es una tarea trivial. Para obtener una representación robusta de esta información, se requiere de una cantidad grande de datos de entrenamiento.

Por todo esto, tomar el camino de no utilizar la representación de los fonemas es una posibilidad más viable. Aún con todos los problemas que ello implica, esta solución evita el reconocimiento de fonemas de cada lenguaje así como la creación de modelos de lenguaje respectivos. Además facilita la introducción de nuevos lenguajes en caso de ser necesario. Por otro lado, esta es la única solución posible para lenguas marginadas sin recursos lingüísticos suficientes. Tal como es el caso de muchas de las lenguas indígenas de México

1.3 OBJETIVO DE LA TESIS

Nuestra investigación está orientada a no depender de la representación fonética de la señal de voz para la identificación del lenguaje hablado, por lo tanto el objetivo de la investigación consiste en desarrollar un nuevo método que, obteniendo información directamente de la señal de voz, nos permita obtener mejores porcentajes de identificación del lenguaje hablado que los métodos propuestos hasta ahora bajo este mismo enfoque, no utilizando información fonotáctica.

Nuestro trabajo aportará:

- Un método para extracción de características acústicas especializado para la identificación del lenguaje hablado.
- Un método de identificación del lenguaje hablado que no utilice reconocimiento fonético.

Con base en el diagrama de componentes básicos para la identificación del lenguaje hablado sin representación fonética -ver figura 1.1-, se realizó el trabajo en dos pasos principales: el primero dedicado al procesamiento acústico de la señal de voz y el segundo a la identificación del lenguaje.

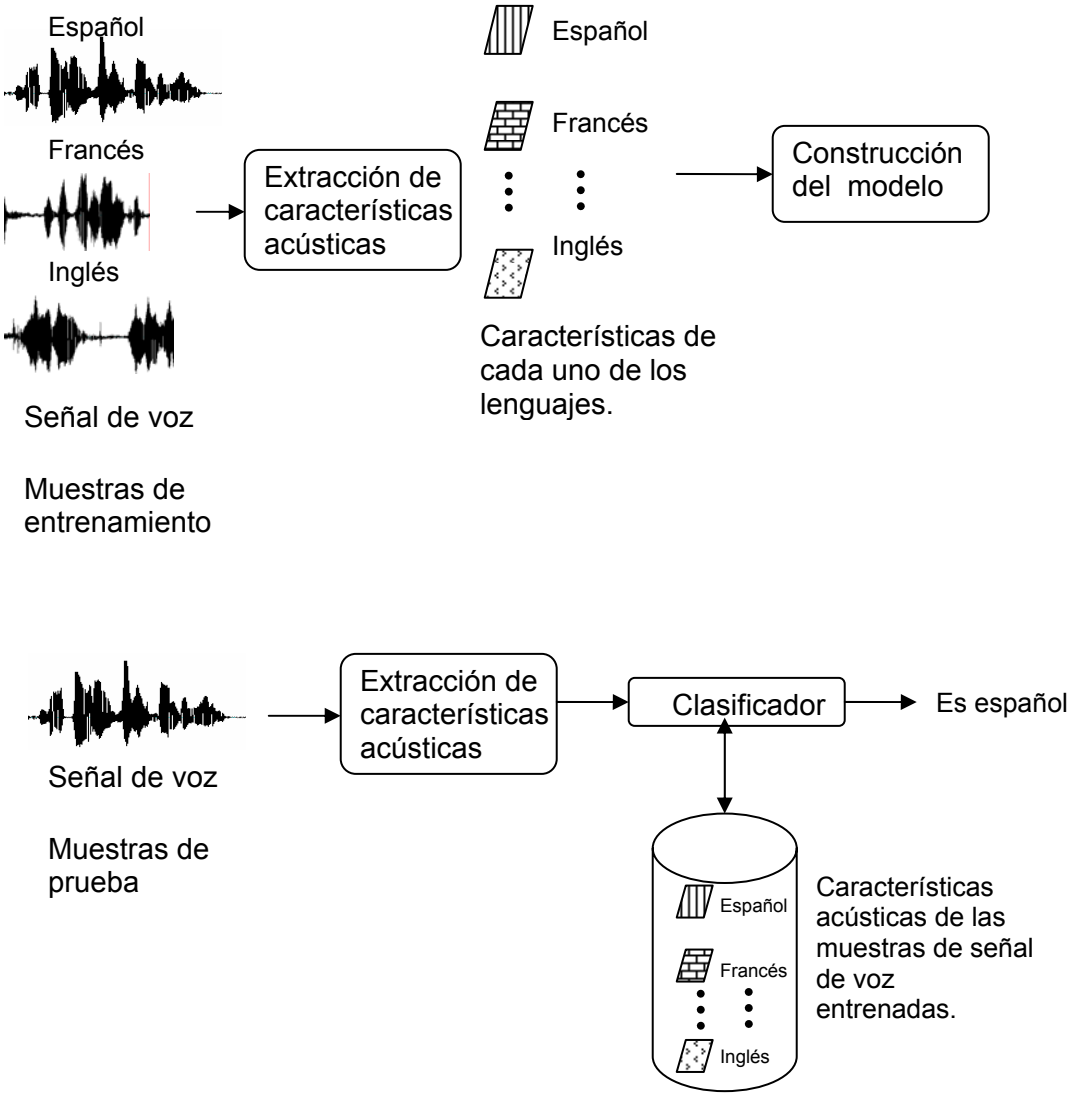

Figura 1.1 Componentes básicos para la identificación del lenguaje hablado sin representación fonética.

1.3.1 OBJETIVOS PARTICULARES

Al no depender de la representación fonética de la señal de voz, el peso del método recae en el procesamiento acústico. Se necesita un nuevo proceso acústico que extraiga las características más representativas para una mejor discriminación de los lenguajes. Por otro lado, tenemos que los lingüistas han estudiado el habla no solo en términos de fonemas, sino que además ellos definen características extralingüisticas al hablar, tales como la prosodia, el ritmo, el tono y la duración. Dichas características las definen como fenómenos fonético-fonológicos (ver capítulo 3), los cuales no pueden segmentarse como los fonemas, porque actúan simultáneamente sobre más de un segmento, es por ello que podemos hablar de fonemas segmentales y suprasegmentales. Entonces tenemos los siguientes objetivos particulares para el procesamiento acústico de la señal de voz:

1. Extracción de características suprasegmentales del habla para la tarea de identificación del habla. Para ello, proponemos capturar los cambios temporales (o deltas Δ) en el espectro de la señal de voz de diferentes coeficientes cepstrales. Es a través de los deltas que se busca capturar la información suprasegmental (la prosodia, el ritmo, la duración y el tono) presente en una elocución. Adicionalmente, se propone aumentar el número de coeficientes cepstrales de frecuencia Mel –ver capítulo 2–. Comúnmente se han utilizado 12 coeficientes con muy buenos resultados para la segmentación de fonemas [17], principalmente en la tarea de reconocimiento del habla. Proponemos aumentar el número de coeficientes centrales Mel a 16 buscando obtener más detalle de las frecuencias.

2. Extracción de características asociadas al ritmo utilizando la transformada Wavelet. La transformada Wavelet no se ha utilizado hasta ahora en la problemática de la identificación del lenguaje, sin embargo, ésta ha sido utilizada en el reconocimiento del habla, obteniendo resultados interesantes [18][19]. En nuestro caso, las wavelet son de gran interés por su capacidad para representar señales con una muy buena resolución en los dominios del tiempo y la frecuencia [20]. La transformada Wavelet permite una buena resolución en las bajas frecuencias que es donde están la

prosodia y el ritmo, elementos que nos permitirán discriminar entre lenguajes. Recordemos que la prosodia ha sido vinculada a la frecuencia fundamental F0 y dicha frecuencia es la más baja, además estudios recientes muestran que el humano no procesa frecuencias individuales independientemente como lo sugiere el análisis acústico, a excepción del ruido blanco que es producido arficialmente; en su lugar escuchamos grupos de frecuencias y aunado a que la prosodia está en la frecuencias bajas, estamos interesados en distinguir el grupo de frecuencias bajas de la señal de voz con una muy buena resolución.

3. Aplicación y evaluación de estos nuevos métodos de caracterización utilizando el corpus OGI multi-language telephone speech OGI_TS [21].
    - Para poder comparar los resultados apropiadamente, realizar pruebas con los lenguajes utilizados en el estado del arte, Cummins et al [15] y Rouas et al [22].
    - Para determinar el alcance de los métodos en lenguas marginadas, realizar pruebas con muestras de lenguas indígenas de México.
    - Para demostrar la pertinencia de la caracterización, realizar experimentos con diferentes clasificadores: estocásticos, máquinas de vectores de soporte y árboles de decisión.

1.4 ORGANIZACIÓN DE LA TESIS

La tesis se ha organizado de la siguiente manera, en el capítulo 2 veremos los fundamentos del tratamiento de la señal de voz, esto es con el fin de familiarizarse con algunos términos. Este capítulo incluye una pequeña introducción al manejo de la señal de voz, así como la transformada de Fourier y la transformada Wavelet; mostrando las diferencias principales entre estas dos transformadas. En el capítulo 3 se detallan los elementos lingüísticos usados para describir el ritmo de un lenguaje: la entonación, la duración, el acento, el ritmo, etc. Además se muestra un estudio del alcance del ser humano en la identificación del lenguaje hablado. En el capítulo 4 se presenta el estado del

arte en la identificación del lenguaje hablado, explicando con más detalle los trabajos contra los cuales comparamos nuestros resultados.

El protocolo de experimentación, tal como el corpus utilizado, los tamaños de muestras y la cantidad de lenguas a identificar, se describe en la primera sección del capítulo 5. El resto del capítulo se dedica a la explicación, instrumentación, discusión y resultados del primer método de caracterización propuesto. De igual forma el segundo método se detalla en el capítulo 6. Finalmente en el capítulo 7 damos nuestras conclusiones generales y trabajo futuro.

Además, en un anexo abordamos el tema de la identificación de las lenguas indígenas de México, su importancia y posibles aplicaciones. Mostramos los resultados obtenidos con los métodos propuestos al aplicarlos a dos lenguas: Náhuatl y Zoque de Oaxaca.

CAPÍTULO 2

FUNDAMENTOS DEL TRATAMIENTO DE LA SEÑAL DE VOZ

En este capítulo veremos en forma general los conceptos y fundamentos asociados a la forma de trabajar la señal de voz por medio de una computadora. Con la idea de tener un panorama más completo de su manejo.

El habla es una señal continua, la cual varía en el tiempo. Esta señal es el producto de las variaciones en la presión del aire cuando hablamos. Un micrófono convierte esas variaciones de presión del aire a variaciones en voltaje, lo que comúnmente llamamos señal analógica. La señal analógica, se puede transmitir a través de un circuito telefónico o puede ser almacenado en una cinta magnética. Sin embargo, para la computadora es necesario digitalizar la señal, convirtiendo la señal analógica a una serie de valores numéricos con una frecuencia regular (frecuencia de muestreo). El número de valores está limitado por el número de bits seleccionados para representar a cada muestra.

La resolución es la cantidad de bits utilizados para almacenar la voz; es decir, cada muestra se representa con un valor digital, limitando el rango de valores discretos

correspondiente al original [23]. Por ejemplo: utilizando 4 bits se pueden representar 16 valores diferentes y con 8 bits se pueden representar 256 valores.

La resolución del teléfono es de 8bits/muestra, es decir, si muestreamos a 8kHz, tenemos 8000 muestras por segundo y así 8000 * 8 = 64000 bits por segundo. En cambio en un CD la resolución es de 16 bits/muestra, por lo tanto, 44100 muestras por segundo * 16 bits = 705600 bits por segundo en modo mono aural, si es estéreo, se duplica.

En general, hemos dividido las características más importantes para el análisis acústico en frecuencia y amplitud, resonancia y percepción auditiva, las cuales se describen a continuación [17][23][24].

2.1 FRECUENCIA Y AMPLITUD

El sonido puede definirse como la decodificación que efectúa nuestro cerebro de las vibraciones percibidas a través de los órganos de la audición. Estas vibraciones se transmiten en forma de ondas sonoras. Todos los sonidos causan movimientos entre las moléculas del aire. Algunos sonidos, tales como los que produce una cuerda de guitarra, producen patrones regulares y prolongados de movimiento del aire. Los patrones de sonidos más simples son los sonidos puros, y se pueden representar gráficamente por una onda senoidal.

La amplitud de una onda sonora corresponde fisiológicamente al movimiento del tímpano de oído. La distancia desde la posición de reposo hasta la de máxima presión alcanzada por una partícula de aire se llama amplitud; la cual es una medida de fuerza de la onda; entonces el volumen de un sonido refleja la cantidad de aire que es forzada a moverse y su unidad es el decibel (dB) [23].

La frecuencia es el número de vibraciones (ciclos) del tono por segundo, por ejemplo, si tenemos 100 ciclos por segundo esto equivale a 100Hz. Los tonos altos se representan por frecuencias altas y los tonos bajos con frecuencias bajas [23].

El humano produce señales de voz desde los 100 Hz. (hombre) ó 200 Hz. (mujer) hasta los 15000 Hz. El teléfono muestrea a 8000 Hz, perceptible pero con baja calidad, comparado con un CD, que muestrea a 44100 Hz. En el caso de los instrumentos musicales el ancho de banda es mayor que para la voz, por lo que la diferencia es audible; esto requiere un mayor espacio para almacenar y transmitir.

El habla no es un tono puro, es una serie de múltiples frecuencias y se representa como una señal compuesta, es decir, el resultado de la adición de un número determinado de ondas sinusoidales simples. El teorema de Fourier (1822) demostró que toda señal compuesta periódica (es decir, que repite periódicamente su perfil) puede descomponerse en un número limitado de ondas sinusoidales simples, ver la figura 2.1; donde tenemos tres ondas periódicas simples de 100,200 y 300 Hz, y en la parte de abajo podemos ver la onda periódica compuesta (la línea de trazo grueso), que es el resultado de la suma algebraica de las ondas simples (las líneas discontinuas). La frecuencia de cada una de esas ondas es múltiplo de la frecuencia fundamental F0, que es la más baja. Las vocales se componen de dos o más ondas simples, son ricas en frecuencias secundarias y tienen estructuras internas que incluyen ondas cíclicas y no cíclicas.

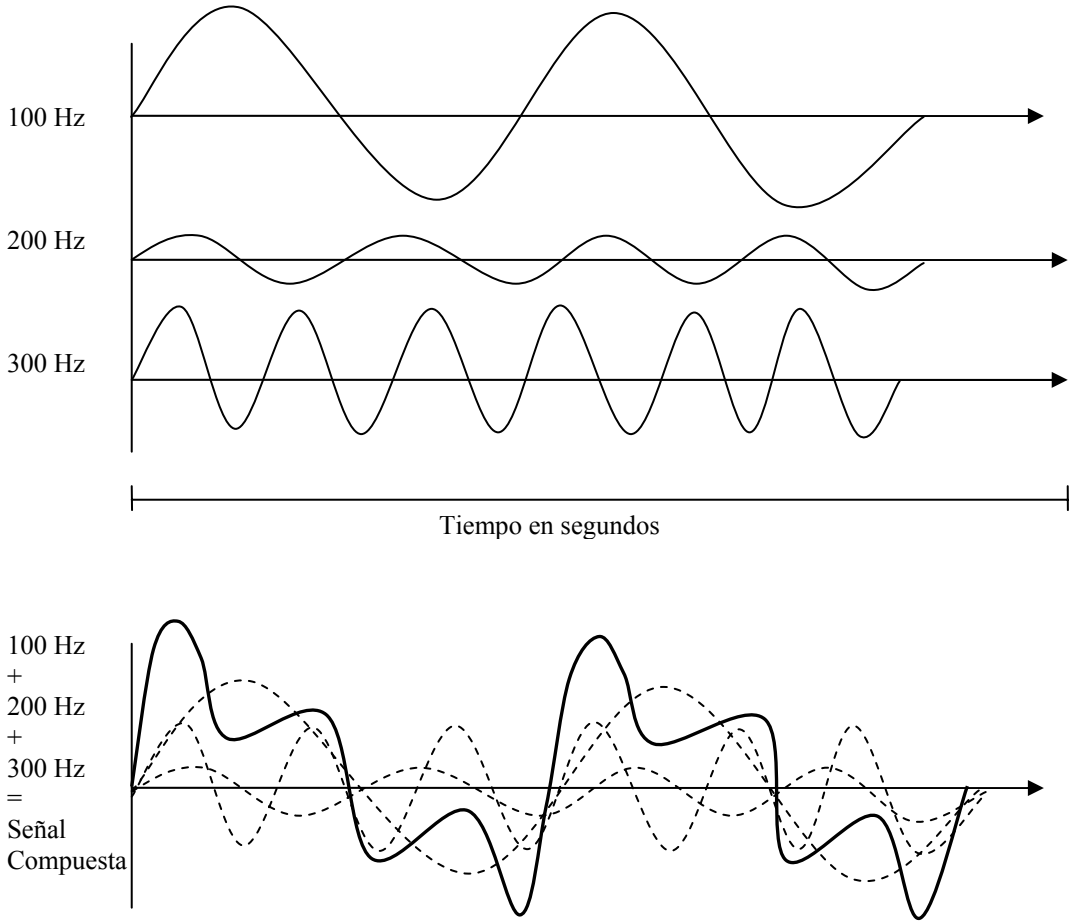

Figura 2.1 Ejemplo de señal compuesta, resultado de la suma algebraica de ondas simples.

Hay básicamente dos tipos de ondas compuestas: las cíclicas (periódicas) y las no cíclicas (aperiódicas). Las ondas cíclicas repiten periódicamente su perfil, debido a cambios regulares en la presión del aire, sus componentes son múltiplos de la frecuencia fundamental, además generan un espectro en línea. Las ondas no cíclicas no repiten periódicamente su perfil, porque hay cambios irregulares en la presión del aire, tienen componentes de todas las frecuencias, y generan un espectro continuo. En la grafica 2.2

podemos ver dos ventanas de análisis de dos sonidos contiguos. La ventana superior es un oscilograma (muestra tiempo y amplitud) y la ventana inferior es un espectrograma (muestra tiempo, frecuencia e intensidad). En estas graficas podemos ver las diferencias entre un sonido periódico (cuando hablamos) o no periódico (no hay señal de voz). El espectro del primer sonido esta constituido por una serie de barras verticales en un tono más fuerte que el del segundo sonido, y en este segundo sonido se puede observar que su intensidad es muy baja y carece de ondas compuestas, puesto que es aperiódico.

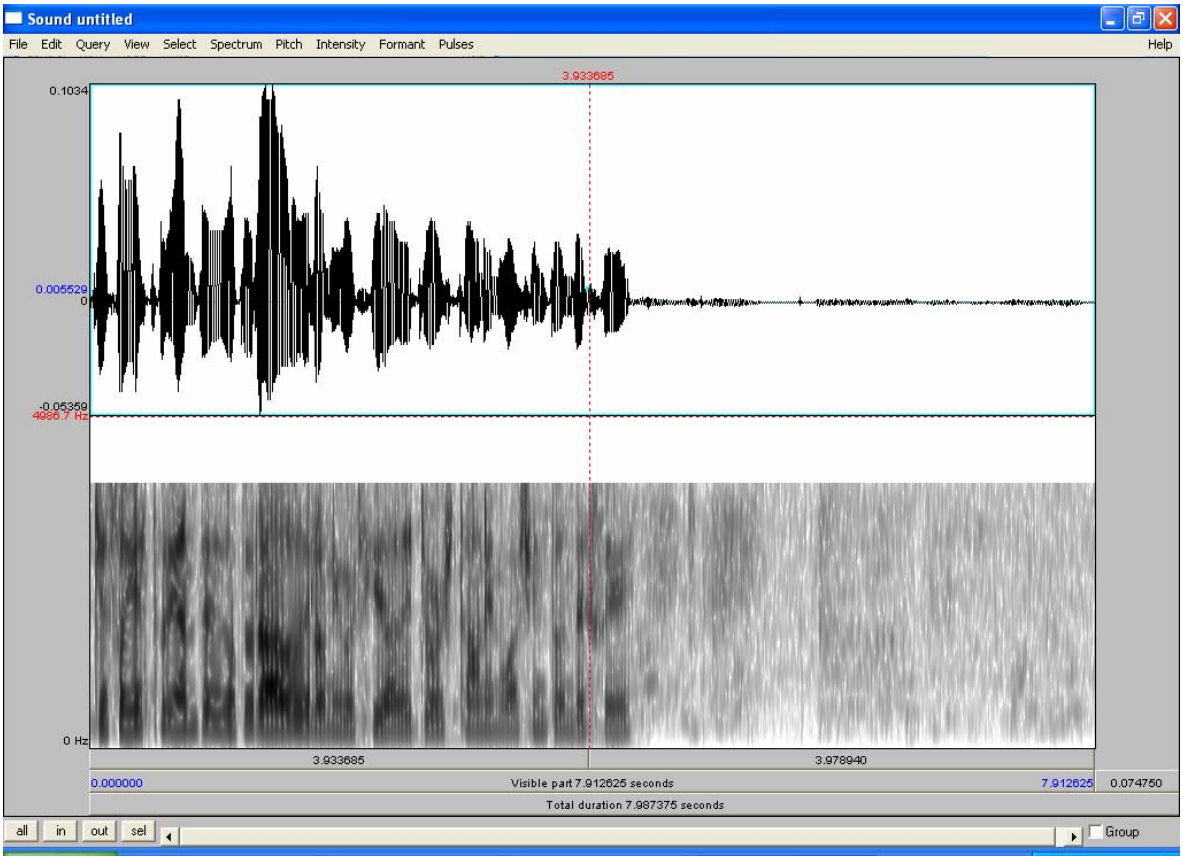

Figura 2.2 Oscilograma y espectrograma de una muestra de señal de voz

2.2 RESONANCIA

La resonancia se define como la habilidad que tiene una fuente vibrante de sonido de causar que otro objeto vibre. Las cámaras de resonancia en instrumentos de música responden a frecuencias específicas o anchos de banda específicos. Al ser estas cajas o cámaras de resonancia más grandes que la fuente del sonido amplifican las frecuencias a las que responden.

La garganta, boca y nariz son cámaras de resonancia que amplifican las bandas o frecuencias formantes contenidas en el sonido generado por las cuerdas vocales, así se originan los distintos sonidos de la lengua. Dicho de otra manera, si la articulación no filtrara constantemente el sonido, cada hablante emitiría siempre un único sonido. Estos formantes amplificados dependen del tamaño y forma de la boca, y si el aire pasa o no por la nariz. Los patrones de los formantes son más fuertes (distinguibles) para vocales que para las consonantes no sonoras. Una consonante puede ser sonora o no sonora. Una consonante es sonora cuando hacer vibrar las cuerdas vocales y es sorda cuando se pronuncia sin hacer vibrar las cuerdas vocales. Por ejemplo, la letra “b” es consonante sonora y la “p” es sorda. En general la función básica de las cámaras de resonancia es reforzar ciertas frecuencias de la onda compleja que le llega.

La mayoría de los sonidos, incluyendo el habla, tienen una frecuencia dominante llamada frecuencia fundamental, comúnmente denotada por F0, y es la más baja [23]; la percibimos como el tono (pitch) combinado con frecuencias secundarias. En el habla, la frecuencia fundamental es la velocidad a la que vibran las cuerdas vocales al producir un fonema sonoro. Sumadas a la frecuencia fundamental hay otras frecuencias que contribuyen al timbre del sonido. Son las que nos permiten distinguir una trompeta de un violín o las voces de diferentes personas. Algunas bandas de las frecuencias secundarias juegan un rol importante en la distinción de un fonema de otro. Se les llama formantes y son producidas por la resonancia.

En la figura 2.3 se muestra en la ventana superior el oscilograma de la señal de voz hablada que representa la secuencia de las siguientes palabras: “Siempre hace calor … y

como a las dos de la tarde mas o menos comienza a llover" en el idioma español. En la ventana inferior se muestra el espectrograma de esa secuencia de palabras, así como los formantes (color rojo), los cuales son cinco, empezando por el formante 1.

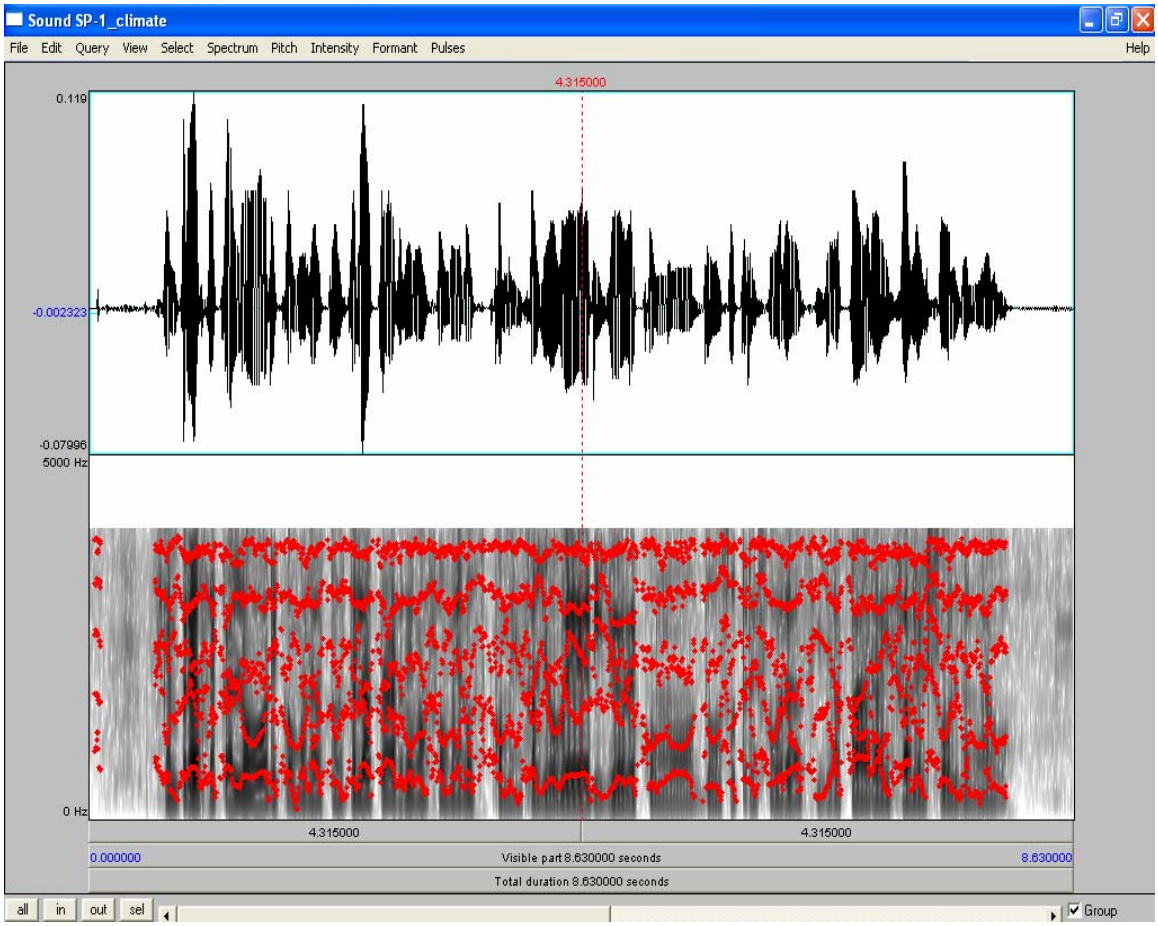

Figura 2.3 Oscilograma y espectrograma de una frase en español, en color rojo se muestran los formantes.

La identidad de las consonantes también se revela por el cambio en las formantes que resultan cuando los articuladores se mueven de un fonema anterior a la consonante y de ella al siguiente fonema, llamadas transiciones de formantes. Estas se analizan

utilizando técnicas como la transformada rápida de Fourier (FFT) generando espectrogramas.

En la figura 2.4 se muestra en la ventana inferior el espectrograma de la figura 2.3 pero mostrando la intensidad (color amarillo) y la frecuencia fundamental “pitch” (color azul). Hay que notar que la intensidad se mide en Pascales (Pa), pero para efectos ilustrativos se muestra en el mismo espectrograma, aun cuando tienen diferentes escalas.

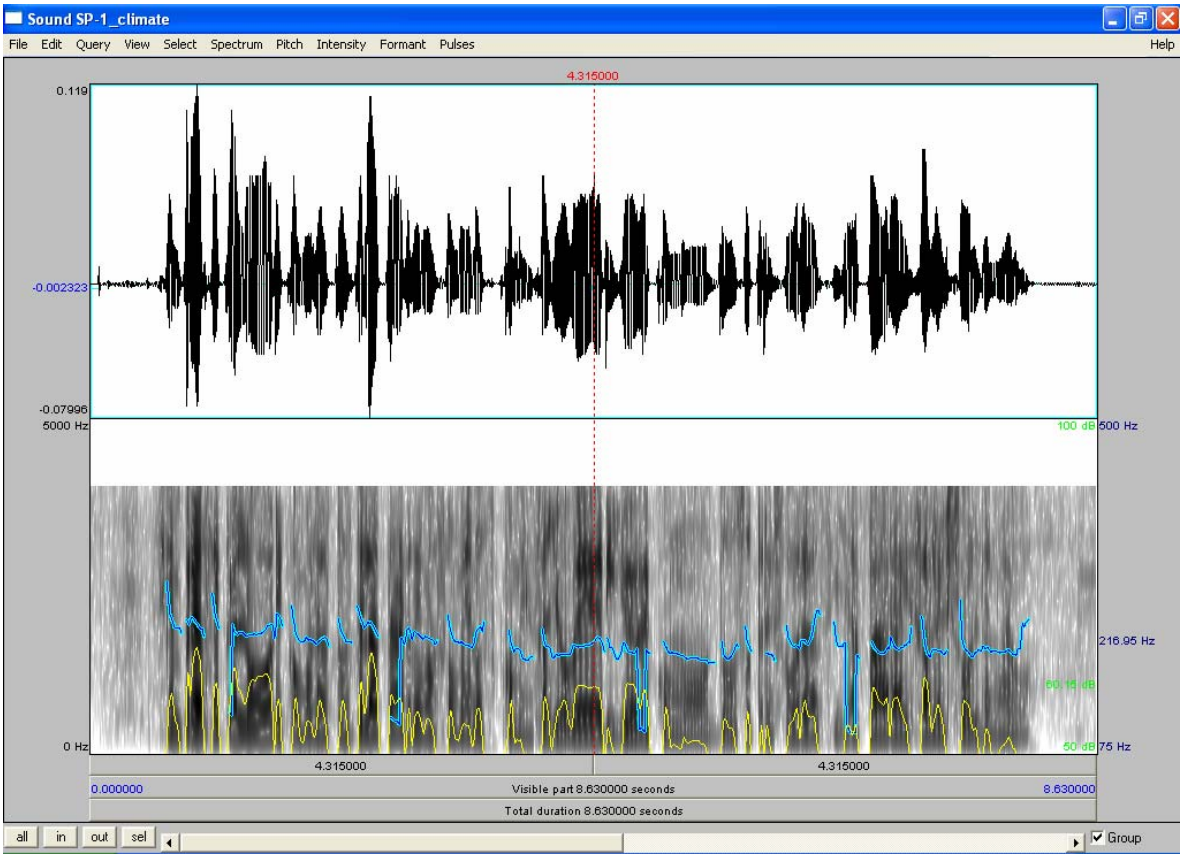

Figura 2.4 Oscilograma y espectrograma de una frase en español, donde se muestra la frecuencia fundamental (pitch, color azul) y la intensidad en color amarillo.

2.3 PERCEPCIÓN AUDITIVA

Los fonemas aparentemente tienen parámetros acústicos claramente definidos, pero más bien, los fonemas tienden a ser abstracciones implícitamente definidas por la pronunciación de las palabras en un lenguaje.

La forma acústica de un fonema depende fuertemente del contexto acústico en el que sucede. A este efecto se le llama coarticulación, por ejemplo, es diferente la pronunciación de “to” en las palabras: *to*do y en es*to*rnudo. Este concepto se utiliza para distinguir entre la característica conceptual de un sonido del habla (fonema) y una instancia o pronunciación específica de ese fonema (fono).

La variabilidad del habla producida por la coarticulación y otros factores, tales como el tipo de locutores, el tipo de conversación, si es espontánea, el canal de transmisión de la señal de voz y el nivel de ruido en la señal; hacen el análisis de la voz extremadamente difícil. La facilidad del humano en superar estas dificultades sugiere que un sistema basado en la percepción auditiva podría ser un buen enfoque. Desafortunadamente nuestro conocimiento de la percepción humana es incompleto. Lo que sabemos es que el sistema auditivo está adaptado a la percepción de la voz.

El oído humano detecta frecuencias de 20Hz a 20,000 Hz, pero es más sensible entre 100 y 6000 Hz. También es más sensible a cambios pequeños en la frecuencia en el ancho de banda crítico para el habla. Además el patrón de sensibilidad a cambios en el tono (pitch) no corresponde a la escala lineal de frecuencia de ciclos por segundo de la acústica.

Para representar mejor el patrón de percepción del oído humano se desarrolló una escala llamada Mel, la cual es una escala logarítmica. Y a partir de dicha escala se crearon los coeficientes cepstrales de Frecuencia Mel. Lo cual se detalla en la sección 2.4.5.

2.4 EXTRACCIÓN DE CARACTERÍSTICAS

Existen diferentes formas de extraer las características de la señal del habla digitalizada, por lo tanto, de acuerdo a qué característica deseamos obtener es que se utiliza una u otra forma de análisis de la señal del habla. En esta sección veremos en forma breve cada uno de los métodos que actualmente se utilizan en el procesamiento de la señal de voz para la extracción de características.

2.4.1 ANÁLISIS LOCAL

Es una técnica de análisis que consiste en tomar intervalos pequeños de la señal digital del habla a analizar e ir procesando dichos intervalos (segmentos), de los que obtendremos distintas propiedades como la energía, la autocorrelación, la densidad espectral, etc. Los segmentos de señal se tomarán mediante la técnica de ventanas, la cual consiste en ir deslizando una ventana de duración L muestras (en t seg.) a lo largo de toda la señal, pudiendo estar dichas ventanas solapadas entres sí con sus adyacentes.

Existen varios tipos de ventana para utilizar en el proceso de análisis local de una señal. Se toma un tipo u otro en función de si estamos en el dominio temporal o espectral.

Es recomendable utilizar la ventana de Hamming cuando se está realizando análisis local en el dominio de la frecuencia, ya que el espectro de la ventana de Hamming está más confinado. El producto de la ventana de análisis por la señal de voz se convierte en una convolución de los espectros de ambas señales en el dominio de la frecuencia [23]. Es recomendable utilizar la ventana rectangular cuando se este realizando el análisis local en el dominio del tiempo, puesto que temporalmente trata por igual a todas las muestras.

El análisis local en el dominio del tiempo básico, para obtener características de la señal de voz, es aquel que localiza la energía, la cual consiste en tomar la suma al cuadrado de las muestras que componen cada una de las ventanas, como resultado obtendremos un vector de muestras en las que los valores máximos se corresponden con los elementos sonoros de una determinada pronunciación.

La función de autocorrelación de una señal nos da una indicación de la dependencia entre muestras sucesivas. La señal de voz presenta correlaciones importantes a corto plazo (muestras próximas) debidas al tracto vocal y correlaciones a largo plazo (muestras lejanas) debidas a la vibración de las cuerdas vocales. Se utiliza la autocorrelación de segmentos de voz de corta duración, para distinguir entre tramos sordos y tramos sonoros estimando la frecuencia fundamental en el caso de los sonoros.

La tasa de cruces por cero es el número de veces que la señal pasa por cero dentro de cada ventana y nos da una información vinculada al contenido espectral de la ventana que estamos analizando, sin necesidad de recurrir a la transformada de Fourier. Hay sonidos con una tasa de cruces por cero baja como las vocales y otros con muy poca energía pero con una tasa de cruces por cero bastante alta. Aunque es sencillo este tipo de extracción de características no obtenemos mucha información de la señal de voz, como cuando usamos la transformada de Fourier.

2.4.2 TRANSFORMADA DE FOURIER (TF)

La transformada de Fourier tiene una multitud de aplicaciones en muchas áreas de la ciencia e ingeniería: la física, la teoría de los números, la combinatoria, el procesamiento de señales, la teoría de la probabilidad, la estadística, la óptica, la propagación de ondas y otras áreas. En procesamiento de señales, la transformada de Fourier suele considerarse como la descomposición de una señal en componentes de frecuencias diferentes.

El análisis de espectros que se define como la transformación de una señal de la representación en el dominio del tiempo hacia la representación en el dominio de la frecuencia, tiene sus raíces a principio del siglo XIX, cuando varios matemáticos lo investigaron desde una base teórica. Jean Baptiste Fourier, matemático francés, estableció que una señal o función podía ser representada como la suma, posiblemente infinita, de series de senos y cosenos. En dicha teoría están basadas casi todas nuestras técnicas modernas de análisis de espectro. Fourier estaba trabajando para Napoleón, durante la invasión de Egipto en un problema de sobrecalentamiento de cañones, cuando dedujo la famosa Serie de Fourier, para la solución de la conducción de calor. La operación de la Serie de Fourier está basada en una señal de tiempo que es periódica. Esto es, una señal de tiempo cuya forma se repite en una cantidad infinita de veces. Fourier demostró que una señal de este tipo es equivalente a una colección de funciones senos y cosenos cuyas frecuencias son múltiplos del recíproco del periodo de la señal de tiempo. El resultado un poco inesperado es que cualquier forma de onda, siempre y cuando no sea infinita en longitud, se puede representar como la suma de una serie de componentes armónicos. Las amplitudes de los varios armónicos se llaman los coeficientes Fourier, y sus valores se pueden calcular fácilmente si se conoce la ecuación para la forma de onda. Fourier más tarde generalizó la Serie de Fourier en la Transformada de Fourier. De acuerdo a esto la transformada de Fourier utiliza dos funciones bases, las cuales son seno y coseno, para poder expandir o representar una señal o función en términos de ellas. En general, dichas funciones tiene la característica de no poseer pendientes abruptas o discontinuidades, no son localizables en el tiempo, y tienen una representación individual de una frecuencia.

Un hecho importante que se puede ver de la Serie de Fourier es que la forma de onda original se puede reconstruir a partir de los coeficientes de frecuencia. En otras palabras, es posible pasar al del dominio de frecuencia y regresar hacia el dominio de tiempo sin que se pierda la información. La Serie de Fourier está perfectamente adaptada para realizar el análisis de frecuencia en formas de ondas periódicas, eso es en señales deterministas.

La transformada de Fourier se utiliza para pasar al "dominio de la frecuencia" para obtener información que no es evidente en el dominio del tiempo.

2.4.3 LA TRANSFORMADA DE FOURIER DISCRETA (DFT)

Debido a que las computadoras trabajan sólo con datos discretos y el procesamiento digital de señales en forma automática se realiza por medio de la Serie de Fourier y la Transformada de Fourier, las cuales son continuas, se desarrolló la transformada de Fourier discreta, designada por la abreviatura DFT (por sus siglas en inglés Discrete Fourier Transform), y a la que en ocasiones se denomina transformada de Fourier finita. La DFT es una transformada de Fourier ampliamente empleada en tratamiento de señales y en campos afines para analizar las frecuencias presentes en una señal muestreada, resolver ecuaciones diferenciales parciales y realizar otras operaciones, como convoluciones.

La DFT opera con una señal de muestras en el dominio del tiempo. A partir de ésta se genera un espectro de muestras en el dominio de la frecuencia. El espectro que resulta es una aproximación de la Serie de Fourier, una aproximación en el sentido que se perdió la información entre las muestras de la forma de onda. La clave hacia la DFT es la existencia de una forma de onda de la que se tomaron muestras, esto es la posibilidad de representar la forma de onda en una serie de números. Para generar esta serie de números desde una señal análoga, se requiere un procedimiento de muestreo, y de conversión de análogo a digital. La señal de la que se tomaron muestras es una representación matemática del nivel de la señal instanciada a intervalos definidos con precisión. No contiene información acerca de la señal entre los tiempos en que se tomaron muestras. Si la proporción de muestreo es lo suficientemente alta como para asegurar una representación razonable de la forma de la señal, la DFT produce un espectro que es muy similar a un espectro teóricamente verdadero. Este espectro también es discreto, y no hay información entre las muestras o "líneas" de espectro.

La DFT $X(k)$ de una señal $x(n)$ se define de la siguiente manera [23]:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j 2\pi kn / N} \qquad (2.1)$$

Donde k=0,…N-1 y $X(k)$ es una señal periódica y compleja con periodo N.

La transformada de Fourier discreta puede calcularse de modo muy eficiente mediante el algoritmo FFT (por sus siglas en inglés Fast Fourier Transform).

La transformada de Fourier de una ventana de señal de voz sólo nos da información de un sonido aislado en la cadena hablada y hay que tener en cuenta que normalmente el tamaño de la ventana es menor que la duración de un fonema. Por este motivo hemos de realizar un análisis tiempo-frecuencia para extraer información de dicha señal.

.

2.4.4 LA TRANSFORMADA DE FOURIER DE TIEMPO CORTO (STFT)

El uso de los análisis de tiempo-frecuencia para el procesamiento y análisis de señales no estacionarias, se remonta a los años 40's. Un ejemplo de señal no estacionaria es la señal de la voz humana, pero para poder analizar la señal de voz necesitamos saber cuando y donde se encuentran las diferentes componentes de frecuencia cuyo contenido espectral varía con el tiempo. Por lo que la transformada de Fourier posee una pobre resolución en el tiempo. A partir de esta necesidad se motivó el desarrollo del espectrograma de sonido usando los conceptos de la transformada de Fourier de tiempo corto (STFT por sus en inglés Short Time Fourier Transform).

La STFT es una adaptación de la transformada de Fourier usando un procedimiento llamado ventaneo. El cual consiste en dividir una señal en pequeños segmentos, a través del tiempo, de tal manera que podamos asumir que para cada segmento la señal es estacionaria, y así calcular la transformada de Fourier clásica para cada segmento de la señal. La forma de dividir la señal se realiza mediante una función tiempo-ventana, cuyo ancho corresponde a la longitud de cada segmentación de la señal. En general este método consiste en fijar un tiempo de interés y volver corriente otro

tiempo. Se aplica una función ventana, que corresponde al producto, que estará centrada alrededor del tiempo de interés.

El ancho de la ventana constituye un parámetro de gran importancia ya que a través de éste podemos establecer el grado de resolución tanto de tiempo como de frecuencia. Si la ventana es muy pequeña tenemos una buena resolución en tiempo pero una mala resolución en frecuencia, ya que veremos solo una pequeña fracción del espectro total de la señal de voz. Por el contrario, si la ventana es muy grande tendremos una buena resolución en frecuencia pero una mala resolución en tiempo. Por lo tanto un defecto de la STFT es que no puede generar una buena resolución tanto en tiempo como en frecuencia de manera simultanea ya que el ancho de la ventana es fijo.

Un espectrograma es una representación de la evolución de las características espectrales de una señal (en nuestro caso voz) a lo largo del tiempo. Con base en los espectrogramas se realizan modelado de señales tales como los de articulación y percepción, para los cuales se realiza análisis de predicción lineal (LPC) y análisis de coeficientes cepstrales en escala de Mel (MFCC) respectivamente.

2.4.5 LOS COEFICIENTES CEPSTRALES DE FRECUENCIA MEL

Para representar mejor el patrón de percepción del oído humano se desarrolló una escala llamada Mel, que es una escala logarítmica. Entonces, los coeficientes cepstrales de frecuencia Mel (MFCC por sus siglas en inglés Mel Frequency Cepstral Coefficients), son coeficientes para la representación del habla basados en la percepción auditiva humana. Los MFCC se derivan de la transformada de Fourier, la diferencia es que en MFCC las bandas de frecuencia están situadas logarítmicamente, según la escala de Mel. Los MFCC modelan la respuesta auditiva humana más apropiadamente que las bandas espaciadas linealmente de FT. Esto permite un procesado de datos más eficiente, por ejemplo, en compresión de audio.

La escala Mel, fue propuesta por Stevens, Volkman y Newman en 1937, el punto de referencia entre esta escala y la frecuencia normal se define equiparando un tono de 1000 Hz, 40 dBs por encima del umbral de audición del oyente, con un tono de 1000 mels. Por encima de 500 Hz, los intervalos de frecuencia espaciados exponencialmente son percibidos como si estuvieran espaciados linealmente. En consecuencia, cuatro octavas en la escala de hercios por encima de 500 Hz se comprimen a alrededor de dos octavas en la escala mel.

2.4.6 LA WAVELET

La wavelet es el resultado de un gran número de investigaciones y constituye una técnica de análisis reciente. Inicialmente un geofísico francés llamado Jean Morlet, en los 80´s, investigaba un método para modelar la propagación del sonido a través de la corteza terrestre. Morlet, desarrolló su propia forma de analizar las señales sísmicas para crear componentes que estuvieran localizados en el espacio, a los que denominó "wavelet de forma constante". Independientemente de que los componentes se dilaten, compriman o desplacen en el tiempo, mantienen la misma forma. Se pueden construir otras familias de wavelet adoptando una forma diferente, denominada wavelet madre, dilatándola, comprimiéndola o desplazándola en el tiempo [25]. Por lo tanto, como alternativa a la transformada de Fourier, Morlet utilizó un sistema basado en una función prototipo, que cumpliendo ciertos requerimientos matemáticos y mediante dos procesos denominados dilatación y traslación, forman un conjunto de bases que permitían representar las señales de propagación con la misma robustez y versatilidad que la transformada de Fourier, pero sin sus limitaciones, como por ejemplo, el ventaneo fijo.

Morlet y Grossmann trabajaron para demostrar que las ondas se podían reconstruir a partir de sus descomposiciones en wavelet. De hecho, las transformaciones de wavelets resultaron funcionar mucho mejor que las transformadas de Fourier, porque eran mucho menos susceptibles a pequeños errores de cómputo. Un error o un truncamiento

indeseados de los coeficientes de Fourier pueden transformar una señal suave en una saltarina o viceversa; las wavelet evitan tales consecuencias desastrosas.

La simplicidad y elegancia de esta nueva herramienta matemática fue reconocida por un matemático francés llamado Yves Meyer, quien descubrió que las wavelets formaban bases ortonormales de espacios ocupados por funciones cuyo cuadrado es integrable, lo que traducido al lenguaje de procesamiento de señales, corresponde a funciones o señales cuyo contenido energético es finito [26]. En este momento ocurrió una pequeña explosión de actividad en esta área. Ingenieros e investigadores comenzaron a utilizar la transformada wavelet para aplicaciones en diferentes campos tales como la astronomía, acústica, ingeniería nuclear, detección de terremotos, compresión de imágenes, reconocimiento del habla, visión humana neurofisiología, entre otras.

El término wavelet se define como una pequeña onda o función localizable en el tiempo, que visto desde una perspectiva del análisis o procesamiento de señales puede ser considerada como una herramienta matemática para la representación y segmentación de señales, análisis tiempo-frecuencia, y con algoritmos computacionales de fácil implementación.

En términos históricos, el desarrollo de las wavelet entronca con varias líneas de pensamiento, a partir del trabajo de Alfred Haar a principios del siglo XX. Contribuyeron de modo notable al avance de la teoría Goupillaud, Grosman y Morlet con su formulación de lo que hoy conocemos como transformada wavelet continua, Jan Olov-Strömberg con su trabajo sobre wavelet discretas en 1983, Ingrid Daubechies, con su propuesta de wavelet ortogonales con soporte compacto en 1988, Stephane Mallat y Yves Meyer, con su marco multiresolución en 1989, Delrat con su interpretación de la transformada wavelet en tiempo-frecuencia en 1991, Newland, con su transformada wavelet armónica, y muchos otros desde entonces.

La idea del análisis *multiresolución* [27], es decir la observación de señales a distintas escalas de resolución, ya era familiar para los expertos en procesamiento de imágenes. La transformada wavelet y la multiresolución están relacionadas por medio de un banco de filtros de dos bandas, el cual se compone de un filtro pasa-bajas y un filtro

pasa-altas. Al hecho de poner en cascada bancos de filtros se le conoce como niveles de descomposición. La figura 2.5 muestra la descomposición y reconstrucción de la señal, donde h_1 es un filtro pasa-bajo y g_1 es un filtro pasa-alta; a_{j+1} representa la aproximación, para la voz son las bajas frecuencias donde se presume está la prosodia y el ritmo. D_{j+1} representa el detalle, siendo las altas frecuencias.

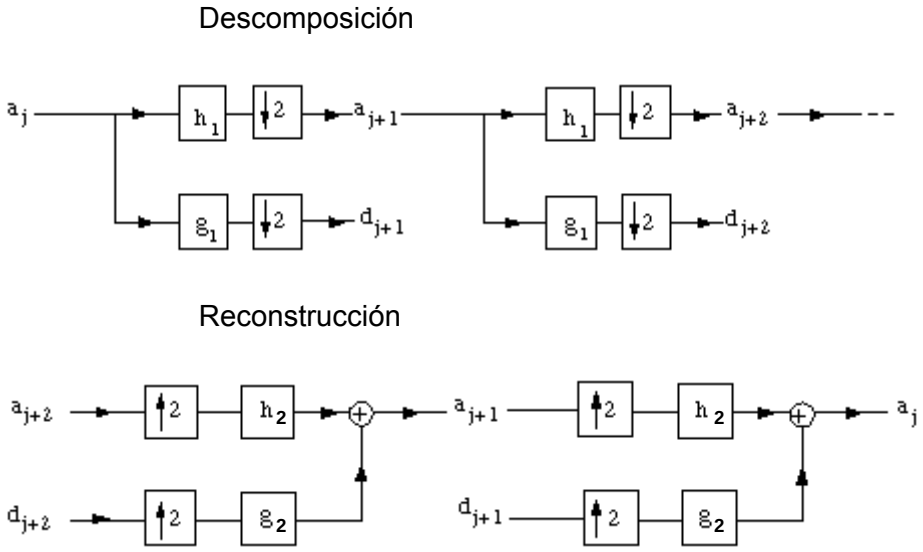

Figura 2.5 Proceso de descomposición y reconstrucción de la señal por medio de Wavelet.

La teoría de wavelet está relacionada con muy variados campos. Todas las transformaciones wavelet pueden ser consideradas formas de representación en tiempo-frecuencia y, por tanto, están relacionadas con el análisis armónico. La transformada wavelet es un caso particular de filtro de respuesta finita al impulso.

Gracias al trabajo de Mallat [27], las wavelet se convirtieron en algo mucho más sencillo. Ya se podía hacer un análisis de wavelet sin necesidad de conocer la fórmula de una wavelet madre. Finalmente en 1987, Ingrid Daubechies [20] descubrió una clase completamente nueva de wavelet, que no sólo eran ortogonales, sino que también se podían implementar mediante sencillas ideas de filtrado digital. Las nuevas wavelet eran

casi tan sencillas de programar y utilizar como la wavelet de Haar, pero eran suaves, sin los saltos de la wavelet de Haar.

2.4.7 LA TRANSFORMADA WAVELET

La transformada wavelet constituye una técnica relativamente nueva que ha sido propuesta por los investigadores como una poderosa herramienta en el análisis sobre el comportamiento local de una señal. Al igual que la STFT, esta transformada utiliza una función ventana que encuadra una señal dentro de un intervalo y focaliza el análisis sólo en ese segmento de la señal.

Las características propias de la transformada wavelet nos otorgan la posibilidad de representar señales en diferentes niveles de resolución, representar en forma eficiente señales con variaciones de picos abruptos, analizar señales no estacionarias permitiéndonos saber el contenido en frecuencia de una señal y cuando estas componentes de frecuencia se encuentran presentes en la señal [25].

La transformada wavelet continua (CWT por sus siglas en inglés Continuous Wavelet Transform) intenta expresar una señal continua en el tiempo, mediante una expansión de términos o coeficientes que se obtienen del producto interno entre la señal y diferentes versiones escaladas y trasladadas de una función prototipo más conocida como wavelet madre. Mediante la variable de escala se puede comprimir o dilatar la wavelet madre, lo que nos dará el grado de resolución con el cual estaremos analizando la señal. Por definición la transformada wavelet continua es más una representación tiempo-escala que una representación tiempo-frecuencia. Para valores pequeños de la variable de escala la CWT obtiene información de la señal que está esencialmente localizada en el dominio del tiempo, mientras que para valores grandes de la variable de escala la CWT obtiene información de la señal localizada en el dominio de la frecuencia. En otras palabras, para escalas pequeñas la CWT nos entrega una buena resolución en el dominio tiempo, mientras que para escalas grandes la CWT nos entrega una buena resolución en el

dominio de la frecuencia. Cuando la variable de escala cambia, tanto la duración como el ancho de banda de la wavelet madre cambian, pero su forma se mantiene igual. En lo anteriormente dicho se encuentra la diferencia principal entre la CWT y la STFT, ya que la primera usa ventanas de corta duración para altas frecuencias y ventanas de larga duración para bajas frecuencias, mientras que la STFT usa una sola ventana con la misma duración tanto para altas frecuencias como para bajas frecuencias. Siendo esto importante para este trabajo. Ver figura 2.6.

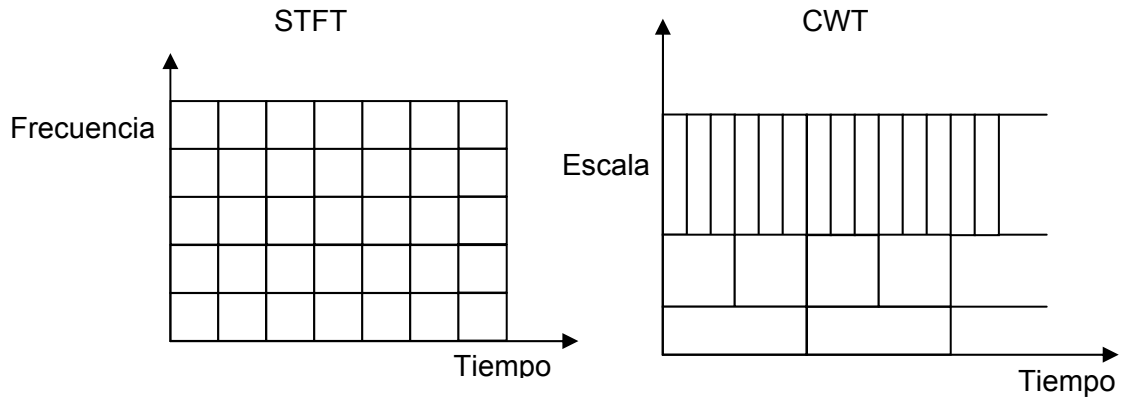

Figura 2.6 Diferencia entre STFT tiempo-frecuencia contra CWT escala-tiempo.

La variable de traslación controla la ubicación de la función en el espacio de tiempo permitiéndonos deslizar la wavelet madre sobre el intervalo de tiempo en el que se haya definido la señal. Un punto importante es que la función wavelet madre se traslada cubriendo toda la señal para cada valor de la variable de escala, es decir, si la escala escogida es pequeña habrá más traslaciones de la wavelet que si la escala escogida es grande [25].

La continuidad de la CWT reside en que tanto la variable de escala como la variable de traslación varían en forma continua. Por lo que para su uso computacional es

necesario discretizar la transformada. Por lo tanto, los valores de las variables tanto de escala como de traslación deben ser discretos. Al hacer esto se creo la Transformada discreta Wavelet (DWT por sus siglas en inglés Discrete Wavelet Transform). La implementación utilizada en nuestro trabajo será la de Daubechies.

En cuanto a sus aplicaciones, la transformada wavelet discreta DWT se utiliza para la codificación de señales, mientras la continua se utiliza en el análisis de señales. Como consecuencia, la versión discreta de este tipo de transformada se utiliza fundamentalmente en ingeniería, mientras que la continua se utiliza sobre todo en la física. Este tipo de transformadas están siendo cada vez más empleadas en un amplio campo de especialidades, a menudo sustituyendo a la transformada de Fourier. Se puede observar este desplazamiento en el paradigma en múltiples ramas de la física, como la dinámica molecular, la astrofísica, la geofísica de los sismos, la óptica, el estudio de las turbulencias y la mecánica cuántica, así como en otros campos muy variados como el procesamiento de imágenes, los análisis de sangre, el análisis de electrocardiogramas, el estudio del ADN, el análisis de proteínas, la meteorología, el procesamiento de señal en general, el reconocimiento del habla, los gráficos por ordenador, el análisis multifractal y en el campo de la biometría.

A diferencia de la DFT, la Transformada discreta wavelet (DWT) es una herramienta muy utilizada en el análisis de señales no estacionarias como la señal del habla e imágenes. La DWT descompone la señal original x(n) en dos señales, x_0(m) y x_1(m). La DWT de x(n) esta dado por [27]:

$$x_0(m) = \sum_{k=0}^{N-1} h_0(k)\, x(2m - k), \qquad (2.2)$$

$$x_1(m) = \sum_{k=0}^{N-1} h_1(k)\, x(2m - k), \qquad (2.3)$$

Donde N es el tamaño de las secuencias $h_0(n)$ y $h_1(n)$. Vale la pena destacar que $h_0(n)$ y $h_1(n)$ no son únicos y que su selección depende en específico de la aplicación.

Las secuencias $h_0(n)$ y $h_1(n)$ son llamadas filtros pasa-bajas y filtros pasa-altas, respectivamente. En la figura 2.5, se muestran estos filtros con h_1 y g_1.

Los más conocidos y utilizados coeficientes del filtro para h_0(n) y h_1(n) fueron descubiertos por Daubechies y son llamados filtros wavelet "db*N*".

Usando los coeficientes $h_0(n)$ y $h_1(n)$, se obtiene la función de escala y wavelet correspondiente de la siguiente manera [20]:

$$\phi(t) = \sqrt{2}\sum_{n=0}^{N-1} h_0(n)\phi(2t - n), \qquad (2.4)$$

$$\psi(t) = \sqrt{2}\sum_{n=0}^{N-1} h_1(n)\phi(2t - n). \qquad (2.5)$$

En este trabajo se utilizo el número de coeficientes del filtro *N* igual a 4 "db4". El valor de los coeficientes se muestra en la tabla 2.1.

n	$h_0(n)$	$h_1(n)$
0	$\frac{1-\sqrt{3}}{4\sqrt{2}}$	$-\frac{1+\sqrt{3}}{4\sqrt{2}}$
1	$\frac{3-\sqrt{3}}{4\sqrt{2}}$	$\frac{3+\sqrt{3}}{4\sqrt{2}}$
2	$\frac{3+\sqrt{3}}{4\sqrt{2}}$	$-\frac{3-\sqrt{3}}{4\sqrt{2}}$
3	$\frac{1+\sqrt{3}}{4\sqrt{2}}$	$\frac{1-\sqrt{3}}{4\sqrt{2}}$

Tabla 2.1 Coeficientes de h_0(n) y h_1(n) de Daubechies db4 (tomada de [20]).

2.4.8 DIFERENCIAS ENTRE LAS TRANSFORMADA DE FOURIER Y WAVELET

La diferencia más importante entre las transformadas Fourier y Wavelet es que las funciones wavelet son localizadas en el espacio. Las funciones de coseno y seno de Fourier no son localizadas en el espacio. Por la forma de caracterizar la señal con wavelet, se pueden realizar muchas funciones y operaciones usando pocas wavelet. Lo que es muy útil en aplicaciones tales como la compresión de datos, detección de características en las imágenes, y en la eliminación de ruido de una serie de tiempo. Es claro que las wavelet no han aparecido como una herramienta que desplaza a la transformada de Fourier, sino más bien como una herramienta que puede complementarse con la TF, o ser una correcta elección dependiendo del tipo de señal a analizar o de la aplicación en la cual se desee utilizar.

Otra diferencia es que la transformada de Fourier corta STFT permite hacer un análisis tiempo-frecuencia de señales no estacionarias, ya que segmenta la señal utilizando una función-tiempo-ventana, donde la ventana puede ser de tipo cuadrada, Hamming, Hanning, entre otras. El problema está en la rigidez del ancho de la ventana que se mantiene fijo durante todo el análisis de la señal y por lo tanto calcula con la misma resolución tanto frecuencias bajas como frecuencias altas. En cambio la transformada wavelet, mediante las variables de escalamiento y de traslación es capaz de hacer un análisis tiempo-frecuencia con una resolución variable, es decir, utiliza ventanas de diferente ancho durante el análisis de la señal.

Una forma de ver la resolución en el tiempo-frecuencia entre las diferentes transformadas de Fourier y Wavelet en el plano de tiempo-frecuencia se muestra en la figura 2.7 para transformada de Fourier y en la figura 2.8 para la transformada wavelet. En la figura 2.7 se muestra la transformada de Fourier, en el dominio de tiempo-frecuencia de una señal simple. Cada uno de los cuadrados muestra una ventana de la señal en las diferentes descomposiciones de frecuencias; como puede observarse se utiliza la misma ventana para todas las frecuencias, teniendo la misma resolución del análisis en todas las locaciones en el plano de tiempo-frecuencia.

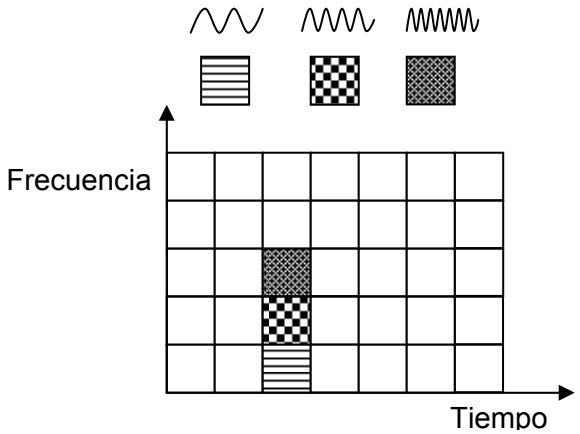

Figura 2.7 Funciones básicas de Fourier, en el plano tiempo-frecuencia.

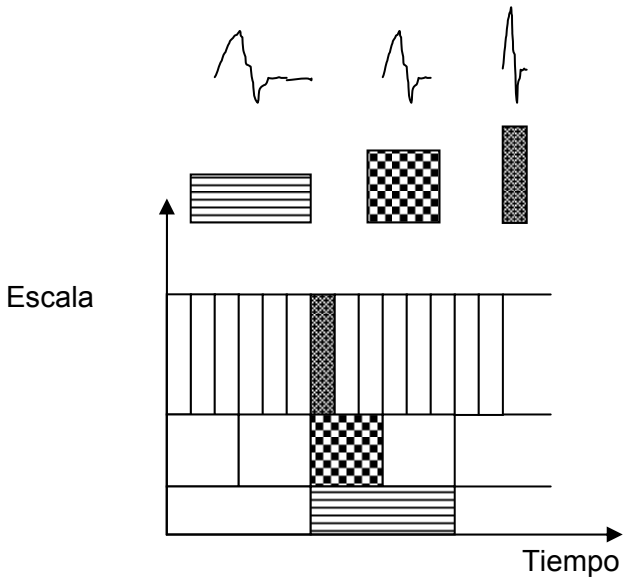

Figura 2.8 Funciones básicas de Wavelet, en el plano escala-tiempo.

Como ya mencionamos, una ventaja de la transformada Wavelet es que las ventanas varían. Para aislar una señal discontinua, uno desearía tener algunas funciones básicas muy cortas. Al mismo tiempo, para obtener un detallado análisis de frecuencias, uno desearía tener algunas funciones básicas largas. Una de las formas para obtener

dichas funciones es tener una combinación de funciones básicas de altas frecuencias cortas y funciones de bajas frecuencias largas. Esta combinación tan deseable es exactamente lo que se obtiene al utilizar la transformada wavelet. La figura 2.8 muestra el alcance en el plano de tiempo-frecuencia con una función wavelet.

Hay que recordar que la transformada Wavelet no tiene un conjunto de funciones básicas como la transformada Fourier, la cual utiliza solamente funciones de seno y coseno. En lugar, la transformada Wavelet tiene un número infinito de funciones básicas.

CAPÍTULO 3

ANTECEDENTES LINGÜÍSTICOS EN LA IDENTIFICACIÓN DEL LENGUAJE HABLADO

Existen importantes estudios desde el punto de vista lingüístico relacionados con la identificación del lenguaje hablado. Los cuales se enfocan en cómo el humano realiza la discriminación entre las lenguas. En esta sección abordaremos el tema de la discriminación de los lenguajes desde este punto de vista.

Los lingüistas, desde un punto de vista diferente al nuestro, han intentado realizar la clasificación de los lenguajes humanos basados en características prosódicas. La prosodia es un término usado típicamente para describir aspectos extralingüísticos del discurso. Ella incluye la entonación, patrones de acentuación, ritmo, melodía, etc. De particular interés para nuestra investigación son los trabajos en la clasificación de los lenguajes basados en su ritmo. Lo cual se detalla en la sección 3.3.1

Los lingüistas parten de fenómenos fonético-fonológicos, los cuales no pueden segmentarse como los fonemas, porque actúan simultáneamente sobre más de un segmento (al menos sobre la sílaba). Estos fenómenos reciben el nombre de suprasegmentales y son tres: el acento, el tono (o la sucesión de ellos, es decir, la entonación) y la duración (o cantidad) 0. El conjunto de estos tres elementos suprasegmentales se denomina prosodia.

La fonología realiza una división entre los fonemas (o fonemas segmentales) y los prosodemas (o suprasegmentos), como el acento, la duración y la entonación [28]. Entre segmentos y suprasegmentos hay una diferencia de clase que resulta evidente: los fonemas son segmentales (o segmentables), uno a uno, mientras que los prosodemas afectan o pueden afectar conjuntamente a varios. Sin embargo, en la realización de los suprasegmentos intervienen índices acústicos y articulatorios que también están presenten en la realización de los segmentos, como:

1. La vibración de las cuerdas vocales que es la fuente de sonoridad de los segmentos sonoros, y también del movimiento del tono fundamental que puede utilizarse en la distinción de las palabras (tono) o de oraciones (entonación).
2. Todo segmento tiene una dimensión temporal, es decir, una duración. Ésta, además, puede desempeñar, en determinadas lenguas, una función distintiva.
3. Todo segmento, al realizarse, ha de tener alguna intensidad. Esta, además, puede desempeñar en algunas lenguas una función distintiva (acento).

Lo anterior muestra las semejanzas entre segmentos y suprasegmentos. Pero entre esos dos elementos hay también una diferencia de grado, que hace que haya que considerarlas como unidades distintas. La diferencia entre dos fonemas no es gradual. Por ejemplo, /p/ se diferencia de /t/ en que una es labial y otra dental. De igual manera, /p/ se diferencia de /b/ por el rasgo de sonoridad. Y un sonido es sonoro o no lo es. Por su parte, el acento, por ejemplo, es gradual: una vocal tona tiene más "fuerza" que una átona, pero no posee ninguna cualidad distinta [28].

Por último, existe una tercera razón para distinguir los segmentos y los suprasegmentos como pertenecientes a dos clases separadas: la función lingüística.

1. La función de los fonemas es distintiva: son unidades que en un contexto dado se excluyen mutuamente (/'pipa/ - /'pepa/ - /'papa/ - /'popa/ - /'pupa/).
2. La función de los suprasegmentos es contrastiva, ya que no pueden alternar en el mismo contexto. En la oposición "amo-amó" lo distintivo es el esquema acentual

/'_ _ / frente a /_ '_/, pero no el acento en sí. El suprasegmento necesita la presencia contrastante de su opuesto en la misma secuencia.

3.1 DEFINICIONES GENERALES

Antes de continuar introduciremos algunos conceptos generales de los estudios lingüísticos del habla 0.

Fonotáctica, trata de la normas y reglas que regulan la combinación de los fonemas de una lengua, por ejemplo, las reglas fonotácticas del español impiden plurales como “clubs” o “films”, siendo las formas correctas, de acuerdo con la fonotáctica de esta lengua clubes y filmes.

Átono, en fonética articulatoria, el adjetivo átono se aplica a las sílabas y vocales que carecen de acento, en las lenguas como el español o el francés las vocales de las sílabas átonas conservan prácticamente el mismo timbre que el de las acentuadas. Pero en otras, como el ruso o el inglés, las átonas tienden a una centralización; en esta última lengua la centralización se materializa en /e/, /i/ o /u/, como se puede comprobar en las sílabas átonas como *language, necessary o plentiful.* Esta característica hace que en la conversación normal se pueda confundir la pronunciación de palabras como *vacation* o *vocation*, aunque en una emisión oral cuidada se puedan diferenciar sin mayor problema.

Tono, existe varias definiciones, una de ellas define al “tono” como la *altura musical* de cada sílaba. Tradicionalmente al tono se le ha llamado acento melódico. Vistos desde la fonética articulatoria, los “tonos” constitutivos de la entonación, se forman por la vibración de las cuerdas vocales y se mueven en la escala de mayor a menor vibración (agudo-grave), mientras que los acentos, componentes de las pautas rítmicas, se mueven en la de mayor a menor intensidad (tónico-átono); cuando mayor sea la vibración, tanto más agudo será el tono. Acústicamente los “tonos” están relacionados con la frecuencia. Los “tonos”

pueden ser estables y dinámicos; los primeros se mueven en la misma línea o dirección, es decir, no cambian de agudo a grave o de grave a agudo y pueden ser altos, medios y bajos. Los segundos también son conocidos con el nombre de tonemas. En los últimos años del siglo XX, el análisis y la representación del "tono", con una nueva aproximación, han sido objeto de estudio de la llamada *fonología autosegmental*. Cada persona tiene su *tono normal* de voz, es decir, la nota que dentro de su registro individual se produce con más naturalidad y menor fatiga. En torno a ella se suceden los movimientos ascendentes y descendentes. Se comprueba que, descartando las diferencias individuales, las gentes de determinadas regiones o países suelen expresarse en un tono normal medio más agudo o más grave.

También se llama "tono" a la función distintiva que cumple la *frecuencia fundamental* en el nivel de la palabra [28]. De la misma manera que en el español el acento tiene una función distintiva, como recurso de diferenciación léxica, por ejemplo, pérdida y perdida, inglés e ingles, entre otras. Existen lenguas como el Chino o el Vietnamita, que se sirven del tono para estos fines. Por ejemplo, /ma/ puede significar varias cosas distintas, desde *madre* hasta *caballo*. Con un "tono" estático alto significa *madre*, con un "tono" dinámico ascendente significa *cáñamo*, con un "tono" dinámico descendente-ascendente significa *caballo*, y con un tono descendente significa *riña* [29]. A las lenguas que usan el tono como recurso en la formación de las palabras se las llama lenguas tonales, por ejemplo: las lenguas de la familia congo-nigeriana, sino-tibetanas y algunas de las lenguas indígenas de México (otimí, mazahua, pame y chichimeca entre otras).

El tono fundamental depende, básicamente, de las vibraciones de las cuerdas vocales; pero, además, hay una serie de factores fonéticos que la condicionan:

1. Existe una relación entre la cualidad o el timbre de la vocal y la altura relativa de su frecuencia fundamental, de modo que las vocales más altas / [i], [e]/ tienen un tono fundamental más elevado.
2. Las frecuencias fundamentales más altas aparecen después de las consonantes sordas, y las más bajas, tras las consonantes sonoras.

3. Además del tono fundamental, la duración y la intensidad también intervienen en la producción y la percepción de la entonación.

Isocronía, isosilabicidad. En fonética articulatoria, se llama isocronía a la tendencia que se tiene en ciertas lenguas, por ejemplo el inglés, a dejar el mismo lapso de tiempo entre dos sílabas tónicas, con independencia del número de sílabas átonas que haya entre ellas. De acuerdo con esta teoría, en los siguientes enunciados habría isocronía: “John said his foot`s bad”, “John said his footman is bad”, “John said his footballer is bad”, O dicho en otras palabras, el hablante nativo del inglés habrá tendido a dejar el mismo tiempo entre John, said, foot`s (footman, footballer) y bad, a pesar de que entre ellas haya dos, tres o más sílabas átonas o ninguna. La cuestión de la isocronía es controvertida, aunque es muy importante dentro de los estudios del ritmo. No todos los lingüistas están dispuestos a aceptar la isocronía total de una lengua, en este caso el inglés, pero la gran mayoría reconocen que existe la tendencia.

La transcripción fonética. Transcribir es representar por medio de signos alfabéticos la complejidad de la cadena hablada. Hay toda una teoría de la transcripción, desarrollada por disciplinas como la etnografía, la musicología, la dialectología, etc.

En nuestro ámbito de estudio, la transcripción puede ser, en primer lugar, fonológica o fonética:

1. La transcripción fonológica refleja la expresión en el plano de la lengua, es decir, la constitución fonológica de la emisión lingüística. Se ocupa, por lo tanto, solamente de lo distintivo.
2. La transcripción fonética refleja la expresión en el plano del habla, es decir, las variantes utilizadas por el hablante, independientemente de su valor distintivo.

La transcripción fonética puede ser estrecha o ancha, la transcripción estrecha intenta recoger la mayor cantidad posible de información fonética. En cambio, la transcripción ancha o amplia anota sólo los rasgos que contribuyen a la significación, se aproxima a la fonológica.

La transcripción, ya sea fonológica o fonética, precisa un alfabeto fonético. No hay un único alfabeto fonético. En el ámbito lingüista, se han utilizado mayormente dos: el de la RFE y el de la IPA. El alfabeto de la RFE (Revista de Filología Española) fue creado en 1915 para ser utilizado por esta revista, y ha sido desde entonces tradicionalmente utilizado por la filología española (y las lenguas de su entorno). Está expresamente diseñado para el español, y resulta muy cómodo de utilizar, pero no permite la comparación con el resto de las lenguas (ni con los trabajos internacionales). El alfabeto de la IPA (International Phonetic Association) fue creado por la Asociación Fonética Internacional en 1889. Pretende servir a todas las lenguas, por lo que admite pequeñas adaptaciones particulares. Ha experimentado continuos cambios, mejoras y adaptaciones. Hoy es comúnmente aceptado como el estándar mundial de transcripción. La última versión es de 1993.

En los últimos años se ha extendido, especialmente en el ámbito de las tecnologías del habla, la utilización del alfabeto SAMPA (Speech Assessment Methods Phonetic Alphabet), que es una especie de versión reducida del IPA, pensada expresamente para ser leída por ordenadores, mediante caracteres ASCII de 7 bits. Para evitar los problemas derivados de la incompatibilidad entre las distintas versiones del SAMPA, y para obtener además otra serie de ventajas, se creó el alfabeto X-SAMPA, que pretende, como el IPA, servir para todas las lenguas.

3.1.1 EL ACENTO

El acento es un rasgo suprasegmental que recae sobre una sílaba de la cadena hablada y la destaca o realza frente a otras no acentuadas (o átonas) [28].

Esta prominencia silábica suele interpretarse tradicionalmente como reflejo de intensidad; por eso, se le ha llamado "acento de intensidad". La realidad, sin embargo, es más compleja: la prominencia resulta de la conjunción de varios factores articulatorios:

1. Una mayor fuerza respiratoria, que genera una mayor intensidad.
2. Una mayor tensión de las cuerdas vocales, que genera una elevación del tono fundamental.
3. Una mayor prolongación en la articulación de los sonidos, que supone un aumento de la duración silábica.

Así pues, la sílaba tónica, habitualmente, es más intensa, más alta y más larga que las sílabas átonas adyacentes. En español, el índice acústico primario del acento es el tono, aunque los otros dos índices (intensidad y duración) también colaboran en la acentuación, en proporciones variables.

La mayoría de las palabras poseen una sílaba tónica y otras átonas. Sólo algunos monosílabos pueden considerarse palabras átonas. Cuando las palabras son más largas, una sílaba posee el acento principal y otra el acento secundario. Dentro de una frase, el último acento principal se denomina acento de frase.

En cuanto a la posición que la sílaba acentuada ocupa dentro de la frase, algunas lenguas son de acento libre, es decir, no hay manera de prever en qué sílaba recae el acento; otras, por el contrario, son de acento fijo, es decir, la posición del acento es siempre previsible. Un ejemplo del primer tipo es el inglés, donde pueden encontrarse formas como "*accent*", que significa una cosa u otra según la posición del acento: significa 'acento' cuando lleva el acento en la primera sílaba y 'acentuar' cuando lo lleva en la segunda. Un ejemplo del segundo tipo lo constituye el francés, donde prácticamente todas las palabras son agudas.

Muchas lenguas no corresponden exactamente a ninguno de esos dos tipos; por ejemplo, el español es de acento libre (pueden incluso presentarse oposiciones del tipo "cántara / cantara / cantará"), pero tiene una marcada tendencia hacia la acentuación llana (casi el 80% de sus palabras se acentúan en la penúltima sílaba).

En las distintas lenguas del mundo, el acento puede tener las siguientes funciones lingüísticas:

1. Contrastiva: distingue sílaba tónicas/átonas en el eje sintagmático. Por ejemplo: "El libro es de él".
2. Distintiva: distingue unidades en el eje paradigmático (en lenguas con acento libre). Por ejemplo: "amo"/"amó".
3. Demarcativo: en lenguas de acento fijo, señala los límites de las unidades en la secuencia. Por ejemplo: el final de una palabra en turco.
4. Culminativa: en las lenguas de acento libre, señala la presencia de una unidad acentual, sin indicar sus límites.

3.1.2 LA ENTONACIÓN

La entonación es uno de los componentes más complejos de una lengua. Se ha definido de muchas maneras, dependiendo básicamente del interés de cada autor: por el tono fundamental, por una conjunción de parámetros acústicos (tono, acento y duración, primordialmente), por su función lingüística, etc.

Quilis [28] define la entonación como "la función lingüísticamente significativa, socialmente representativa e individualmente expresiva de la frecuencia fundamental en el nivel de la oración".

La entonación, como todo enunciado lingüístico, presenta una sustancia y una forma. La sustancia es un *continuum* en el que hay que delimitar las unidades de entonación, de modo que se obtengan elementos discretos para establecer así sus patrones melódicos y la naturaleza de sus elementos.

Entre el nivel de la sustancia y el nivel de la forma, nos encontramos con toda una serie de niveles o grados de abstracción, elegidos arbitrariamente por cada investigador según el fin que se proponga.

Desde el punto de vista articulatorio, el tono depende básicamente de las cuerdas vocales: de su longitud, su grosor su tensión.

Según la utilización lingüística del tono, las lenguas se dividen en tonales y entonadas:

1. Las lenguas tonales utilizan los tonos para distinguir significados. Cumple, entonces una función distintiva en el léxico. Por ejemplo, el chino, el tailandés.

2. Las lenguas entonadas utilizan la sucesión de tonos, es decir, la curva melódica de la entonación, no ya para distinguir significados léxicos, sino para modificar significaciones secundarias (expresividad, intencionalidad, etc.). Cumple, entonces una función expresiva en la frase. A este tipo de lenguas pertenecen todas las románicas.

3.1.3 LA DURACIÓN

La duración es también un fenómeno segmental, puesto que cada sonido posee una duración propia. Así por ejemplo, es sabido que las consonantes fricativas son más largas que las oclusivas, que las sordas son las más largas que las sonoras, etc.

Algunas lenguas poseen pares de fonemas en función de la duración. Por ejemplo, el italiano distingue entre ciertas consonantes breves y largas o "dobles". El latín clásico distinguía entre vocales breves y largas.

De acuerdo a la articulación, la duración se basa en el mantenimiento por más o menos tiempo de una determinada configuración articulatoria. Por el fenómeno de la coarticulación (la cual describimos con un ejemplo dado anteriormente, es diferente la pronunciación de “to” en las palabras: *to*do y en es*to*rnudo), dicha configuración (y, consiguientemente, la duración) se ve alterada en función del contexto.

Como elemento suprasegmental, tanto las sílabas tónicas como las pertenecientes al fonema suelen ser más largas. Anteriormente se han explicado las unidades fonéticas, segmentales y suprasegmentales. Todos esos elementos se combinan en la cadena hablada, dando lugar a una serie de fenómenos de gran complejidad, que se suelen englobar bajo el nombre de fonosintaxis.

La fonosintaxis es el estudio de las modificaciones que sufren los fonemas al agruparse en la cadena hablada. El concepto básico aquí es el de coarticulación: los sonidos no se pronuncian aislados, y la proximidad articulatoria de unos con otros hace que se influyan mutuamente.

Dentro de la cadena hablada, los segmentos se agrupan en unidades cada vez mayores: sílabas, palabras y enunciados.

A pesar de que resulta difícil definir la sílaba, hay pruebas evidentes de que el hombre ha sentido y manifestado la existencia de la sílaba: la escritura fue silábica antes que fonológica. Además, los semianalfabetos dividen las palabras (o las frases) en sílabas sin titubeos. Y en sílabas fonéticas (por ejemplo: "es-tá-re-na-pu-ros"), no en las derivadas de la ortografía ("es-tár-en-a-pu-ros"). Por otro lado, las palabras cantadas se dividen en sílabas, nunca en fonemas; y el ritmo poético descansa sobre el número de sílabas, no de fonemas [30].

Desde la antigüedad se ha considerado que la capacidad de formar sílabas (o, más exactamente, el núcleo de las mismas) es una de las características básicas a la hora de diferenciar entre vocales y consonantes. Sin embargo, no sólo las vocales pueden formar sílabas (o su núcleo): hay muchas lenguas en las que ciertas consonantes (líquidas,

nasales) pueden hacerlo. Por ejemplo, el núcleo de la última sílaba de la palabra inglesa "*people*" lo constituye la [*l*].

Desde el punto de vista acústico, los fonemas situados antes del núcleo silábico presentan un aumento de su intensidad, sonoridad y perceptibilidad, hasta llegar al máximo que constituye el núcleo. De igual manera, los fonemas que se encuentran detrás del núcleo presentan una disminución de dichas características, a partir del máximo constituido por el núcleo.

Desde el punto de vista articulatorio, los fonemas anteriores al núcleo silábico experimentan una abertura gradual de los órganos articulatorios, hasta llegar al máximo del núcleo, a partir del cual los fonemas experimentan un cierre. Lo mismo cabe señalar de la tensión articulatoria y de la presión del aire aspirado.

De acuerdo a lo anterior, la frontera o el límite silábico ha de estar situado donde se produce un mínimo entre dos máximos (es decir, los núcleos de las dos sílabas entre las que se establece el límite):

Los mínimos y máximos, como se ha dicho, corresponden a la intensidad, a la sonoridad, a la presión respiratoria, a la tensión muscular e, incluso, a la energía articulatoria general.

El límite silábico desempeña una función distintiva en las lenguas en las que forzosamente coincide con el límite entre morfemas. Por ejemplo, en inglés es el límite silábico el que distingue entre [ə 'neim] (*a name*, 'un nombre') y [ən 'eim] (*an aim*, 'un objetivo').

Los sonidos se agrupan, como hemos visto, en unidades cada vez mayores: la sílaba -que no suele considerarse objeto específico de la fonosintaxis-, la palabra y la oración. Sin embargo, la fonosintaxis distingue otra unidad, intermedia entre las dos últimas: el sirrema.

3.1.4 EL SIRREMA

El sirrema es "la agrupación de dos o más palabras que constituyen una unidad gramatical, unidad tonal, unidad de sentido y que, además, forman la unidad sintáctica intermedia entre la palabra y la frase" [28].

Las palabras que constituyen el sirrema permanecen siempre unidas: entre ellas no puede haber pausa. La razón de ser de dicha unidad es acentual: el sirrema aglutina a una serie de elementos silábicos átonos que no pueden producirse aislados, sino en torno a alguna otra sílaba acentuada, para formar con ella una unidad indisoluble.

En general, cada lengua tiene su propio inventario de las partes de la oración que forman sirrema. Fuera de esas combinaciones, las demás agrupaciones están sujetas a una gran variabilidad en lo referente a pausas y entonación. En español, forman sirrema las siguientes partes de la oración:

El artículo y el sustantivo. Por ejemplo: el carro (/el'kaʁo/).

1. El pronombre átono y el elemento gramatical que le antecede. Por ejemplo: dile que venga (/dile ke 'beNga/).
2. El adjetivo y el sustantivo, o viceversa. Por ejemplo: perro blanco (/'peʁo'blaNko/).
3. El sustantivo y el complemento determinativo. Por ejemplo: el perro de Javier(/el 'peʁode'javier/).
4. Los tiempos compuestos de los verbos. Por ejemplo: he comido (/'eko'mido/).
5. Los elementos de una perífrasis o una frase verbal. Por ejemplo: hemos dejado de ser (/'emosde'xadode'ser/).
6. El adverbio y verbo, adjetivo o adverbio. Por ejemplo: los más destacados alumnos (/los'masdesta'kadosaluNnos/).
7. La conjunción y la parte del discurso que la introduce. Por ejemplo: Juan y Pedro (/'xuan i'pedRo/).
8. La preposición con su término. Por ejemplo: voy con Juan (/'boi koN'xuan/).

3.2 EL RITMO

El término ritmo puede tener en lingüística, al menos, dos acepciones 0:

1. En un sentido amplio se llama ritmo a las sensaciones auditivas que se perciben a los intervalos regulares de tiempo, producidas por repeticiones isofónicas de cualquier recurso prosódico del lenguaje, como puede ser la rima, la censura, etc.

2. En un sentido estricto, el ritmo es un prosodema básico de la cadena hablada, junto con la entonación y el acento. Aún siendo conscientes, por una parte, de que lo que realmente se percibe auditivamente es una prominencia, conviene separar, en lo posible, los rasgos de tensión y los de melodía que se manifiestan en la cadena hablada; los rasgos de melodía corresponden a la entonación y los rasgos de tensión corresponden al ritmo (también llamado ritmo verbal para diferenciarlo del ritmo musical).

El ritmo de un grupo fónico es la pauta de tensión formada en el mismo por la combinación de sílabas tónicas y átonas, y largas y breves. El ritmo es uno de los prosodemas o fonemas prosódicos (o suprasegmentales) más característicos de una lengua. Como no todas las lenguas hacen el mismo uso de las sílabas largas y breves, y de las tónicas y átonas, habrá distintos tipos de ritmos; los más importantes son el acentual y el silábico.

Ritmo acentual (o *stress-timed*) quiere decir que las pautas que se forman en el grupo fónico tienen un *tempo* marcado por el acento, o sea, están acompasadas por el acento, mientras que en el **ritmo silábico** (o *syllable timed*) es la sílaba la que sella el *tempo*, es decir, el ritmo está acompasado por la sílaba; el del inglés es de tipo acentual, mientras que el del español es silábico.

Las vocales del inglés poseen dos rasgos peculiares que no existen en el español, y que contribuyen a que las diferentes pautas rítmicas sean distintas a las del español:

a) pueden ser largas y breves, con lo que unas sílabas tendrán mayor duración que otras; y
b) en las sílabas átonas pierden su timbre pleno.

En cambio, en español los fonemas vocálicos poseen la misma duración aproximadamente, y la diferencia entre los timbres de las vocales tónicas y átonas apenas es perceptible. Como ya mencionamos anteriormente, para el oído inglés el "ritmo" español resulta marcial, ya que da timbre pleno a todas las vocales de las sílabas. En cambio, al español, el "ritmo" inglés le produce un efecto "entrecortado" y sujeto a "tirones", el de esta lengua se caracteriza, además, por la *isocronía*, es decir, por la tendencia a dejar el mismo tiempo entre dos sílabas tónicas, con independencia del número de sílabas átonas que hay entre las dos tónicas.

El "ritmo" es probablemente el rasgo de la *base articulatoria* de una lengua cuya adquisición o dominio resulta más difícil al estudiante adulto de un idioma extranjero y, aunque la inteligibilidad depende en gran parte de su correcta emisión, a éste no se le presta la atención debida en la enseñanza de idiomas extranjeros [12]; por ejemplo, aunque el hablante inglés haga esfuerzos por entender al extranjero que use pautas rítmicas diferentes a las inglesas se corre el riesgo de que la comprensión del mensaje quede interrumpida, si se aplican indiscriminadamente pautas rítmicas castellanas al hablar inglés. Y, paradójicamente, es de los primeros rasgos que se aprenden en la infancia durante la adquisición de la lengua materna.

3.2.1 CLASIFICANDO LOS LENGUAJES HUMANOS A TRAVÉS DE SU RITMO

Los lingüistas, desde un punto de vista diferente al nuestro, han intentado realizar la clasificación de los lenguajes humanos basados en las características suprasegmentales, específicamente el ritmo. El ritmo lo definimos como la organización

sistemática de unidades prominentes y no prominentes del habla en unidades de tiempo [31]. Una unidad, dependiendo del lenguaje, puede ser una sílaba o un intervalo vocálico; y la prominencia está dada por su duración, su intensidad o una frecuencia fundamental alta. Es posible hablar de patrones rítmicos en el habla específicos o no a un lenguaje. Nuestro interés recae en los patrones rítmicos distintivos de un lenguaje. Los primeros intentos por clasificar a los idiomas con base en el ritmo recaen en la habilidad documentada en infantes para distinguir lenguajes [32]. Estos primeros intentos distinguían dos clases [12]:

- los lenguajes mostrando patrones de igual duración entre sílabas prominentes (sílabas acentuadas) a esta clase se le conoce como “stress timed”, (por ejemplo, el inglés o el alemán) y

- los lenguajes con sílabas de igual duración o “syllable timed”(por ejemplo, el francés o el español)

Sin embargo, investigaciones posteriores encontraron que esta clasificación era demasiado restrictiva. Lo que es más se llegó afirmar que el ritmo era sólo un fenómeno perceptual imposible de extraer de la señal acústica. Trabajos recientes han buscado redefinir el concepto de ritmo para clasificar los lenguajes. Ramus et al [33] introdujo una nueva forma de medición del ritmo. Esta se basa en la duración de los intervalos vocálicos y consonánticos. Un intervalo vocálico es un fragmento de la señal de habla asociado a una o varias vocales (diptongos) del idioma en cuestión. De manera similar, un intervalo consonántico es un fragmento donde se realizan únicamente consonantes. En particular los trabajos basados en la proporción relativa de las vocales han obtenido resultados interesantes [33]. En estos trabajos las medidas para describir el ritmo se basan en tres conceptos principales:

- La proporción de los intervalos vocálicos en una oración, que es la suma de los intervalos vocálicos entre la duración total de la oración (%V).

- La desviación estándar de la duración de los intervalos vocálicos en cada oración (ΔV).

- La desviación estándar de la duración de los intervalos consonánticos en cada oración (ΔC).

A partir de ellas Ramus et al [33] pudieron determinar el ritmo de algunos lenguajes. La gráfica 3.1 muestra la distinción de los lenguajes en función de la proporción relativa de los intervalos vocálicos (%V) y su desviación estándar (ΔV). A partir de las ideas de este trabajo nuevos esquemas de clasificación han sido planteados [34][35]. Estas ideas han sido probadas por Rouas et al [22] en 2002 y [36] en 2005. Es por ello que es uno de los sistemas con los cuales comparamos nuestros resultados.

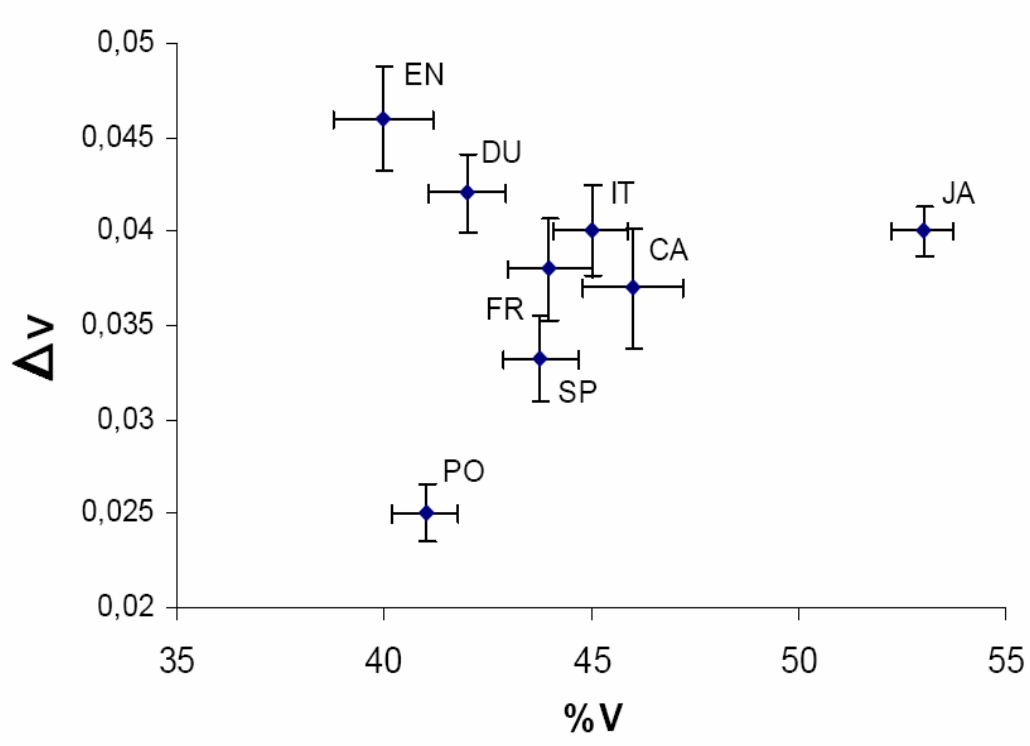

Figura 3.1 Distinción rítmica de los lenguajes en función de sus intervalos vocálicos (tomada de [33])

Es a partir de estos estudios que podemos orientar nuestra investigación para la integración del ritmo en la caracterización de la señal acústica. Por supuesto, aun falta trabajo por realizar. En todos estos trabajos el interés no es crear un sistema automático, de ahí que la segmentación e identificación de fonemas (más aun la definición de un fonema) se realizó de manera manual sobre un conjunto reducido de grabaciones (exceptuando el caso de Rouas). Otros trabajos relacionados con ritmo son los de Galves y Steiner acerca de la sonoridad [37][38]. Pero nuestro trabajo pretende identificar un idioma sin hacer un reconocimiento de los intervalos vocálicos y consonánticos de la señal, por lo que pretendemos basarnos sólo en la señal de voz, extrayendo el ritmo de otra manera.

3.3 LA IDENTIFICACIÓN DE IDIOMAS POR LOS SERES HUMANOS

Otro estudio relacionado con la discriminación de idiomas es el llevado por Muthusamy. En él se comprueba la capacidad de los seres humanos para la identificación del lenguaje. Con escuchar la voz unos segundos, la gente es capaz de determinar de que lenguaje se trata, siempre y cuando conozcan el lenguaje en particular; y en el caso de que sea un lenguaje que ellos no están familiarizados, pueden realizar un juicio subjetivo de acuerdo a los lenguajes similares que ellos conocen, por ejemplo, suelen decir: “suena parecido al alemán”. De acuerdo a esto, Muthusamy en 1994 [39] realizó un estudio para obtener las mejores marcas que tienen los humanos en la identificación del lenguaje hablado. Sus pruebas consistieron en dos casos: la identificación del lenguaje hablado por personas monolingües, es decir, que sólo conocen su lengua materna; y el segundo caso por personas que conocen varios lenguajes. Para el experimento se analizaron 28 personas (14 mujeres y 14 hombres); de los cuales fueron 10 hablantes nativos del lenguaje inglés y 2 personas para cada uno de los 9 lenguajes restantes. Todas las personas conocían el lenguaje inglés. Las personas podían escuchar 10 segundos de señal de voz espontánea, del corpus OGI_TS [21], tantas veces como ellos desearan.

Después de dos o tres días las personas tenían que identificar el lenguaje de una muestra específica. Los resultados se muestran en las figura 3.2.

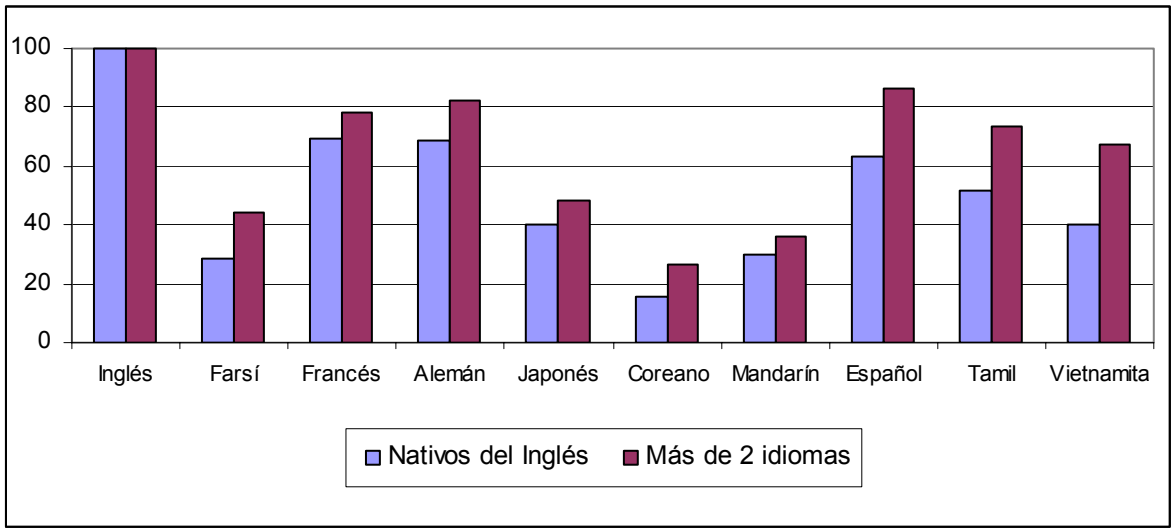

Figura 3.2 Porcentaje de identificación del lenguaje hablado por hablantes nativos del inglés y los que hablan más de dos idiomas en 6 segundos de señal de voz (tomada de [39]).

El porcentaje de reconocimiento para personas que conocían 4 idiomas fue de 66.7%, para personas que conocían 3 idiomas 57.9%, para las personas que conocían 2 fue de 51.1% y las que sólo conocían uno fue de 44.1%. Ver figura 3.3. Con esto Muthusamy [39] concluyó que el porcentaje de discriminación de los lenguajes aumenta cuando se tiene conocimiento de más lenguas. Lo importante de esto es notar que incluso la identificación del lenguaje hablado hecho por los humanos no tiene grandes porcentajes de discriminación, aún en el caso de personas que dominan cuatro idiomas.

3.4 CONCLUSIONES

Podemos concluir, de acuerdo a los lingüistas, que existe información importante para la distinción de los idiomas en el ritmo, la entonación y la duración, conceptos muy ligados entre sí. De ahí que un esquema de discriminación que incluya información suprasegmental tendrá una mayor pertinencia que limitarse únicamente a fonemas segmentales.

Los lingüistas mencionan que la frecuencia fundamental de la voz depende, básicamente, de las vibraciones de las cuerdas vocales, definiéndolo como tono fundamental; el cual tiene una serie de factores fonéticos que la condicionan. Como por ejemplo, las vocales más altas tienen un tono fundamental más elevado; o que las frecuencias fundamentales más altas aparecen después de las consonantes sordas, y las más bajas, tras las consonantes sonoras. Además del tono fundamental, la duración y la intensidad también intervienen en la producción y la percepción de la entonación.

Por otro lado, de acuerdo al procesamiento de la señal de voz, la frecuencia fundamental comúnmente llamada “picth” en inglés, es la frecuencia mas baja –véase capítulo 2. Entonces podemos asumir que en las frecuencias bajas hay información relevante para la identificación del lenguaje hablado. Las cuales podrían representar al ritmo, la entonación y la duración, en general las características suprasegmentales que usamos al hablar.

CAPÍTULO 4

ESTADO ACTUAL EN LA IDENTIFICACIÓN DEL LENGUAJE HABLADO

La investigación en la identificación automática del lenguaje hablado tiene una historia que abarca desde los años 70´s, a lo largo de esos años se han desarrollado diferentes enfoques de cómo resolver esta tarea; todos esos diferentes enfoques tienen en forma general dos fases para la identificación del lenguaje hablado, los cuales se muestran en la figura 4.1 (Zissman y Berkling 2001 [40]). Durante la fase de entrenamiento, el típico sistema es presentado con ejemplos de muestras de voz de una gran variedad de lenguajes. Cada muestra de entrenamiento es convertida a una cadena de vectores de características. Cada vector de características proviene de ventanas de la señal digitalizada (por ejemplo ventanas cortas de 20ms) durante los cuales la señal de voz se asume estática. Los vectores de características contienen información espectral o cepstral acerca de la señal de voz (el cepstrum es el inverso de la transformada de Fourier de magnitud logarítmica del espectro, véase el capítulo 2). El algoritmo de entrenamiento analiza una secuencia de dichos vectores y produce uno o más modelos para cada lenguaje. Esos modelos capturan las características propias del lenguaje. Durante la fase de reconocimiento, el vector de características procesado de una nueva señal de voz es comparada con cada modelo. Después se calcula la probabilidad de que esta nueva señal

pertenezca a alguno de los modelos obtenidos al entrenar. Este cálculo se repite para cada uno de los lenguajes entrenados y finalmente se considera el lenguaje de la muestra aquel con la máxima probabilidad.

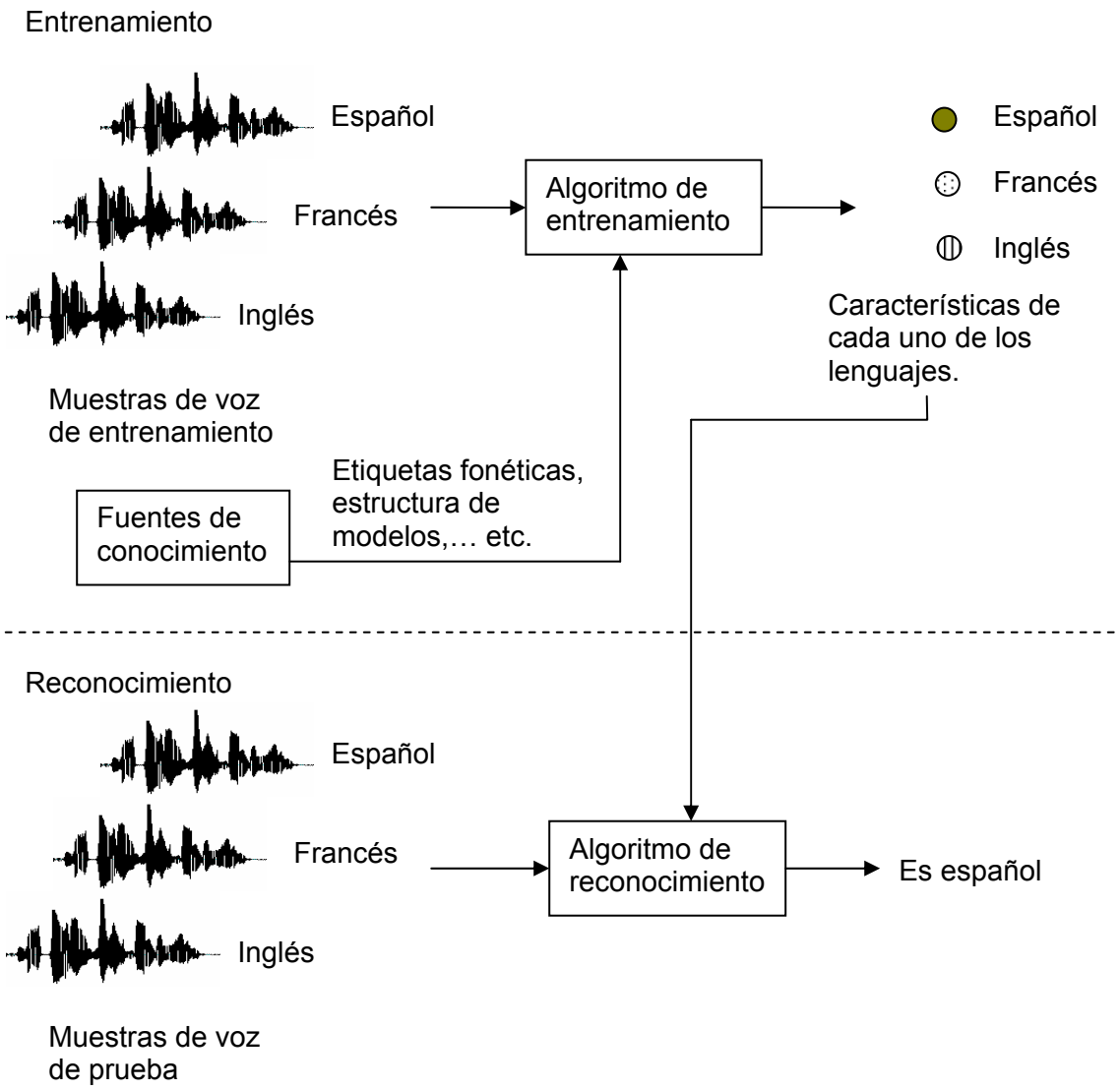

Figura 4.1 Las dos fases en la identificación del lenguaje hablado (tomado de [40]).

Los sistemas varían primeramente de acuerdo al método que utilizan para modelar los lenguajes. Por lo que existen diferentes métodos para extraer características de la señal de voz, esto ha incrementado la cantidad de conocimiento haciendo que los sistemas de identificación del lenguaje hablado sean cada vez más complejos. Durante el entrenamiento, algunos sistemas requieren solamente la señal de voz digitalizada y sus correspondientes características de cada uno de los lenguajes, porque los modelos de lenguaje están basados simplemente en la representación de la señal de voz. Los sistemas de identificación del lenguaje hablado más complejos usan fonemas para modelar la señal de voz, esto puede requerir entre otras cosas: una transcripción fonética (secuencia de símbolos representando los sonidos al hablar), o una transcripción ortográfica (el texto de las palabras) junto con un diccionario de la transcripción fonética (trazando las palabras con un prototipo de pronunciación) para cada muestra de señal de voz de entrenamiento. Producir esas transcripciones y diccionarios es muy complejo, el proceso consume mucho tiempo y usualmente requiere de lingüistas expertos en el idioma de interés.

En los comienzos de la identificación automática del lenguaje hablado, las investigaciones eran capitalizadas en las diferencias espectrales contenidas en los diferentes lenguajes, explotando el factor de que los idiomas contienen diferentes fonemas. A este tipo de sistemas se les conoce como ***Procesos basados en similitudes espectrales*** [40]. Para entrenar esos sistemas, se procesa un conjunto de muestras de señal de voz obteniendo un conjunto de espectros prototipos de cada lenguaje. Durante el reconocimiento, ese conjunto de espectros prototipos de las muestras de entrenamiento es comparado con el espectro de la muestra de prueba. El lenguaje es identificado de acuerdo al que obtuvo mejores resultados al compararlo con el conjunto de espectros prototipos.

Existen variaciones en la obtención del conjunto de espectros prototipos, pueden ser obtenidos directamente como vectores de características o pueden ser usados para procesar un formato base de características.

El cálculo de la similaridad entre una muestra de señal de voz de prueba contra una muestra del modelo de entrenamiento, en los inicios de los sistemas de similitud-espectral

consistía en calcular la distancia entre cada una de las muestras de entrenamiento y la muestra de prueba. El resultado de la distancia entre cada uno de ellos era acumulado en una distancia total. Estos sistemas basados en la cuantificación-vectorial aplicados para la identificación del lenguaje hablado fueron introducidos por Riek et al [41], Nakagawa et al [42].

Considerando que los sistemas de identificación del lenguaje hablado descrito anteriormente realizan la clasificación principalmente estática, los modelos de Markov (Hidden Markov Models - HMMs por sus siglas en inglés) fueron aplicados en la Identificación del lenguaje hablado, ya que los HMMs tienen la habilidad de modelar características secuenciales de la producción del habla. El uso de los modelos de Markov fue propuesto inicialmente por House y Neuburg [43]. Anteriormente Savic et al [44] y Zissman [45] aplicaron HMMs a los vectores de características espectrales y cepstrales. En dichos sistemas, los HMM de entrenamiento fueron realizados sobre muestras de señal de voz no etiquetadas. Riek y Zissman encontraron que los sistemas entrenados con HMM no supervisados no realizaron la tarea tan bien como algunos de los clasificadores estáticos. Aunque Nakagawa et al [46] eventualmente obtuvieron mejores resultados para sus HMM que para sus clasificadores estáticos.

Después de los procesos basados en similitudes espectrales, se desarrollaron diferentes enfoques. Para generalizar hemos divido en dos enfoques la forma de resolver el problema de identificar automáticamente los idiomas, uno que se basa en la representación fonética de la señal de voz, es decir en la segmentación de fonemas y sus subsecuentes procesos, y el otro en donde sólo se utilizan las características acústicas de la señal de voz para la identificación de los idiomas, y como se ha mencionado anteriormente, este último hasta nuestros días no ha tenido resultados comparables a los del primer enfoque. Por lo tanto, se ha divido el estado del arte en estos dos enfoques: los que utilizan el reconocimiento fonético y los que no lo utilizan. En cada uno de ellos intervienen diferentes formas de solucionar la tarea. Las siguientes dos secciones tratan a detalle cada uno de estos enfoques.

4.1 SISTEMAS CON RECONOCIMIENTO FONÉTICO

La figura 4.2 muestra los componentes básicos para la identificación del lenguaje hablado basado en la representación fonética de la señal de voz. La identificación del lenguaje hablado se realiza en tres etapas, en forma general:

- **El procesamiento acústico de la señal de voz,** el primer proceso de la identificación del lenguaje es la transformación de la señal de voz continua en una secuencia de eventos discretos. Es a partir de estas secuencias de eventos que será posible identificar los fonemas dentro de la señal. La señal de voz refleja la configuración del tracto vocal del hablante. Por lo tanto, cuando producimos diferentes sonidos generamos variaciones pequeñas sobre periodos cortos de tiempo durante configuraciones de tracto vocal uniformes. Las formas de las ondas acústicas pueden entonces ser segmentadas en pequeñas variaciones. El resultado es un conjunto de intervalos segmentados de los sonidos, los cuales son llamados fonemas y corresponden directamente a diferentes sonidos en el lenguaje de cada hablante. Además cada fonema puede contener diferentes pronunciaciones o alófonos los cuales son propios de cada lenguaje.

- **Alineamiento de fonemas**, el alineamiento de una señal de voz continua es el primer paso hacia la captura de las características estructurales de niveles altos. Por lo tanto, es importante encontrar un conjunto de unidades del habla que capturen la cantidad apropiada de detalle para la aplicación. Existen diferentes métodos para el alineamiento de fonemas para múltiples lenguajes, tales como, los *clusters* de fonemas, mapeo de fonemas, mezcla de lenguajes, etc.

- **Extracción de características estructurales**, después de que la señal de voz continua es decodificada en una secuencia de eventos discretos, el siguiente paso generalmente captura las características estructurales, tales como, las palabras, subpalabras, sílabas, etc. En el método basado en la dependencia del lenguaje se obtienen secuencias de fonemas válidas de cada lenguaje o categorías de fonemas.

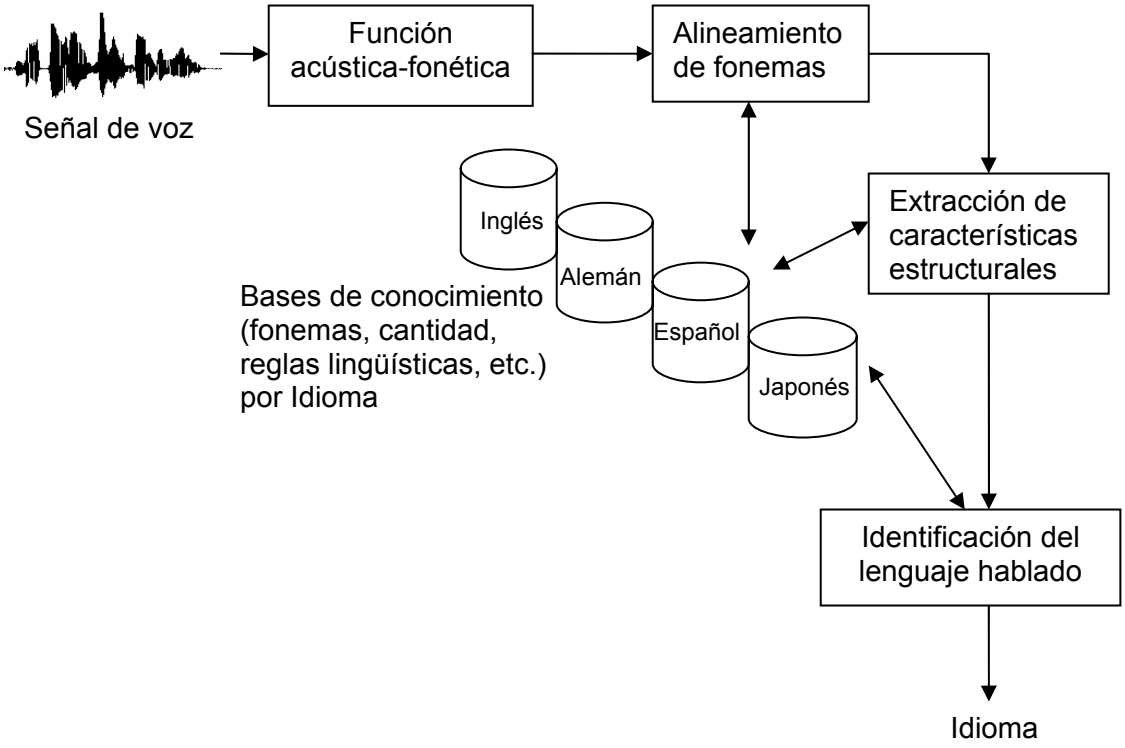

Figura 4.2 Componentes básicos de un sistema de identificación del lenguaje hablado basado en reconocimiento fonético.

Como podemos observar, Este tipo de sistemas utiliza dos tipos de información: la fonética y la fonotáctica, y consta de dos pasos básicos:

- Reconocimiento fonético. Se utilizan para cada lenguaje a identificar un conjunto de modelos acústicos (uno para cada fonema) para transformar la señal acústica en la secuencia de fonemas más probable. Para la construcción de dichos modelos acústicos es necesario contar con un gran número de grabaciones. Estas grabaciones deben cubrir todos los fonemas del idioma así como sus alófonos

(variantes de los fonemas dado su contexto, el acento regional, etc.). Ahora bien, dichas grabaciones deben estar etiquetadas manualmente por expertos a nivel fonético, es decir, cada grabación debe ser escuchada por el experto para determinar las fronteras entre los fonemas. Esta segmentación manual de la señal es una tarea muy costosa.

- Modelado del lenguaje. Es a partir de la secuencia de fonemas que propiamente se realiza la tarea de identificación. Para ello se comparan las proporciones de los fonemas en la secuencia contra los modelos de los lenguajes a identificar. Un modelo de lenguaje es un listado de todas las posibles combinaciones de dos, tres o *n* fonemas con sus respectivas probabilidades (v. g. las características fonotácticas del idioma). Para la creación de los modelos de lenguaje son necesarios grandes corpus de texto (y/o transcripciones ortográficas de grabaciones) para la estimación de dichas probabilidades. Los modelos pueden ser entrenados usando modelos ocultos de Markov (HMM´s), redes neuronales o mezclas gaussianas, entre otros.

Dada la clara diferencia de los lenguajes por sus fonemas y en el uso de éstos, ***los sistemas basados en el reconocimiento fonético*** son los más populares, alcanzando un importante desarrollo. Los fonemas al hablar están dados como una función del tiempo y determinan la base del lenguaje a identificar basado en las estadísticas de la secuencia de fonemas. Por ejemplo, Lamel construyó dos reconocedores de fonemas basados en HMM: uno para el inglés y otro para el francés (Lamel y Gauvain [47]). Dichos reconocedores de fonemas trabajan con datos de prueba en inglés o francés. Lamel encontró que la probabilidad obtenida por los reconocedores de fonemas dependientes del lenguaje podrían ser utilizados en la discriminación entre los dos idiomas: inglés y francés. Muthusamy realizó algo similar para el inglés y japonés [48].

La novedad en los sistemas basados en reconocimiento de fonemas fue la incorporación de más conocimiento dentro del área de identificación del lenguaje hablado. Ambos Lamel y Muthusamy probaron sus sistemas con un corpus multi-lenguaje etiquetado fonéticamente, generando más información para el proceso de entrenamiento,

es decir, un entrenamiento con muestras de señal de voz fonéticamente etiquetadas en cada lenguaje. El único problema es que en estos sistemas es más difícil agregar nuevos lenguajes dentro del proceso de identificación del lenguaje hablado. En comparación con los sistemas que usan similitud-espectral, los cuales no requieren de etiquetado.

Una aportación importante en esta área de investigación fue que los sistema de LID podían ejecutarse satisfactoriamente aun cuando los modelos de reconocedores fonéticos no hayan sido entrenados con las muestras de señal de voz de los lenguajes a reconocer. A este tipo de sistemas se les conoce como PRLM por sus siglas en inglés (Phone Recognition followed by Language Modelling). El entrenamiento de estos sistemas se basa en los fonemas de un sólo lenguaje y se utiliza la información fonotáctica –el conjunto de reglas dependientes del lenguaje las cuales determinan las combinaciones válidas de los fonemas– para la discriminación entre los idiomas.

Por ejemplo, en alemán la palabra "*spiel*" se pronuncia /sh p iy l/, cuya pronunciación similar en inglés podría ser "spelled" (deletrear). Como la palabra "*spiel*" comienza con un conjunto de consonantes /sh p/ , lo cual ocurre muy poco en inglés (sólo en casos cuando una palabra termina en /sh / y la siguiente empieza por /p /, o en una palabra compuesta como "*flashpoints*"), entonces podemos obtener una característica que diferencia a estos dos idiomas en particular. Este tipo de datos fonotácticos fueron utilizados por Damashek [49], el cual usaba un análisis de n-gramas obtenido de documentos en texto para la identificación del lenguaje. Otro ejemplo, es el reportado por Caseiro y Troncoso [5] donde la identificación entre el par de idiomas: Español y Alemán pudo ser realizada usando los modelos acústicos del Portugués. Este trabajo se extendió usando el corpus SpeechDat-M [50], alcanzando los mejores resultados hasta hoy en este tipo de sistemas. Este último trabajo utilizó información de un sólo idioma, el portugués, para construir los reconocedores fonéticos de los otros idiomas. Los modelos de los otros idiomas están basados en una interpolación de las probabilidades de un fonema tipo bi-grama. Sus resultados alcanzaron 79.6% de exactitud usando 10 segundos de señal de voz, para 6 lenguajes a identificar.

Zissman y Singer [51] probaron con el corpus OGI_TS [21] usando una pequeña variación a este enfoque: ellos explotan la posibilidad de que un reconocedor fonético, para

un lenguaje, puede ser desarrollado reutilizando los modelos acústicos de un lenguaje diferente. Esto tiene la ventaja de que los reconocedores fonéticos no necesitan ser desarrollados para todos los lenguajes a identificar. Este sistema [51] obtiene un 79% precisión en la identificación de 11 lenguajes usando 50 segundos de señal de voz y 70% usando 10 segundos de señal de voz. Yan y Barnard [52] realizaron una extensión a este tipo de sistemas conteniendo múltiples lenguajes, en donde necesitaron obtener todos los reconocedores fonéticos para cada lenguaje a ser identificado. La figura 4.3 muestra la estructura de este tipo de sistemas, los cuales son conocidos como PPRLM por sus siglas en inglés (Parallel PRLM).

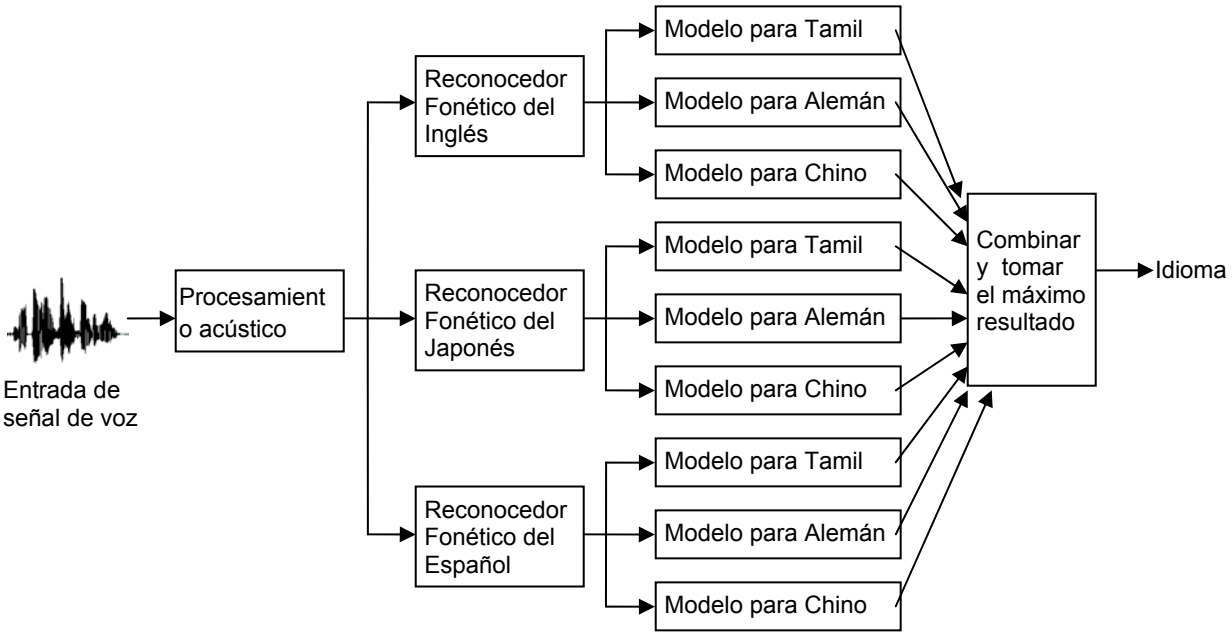

Figura 4.3 Sistema que utiliza diferentes reconocedores de fonemas en paralelo PPRLM

Incluimos en esta categoría a los “tokens” de una señal de voz. El proceso para obtener los “tokens” consiste en convertir una muestra de señal de voz en una secuencia de fonemas simbolizados. Los valores resultantes de la secuencia de fonemas

simbolizados fueron utilizados en la identificación de los lenguajes hablados por Zissman y Singer [51], una versión más actualiza de estos “tokens” fue hecha por Torres-Carrasquillo et al [8], al utilizar los modelos de mezclas gaussianas -GMM Gaussian Mixture Model- con dichos “tokens”, uno para cada lenguaje a identificar; utilizando 12 lenguajes, obtuvo porcentajes de discriminación no mayores al 80%. El uso de GMM en la identificación del lenguaje hablado fue introducido por Zissman en 1993 [45].

Por otro lado, también se han realizado importantes avances en la tarea de verificación de idiomas, principalmente impulsados por el foro de evaluación del NIST. En esta tarea, a diferencia de la tarea de identificación, se desea comprobar si una muestra es o no de un lenguaje específico. Es decir, el sistema de verificación recibe como entrada la muestra y el idioma que se desea verificar y tiene como salida la confirmación o el rechazo de que la muestra es de dicho idioma. El trabajo de Singer et al [10] usando “tokens” y GMM ha sido evaluado en este foro alcanzando una taza de error (EER) de 4.8%. Campbell, Singer, Torres-Carrasquillo y Reynolds en 2004 [53] propusieron el uso de las máquinas de vectores de soporte para la clasificación de las secuencias de “tokens” fusionadas con las GMM mejorando la tasa de error en 3.2%.

Entre los sistemas basados en el reconocimiento de fonemas y los sistemas específicos para el reconocimiento del habla con grandes vocabularios, se encuentran ***los sistemas basados en el reconocimiento de palabras*** [40] para la identificación del lenguaje hablado. Estos sistemas utilizan una secuencia de modelos más sofisticada que los que utilizan los modelos fonotácticos (reconocer los fonemas), pero no emplean la conversión completa de los sistemas de voz a texto. Mendoza et al [55], Kadambe y Hieronymus [54] propusieron el uso de un modelo léxico, basándose en que cada idioma tiene sus propias reglas léxicas; la muestra de voz es procesada por cada uno de los modelos creados para cada lenguaje reconociendo los fonemas, esto se hace para cada modelo de fonemas en paralelo. Resultando de esto una secuencia de fonemas de un lenguaje en particular con la máxima probabilidad. Estos sistemas funcionan en la siguiente secuencia: se reconocen primero los fonemas, después las palabras y eventualmente el idioma.

4.1.1 DISCUSIÓN

La representación o modelado del habla usada en los sistemas de identificación del lenguaje con reconocimiento fonético determina el detalle con el cual las características estructurales son extraídas. Las investigaciones se han concentrado en:

(i) la búsqueda de una "*buena*" representación con respecto a los resultados (una correcta discriminación de los lenguajes).
(ii) una representación previa de los lenguajes no vistos.

La tendencia general en los recientes años ha sido hacia la modelación fonética fina. Dicha representación es altamente dependiente del lenguaje y requiere un etiquetado manual de datos de entrenamiento. Ahora las investigaciones se enfocan al uso de modelos con selección detallada del habla para reducir la complejidad de los sistemas y generalizar los datos etiquetados, como por ejemplo, Yan [4] desarrolló un conjunto de fonemas representativos para cada lenguaje, o Berkling [56] que para conseguir la independencia del lenguaje, desarrolló unos fonemas generales por medio de clusters. Como se mencionó anteriormente, los sistemas con reconocimiento fonético, son los que más se han investigado y los de mejores resultados. Tres elementos son fundamentales en este tipo de sistemas:

- Un módulo para la segmentación de la señal de voz en fonemas. Esta segmentación debe ser realizada con buenos resultados, es decir, que sí se logren capturar los fonemas, ya que el siguiente módulo depende de su correcto funcionamiento.
- Un módulo para la extracción de características estructurales de las palabras. En donde se utiliza el modelado del lenguaje, recordemos que un modelo de lenguaje es un listado de todas las posibles combinaciones de dos, tres o *n* fonemas con sus respectivas probabilidades. Para la creación de los modelos de lenguaje son necesarios grandes corpus de texto (y/o transcripciones ortográficas de grabaciones) para la estimación de dichas probabilidades. Por supuesto, se da por hecho que se trata de un idioma con convenciones claramente establecidas para su

escritura. Situación que no es del todo cierta para todas las lenguas humanas, por ejemplo, muchas de las lenguas indígenas mexicanas.

- Un proceso de etiquetado manual de los datos de entrenamiento para cada lenguaje a identificar. Este etiquetado manual debe ser realizado por expertos a nivel fonético para determinar las fronteras entre los fonemas. Además de que dicha segmentación manual de la señal de voz es una tarea muy costosa, damos por hecho, que el idioma en cuestión ha sido previamente sistematizado por lingüistas definiendo claramente su conjunto de fonemas.

4.2 SISTEMAS SIN RECONOCIMIENTO FONÉTICO

En la figura 4.4 se muestra la estructura de los sistemas que no utilizan representación fonética, en la cual se eliminan los módulos que tienen que ver con el alineamiento de fonemas, y con la extracción de características estructurales, es decir, se eliminan los módulos que concentran la información lingüística del lenguaje a identificar. Por supuesto, la etapa de extracción de características acústicas de la señal de voz es diferente a los sistemas anteriores al aprovechar otras características del lenguaje, tales como la prosodia, el ritmo, la entonación, etc.

Li [57] propuso el uso de ciertas características para la identificación del lenguaje hablado, sin recurrir al uso de la segmentación de fonemas. En su sistema, el núcleo-silábico (por ejemplo las vocales) de cada muestra de señal de voz era localizado automáticamente. Después, cada vector de características conteniendo información espectral era procesado por regiones cercanas al núcleo silábico. Cada uno de esos vectores consistía de sub-vectores espectrales calculados en los marcos vecinos (pero no necesariamente adyacentes) de los datos de la señal de voz. Más que colectar y modelar dichos vectores en un entrenamiento total de la señal de voz, Li guardó por separado las colecciones de los vectores de características para cada una de las muestras de entrenamiento. Durante la prueba, se localiza el núcleo silábico de las muestras de prueba y se obtiene el vector de

características. Cada conjunto de vectores de características de entrenamiento es comparado con el del conjunto de pruebas y se busca al vector de características más similar. Ya encontrado entonces se dice que el idioma del vector de características más similar es hipotéticamente el idioma de la muestra de prueba.

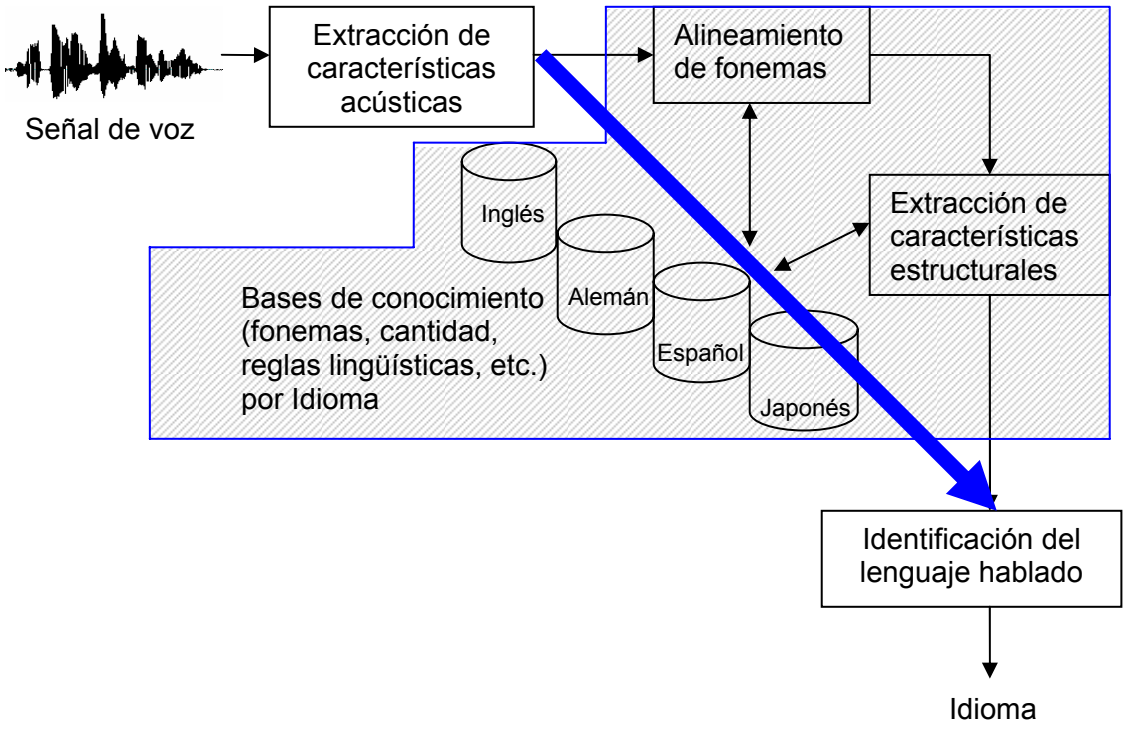

Figura 4.4 Componentes básicos para la identificación del lenguaje sin representación fonética.

La prosodia de los lenguajes, compuesta por la entonación, el ritmo, la duración, etc.; ha sido utilizada en la identificación del lenguaje hablado. Su uso fue motivado en parte, por los estudios hechos por Muthusamy [39] donde mostraba que los humanos utilizamos las características de la prosodia en la identificación de los idiomas. Por ejemplo, Itahashi et al [13][58] construyeron su sistema basado en el uso simple del pitch (frecuencia fundamental

F0). Su argumento es que la estimación del pitch es más robusta en ambientes ruidosos que los parámetros espectrales. Posteriormente, Thyme-Gobbel y Hutchins [14] también utilizaron las características de la prosodia para la identificación del lenguaje hablado. Sus parámetros están basados en la captura de los contornos del pitch y la amplitud entre una sílaba y otra. Dichos parámetros fueron normalizados para ser insensibles a la proporción de la amplitud total, el pitch y el hablante. Sus resultados se muestran en la tabla 4.1 y 4.2 para dos características definidas por ellos: (i) la característica dP (*Diferencial Pitch*), que trata de capturar las diferencias del contorno del picth entre una sílaba y la siguiente; (ii) la característica del ritmo, la cual Thyme-Gobbel la definió: "como la baja frecuencia de la FFT sobre la amplitud" [14]. Este trabajo demostró que era posible utilizar parámetros prosódicos en la discriminación entre una lengua y otra.

	Español	Japonés	Mandarín
Inglés	61	57	63
Español	-	61	65
Japonés	-	-	67

Tabla 4.1 Porcentajes de discriminación obtenido por Thyme-Gobbel et al [14], utilizando la característica "dP".

	Español	Japonés	Mandarín
Inglés	63	58	59
Español	-	59	62
Japonés	-	-	60

Tabla 4.2 Porcentajes de discriminación obtenido por Thyme-Gobbel et al [14], utilizando la característica "ritmo".

Trabajos posteriores a Thyme-Gobbel y Hutchins han alcanzado mejores resultados. A continuación se resumen los tres trabajos con mejores resultados en la

identificación del lenguaje sin reconocimiento fonético y es precisamente con estos trabajos que comparamos los métodos propuestos en esta tesis:

- Cummins et al [15] en 1999 utilizó la prosodia en el procesamiento acústico para extraer las características de la señal de voz de cada uno de los lenguajes a identificar. El trabajo recae en la suposición de que las variaciones de amplitud en la frecuencia fundamental son importantes para percibir el ritmo en el habla. Para ello utilizó el filtro pasa-bandas de bajo orden Butterworth, centrado a 1000 Hz con un ancho de banda de 500 Hz. El logaritmo de la frecuencia fundamental Log F0 fue calculado para cada 1 ms de la señal de voz, el cual fue obtenido con una ventana rectangular. También usó la energía y la derivada de la energía, conocida como delta de la energía (DEnergía). A partir de esta caracterización usó una red neuronal para clasificar pares de idiomas. Para la red neuronal usó el modelo LSTM (Long Short-Term Memory) y para el entrenamiento utilizó una combinación de backpropagation truncada a través del tiempo y aprendizaje recurrente en tiempo real. Sus pruebas fueron hechas con el corpus OGI_TS [21] -ver sección 5.1.1.- tomando cinco idiomas: inglés, japonés, alemán, español y chino mandarín, con 50 hablantes diferentes por idioma para entrenamiento y 20 para prueba. Sus resultados se muestran en la tabla 4.3, para pares de lenguajes en señales de voz de 50 segundos.

	Alemán	Español	Japonés	Mandarín
Inglés	52	62	57	58
Alemán	-	51	58	65
Español	-	-	66	47
Japonés	-	-	-	60

Tabla 4.3 Porcentajes de discriminación obtenido por Cummins et al [15].

- Samouelian [59] para el procesamiento acústico realizó los siguientes pasos: fragmentó la señal, obtuvo 12 coeficientes cepstrales de frecuencia Mel, calculó su delta (el cambio de cada coeficiente entre dos segmentos contiguos en el tiempo), y finalmente aplicó la derivada de la energía, DEnergía. Para el proceso de

aprendizaje usó el algoritmo C4.5 para generar un árbol de decisión a partir de un corpus de entrenamiento de 50 hablantes de tres idiomas (inglés, alemán y japonés), obtenidos del corpus OGI_TS [21] con muestras de señal de 45 segundos. Las pruebas se realizaron con muestras de 45 segundos y de 10 segundos obteniendo exactitudes de 53% y 48.6% respectivamente.

- Rouas en 2003 [22] y 2005 [36] propone un método para identificar los lenguajes en base a su entonación y ritmo, retoma las teorías lingüísticas (vistas en el capítulo 3) y trata de caracterizar el ritmo en función de intervalos vocálicos y consonánticos. Su modelo parte de segmentar la señal de voz en intervalos formados por vocales e intervalos formados por consonantes, para obtener los parámetros del ritmo. Dicho modelo se basa en los estudios previamente hechos por Ramus et al [33]. Hay que recordar que Ramus realizó su experimento con muy pocas muestras y donde lo que se trataba de demostrar era una nueva forma de clasificar los idiomas. Rouas utilizó la frecuencia fundamental F0 de cada segmento, obteniendo cuatro parámetros, para determinar los intervalos vocálicos y consonánticos. Dichos parámetros son modelados usando los modelos de mezclas de Gaussianas (GMM). Sus experimentos fueron hechos con el corpus OGI_TS usando 10 lenguajes. Sus resultados se muestran en la tabla 4.4. para pares de lenguajes en la tarea de identificación del lenguaje hablado [22]; ya que los resultados mostrados en [36] fueron para la tarea de verificación del idioma.

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	60	68	75	68	68	79	77	76
Alemán	_	59	62	66	66	71	70	72
Español	_	_	81	62	63	76	65	67
Mandarín	_	_	_	50	51	74	74	76
Vietnamita	_	_	_	_	69	56	71	67
Japonés	_	_	_	_	_	66	59	67
Coreano	_	_	_	_	_	_	62	75
Tamil	_	_	_	_	_	_	_	70

Tabla 4.4 Porcentajes de discriminación obtenido por Rouas et al [22].

A manera de resumen, las tablas 4.5 y 4.6 muestran un concentrado de los resultados obtenidos por los trabajos más significativos usando los diferentes enfoques: con reconocimiento fonético y sin él.

Autor /año	Tipo de procesamiento acústico	Tiempo de muestra (segundo)	Cantidad de idiomas a verificar/ identificar	Recursos lingüísticos		Método de clasificación	Corpus	Resultados	
				Reconocimiento fonético	Modelos de lenguaje			Exactitud en reconocimiento	Error en verificación (EER)
Casseiro 2000	12 DMFCC	10	6 lenguajes	Sólo uno (para el portugués)	Uno para cada lenguaje a reconocer	Interpolación de probabilidades de los bi-gramas	SpeechDat corpus	79.6%	
Torres/ Singer 2002	SDC Delta-cepstral coeficientes	45	12 lenguajes	tokens en lugar de fonemas	Uno para cada lenguaje a reconocer	GMM-Tokenization	OGI_TS		6.7%

Tabla 4.5 Comparativo de métodos para la identificación del lenguaje hablado con reconocimiento fonético.

Autor /año	Tipo de procesamiento acústico	Tiempo de muestra (segundos)	Cantidad de idiomas a reconocer	Tipo de Clasificación	Método de clasificación	Tamaño de Corpus	Porcentaje de identificación
Samouelian 1998	12 MFCC 12 DMFCC 1 DEnergía	45 seg entrenamiento	3 lenguajes	Multiclase	árbol de decisión C4.5	OGI_TS 50 hablantes c/idioma	45s: 53% 10s: 48.6%
Cummins 1999	Delta F0 DEnergía	50	5 lenguajes	Binaria	Red neuronal back-propagation LSTM	OGI_TS 50 hablantes c/idioma	*Ver tabla 4.3*
Rouas 2003 y 2005	Intervalos de vocales y consonantes	45 seg 10 seg	10 lenguajes	Binaria	GMM-modelo prosódico	OGI_TS	*Ver tabla 4.4*

Tabla 4.6 Comparativo de métodos para la identificación del lenguaje hablado sin reconocimiento fonético.

De la tabla 4.5 podemos ver que los métodos que utilizan representación fonética hasta hoy son lo que mejores resultados han dado. El trabajo de Torres-Carrasquillo, evaluado en el NIST, en la tarea de verificación del lenguaje hablado, muestra sus resultados en forma diferente que los de Casseiro. Hay que recordar que la tarea de verificación consiste en dar una muestra del idioma, por ejemplo el francés, y el sistema tiene que dar la repuesta si es francés o no lo es. Por otro lado, la tarea de identificación consiste en determinar el idioma de una muestra, de entre un conjunto de idiomas conocidos.

En el caso de los sistemas sin representación fonética, Samouelian probó con pocos idiomas pero su evaluación fue multiclase, como los realizan los sistemas con representación fonética. Y los casos de Cummins y Rouas sus evaluaciones fueron con clasificadores binarios. Se puede construir clasificadores multiclase a partir de los binarios por medio de métodos de combinación de clasificadores, se selecciona el resultado del clasificador binario que sea más alto que los otros resultados .

CAPÍTULO 5

INCLUSIÓN DE INFORMACIÓN SUPRASEGMENTAL

De acuerdo a los lingüistas, existe información suprasegmental en el habla, tales como la prosodia, el ritmo, la entonación y la duración; los cuales son elementos distintivos de los idiomas. Dichas características, llamadas suprasegmentales, afectan conjuntamente a varios fonemas, es decir, no se pueden analizar cuando separamos el habla en fonemas. Por lo tanto, uno de los objetivos es extraer las características suprasegmentales del habla para la tarea de identificación automática del lenguaje hablado.

Como se mencionó en el estado del arte, Cummins et al [15] basan el procesamiento acústico solamente en el uso de la frecuencia fundamental. En este trabajo se planea ampliar las características al usar las frecuencias secundarias. Por esta razón se plantea el uso de coeficientes cepstrales de frecuencia Mel, de manera similar a la utilizada por Samouelian [59]. Sin embargo, proponemos capturar los cambios temporales en el espectro de la señal de voz por medio de los cambios (deltas Δ) en los coeficientes cepstrales de frecuencia Mel –que a diferencia de Samouelian– los cambios no son calculados únicamente en ventanas adyacentes. Por otro lado, la utilidad de los coeficientes cepstrales de frecuencia Mel ha sido demostrada tanto en la identificación del lenguaje hablado con representación fonética, como en la identificación sin reconocimiento fonético. Bajo este último enfoque, se han utilizado 12 coeficientes cepstrales para el

procesamiento acústico, la cantidad comúnmente usada para el tratamiento de voz [17]. Nosotros además de pretender capturar la información suprasegmental con el uso de los deltas de los coeficientes, hemos extendido el número de coeficientes cepstrales a 16. El objetivo es obtener más detalle de las frecuencias que componen la señal de voz, ya que se desea ir más allá de una segmentación fonética.

Antes de detallar el método propuesto que se encuentra en las secciones 5.2 y 5.3 es necesario explicar las condiciones bajo las cuales se realizó la experimentación y las herramientas utilizadas para tales propósitos, las cuales se encuentran en la siguiente sección.

5.1 PROTOCOLO DE EXPERIMENTACIÓN

Esta sección describe: el corpus utilizado, la cantidad de hablantes de cada idioma, los idiomas a identificar, los tamaños de muestra. Así como los clasificadores utilizados en el proceso de aprendizaje.

5.1.1 CORPUS OGI_TS

Para tener la misma base de evaluación se utilizo el corpus: The Oregon Graduate Institute Multi-language Telephone Speech (OGI_TS por sus siglas en inglés) [21]. El cual se creó formulando una serie de preguntas a un interlocutor. Las grabaciones se realizaron por teléfono, es decir a 8Khz, y divididas por cada una de las 10 preguntas formuladas. Cuatro de las preguntas tienen respuestas cerradas, por ejemplo: “Por favor recite los siete días de la semana”, “por favor diga los números del cero al 10”. Y las restantes seis tienen

respuestas abiertas, es decir no hay texto predeterminado como respuesta, por ejemplo: “Describe el cuarto donde estás en este momento”, “Describe el trayecto a tu casa”, “Habla de cualquier tema que desees”. Las diez respuestas contienen mensajes diferentes con un total aproximado de dos minutos de voz por cada persona. El número de personas varía de 70 a 100 dependiendo del idioma. El corpus comprende un total de 22 idiomas.

De este corpus se tomaron nueve idiomas: Inglés, Alemán, Español, Japonés, Chino Mandarín, Coreano, Tamil, Vietnamita y Farsi. Los mismos usados en los trabajos de Cummins et al [15] y Rouas et al [22] con el fin de tener puntos de comparación de nuestros métodos. No fue posible realizar la comparación con el Francés dado que recientemente fue eliminado del Corpus OGI_TS.

En los experimentos se tomaron 50 hablantes diferentes para cada idioma. Recordemos que las grabaciones son de conversaciones telefónicas, donde las personas hablaron espontáneamente, respondiendo a preguntas. Al tratarse de habla espontánea se presentan fenómenos de co-articulación así como pausas. Las pruebas fueron hechas con diferentes tamaños de muestras de 7, 10, 30 y 50 segundos, tanto para entrenamiento como para prueba. Teniendo en total 450 hablantes diferentes y 22500 segundos de señal de voz, para el experimento de 50 segundos por hablante.

Las pruebas con 50 segundos fueron realizadas sobre las muestras de voz con las grabaciones de las respuestas a la pregunta: “Habla sobre cualquier tema que desees”, tomando los primeros 50 segundos de muestra de voz. El OGI_TS se refiere a estas muestras como “historias después del tono”, y son denotadas por el indicador *story-bt*, donde *bt* es el número de la persona que realizo dicha grabación.

Las pruebas con 30 segundos, son el resultado de cortar a 30 segundos las mismas muestras de voz de las historias; de la misma manera se tienen muestras de 10 y 7 segundos.

5.1.2 ALGORITMOS DE APRENDIZAJE

Existen dos tipos de clasificadores: los supervisados y los no supervisados. En el contexto de aprendizaje automático entendemos por clasificación supervisada, la clasificación sabiendo la existencia de ciertas clases y teniendo ejemplos de cada una de ellas. En nuestro caso utilizamos la clasificación supervisada, porque tenemos ejemplos de la señal de voz de cada uno de los idiomas a reconocer.

Clasificar consiste en establecer una regla para ubicar nuevas observaciones en alguna de las clases existentes. Un constructor de clasificadores produce una regla de clasificación a partir de datos previamente clasificados, esta regla provee al sistema de capacidad de predicción, es decir después de entrenar al sistema con los datos de las características de la señal de voz, el sistema tienen la capacidad de predecir el idioma correspondiente a los datos de prueba.

Para demostrar la pertinencia de nuestros métodos se realizaron experimentos con diferentes clasificadores, de acuerdo a diferentes representaciones: estocásticos, máquinas de vectores de soporte y árboles de decisión. Los clasificadores utilizados fueron:

- Vecinos más cercanos (KNN): NNge (nearest-neighbor general) instrumentado en la herramienta WEKA [61].
- El clasificador Naïve-Bayes
- Máquinas de vectores de soporte (SVM): SMO (Sequential Minimal Optimization) que es una máquina de vectores de soporte instrumentado por Platt [60] en la herramienta WEKA [61]).
- Árboles de decisión: C4.5

Para la extracción de la información acústica de la señal de voz se utilizó Praat [62] un editor de audio y extractor de características acústicas de la señal de voz. Finalmente las pruebas de clasificación se realizaron con el sistema Weka [61].

5.1.3 EVALUACIÓN

Para validar el comportamiento de un modelo de aprendizaje se tienen que usar datos independientes de los usados para construir el modelo. Es decir, necesitamos datos de entrenamiento y datos de prueba. El clasificador predice la clase de cada dato de prueba: si es correcta se cuenta como éxito, en caso contrario como error.

Del conjunto de datos se toma una parte para los datos de prueba y el resto para el entrenamiento. Mientras mayor es el conjunto de entrenamiento tenemos un mejor clasificador y mientras sea mayor el conjunto de pruebas más exacto es el estimado de error [63]. En el caso de que la cantidad de datos para entrenamiento y prueba sean limitados, como en nuestro caso que tenemos 50 hablantes por cada idioma sin contar con las variantes y regionalismos que tiene cada idioma, existen varias soluciones. La más sencilla es reservar una parte de los datos para prueba, por ejemplo el 80% para entrenamiento y 20% para pruebas. Pero pudiera ser que ambos conjuntos no sean representativos. Además, cada clase en todo el conjunto de datos debería estar representada, en la misma proporción para los datos de entrenamiento y para los datos de prueba.

La validación cruzada de k pliegues (*k-fold cross validation*) [63] es el método de evaluación más utilizado en el aprendizaje automático. En este método, los datos disponibles se dividen aleatoriamente en un conjunto de entrenamiento y un conjunto de prueba, de la siguiente manera: se divide el conjunto de datos de que se dispone en *k* conjuntos disjuntos de igual tamaño T_1,..., T_k . Se realizan *k* experimentos, usando como conjunto de entrenamiento en la iteración i-ésima $\bigcup_{j \neq i} T_j$ y como conjunto de prueba T_i . Los resultados de cada uno de los k experimentos son promediados para obtener una sola estimación.

Las pruebas realizadas en nuestros experimentos fueron hechas con el método de validación cruzada con 10-pliegues. Es decir, los datos se dividen en 10 partes, cada clase

(idioma) está representada en la misma proporción, en cada iteración. Recordemos que son 10 iteraciones en total, se retienen 1/10 de los datos diferentes para prueba y 9/10 se utilizan para entrenamiento, se calcula el error que genera cada iteración y se promedian los resultados.

Para nuestra tarea, la identificación entre pares de idiomas (problema de dos clases), se tienen dos clases idioma1 e idioma2 con cuatro posibles resultados [63].

- Positivos verdaderos (clasificación correcta).
- Falsos positivos (el resultado es idioma1 cuando debería ser idioma2).
- Negativos verdaderos (clasificación correcta).
- Falsos negativos (el resultado es idioma2 cuando debería ser idioma1).

	Idioma1	Idioma2
Idioma1	Positivo verdadero	Falso negativo
Idioma2	Falso positivo	Negativo verdadero

Tabla 5.1 Matriz de confusión para el problema de dos clases.

5.1.4 REDUCCIÓN DE DIMENSIONALIDAD

Nuestros conjuntos de características de representación del habla para cada uno de los hablantes de cada idioma no son pequeños. Por ejemplo para el caso de muestras de señal de voz de 50 seg, tenemos alrededor de 20,000 atributos por cada muestra de señal de voz (más adelante se detalla la extracción de características). Así pues, debido a los problemas causados por la alta dimensionalidad de la representación del conjunto de

características de la señal de voz, existe la necesidad de reducir el conjunto de características original, es decir, hacer una reducción de dimensionalidad. En nuestro trabajo utilizamos la ganancia de información (IG *Information Gain* por sus siglas en inglés).

Una definición de ganancia de información es la diferencia entre la entropía de un atributo y la de clase, causada por dividir los ejemplos de acuerdo a dicho atributo. En el fondo no es más que una heurística, que nos servirá para la elección de lo mejores atributos en cada clase, de acuerdo a un umbral.

De manera general, se define la entropía como la medida de la incertidumbre que hay en un sistema. En otras palabras, la probabilidad de que ocurra cada uno de los posibles resultados, ante una determinada situación.

De acuerdo a Tom Mitchell [63] en forma general: Dada una colección de datos S, donde el atributo objetivo puede tomar *c* diferentes valores (*c* clases), entonces la entropía de S relativa a dicha clasificación *c* es definida de la siguiente manera:

$$Entropía(S) \equiv \sum_{i=1}^{c} - p_i \log_2 p_i$$

Donde p_i es la proporción de S perteneciente a la clase *i*.

Entonces la ganancia de información de un atributo A relativo a una colección de datos S, se define de la siguiente manera:

$$Ganancia\ (S, A) \equiv Entropía\ (S) - \sum_{\upsilon \in Valores\ (A)} \frac{|S_\upsilon|}{|S|} Entropía\ (S_\upsilon)$$

Donde *Valores(A)* es el conjunto de todos los posibles valores para el atributo A y $S\upsilon$ es el subconjunto de S para el cual el atributo A tiene el valor υ.

5.2 MÉTODO DE REFERENCIA

Como primer paso para la extracción de características suprasegmentales, proponemos ampliar el número de coeficientes cepstrales de frecuencia Mel (MFCC) que los usualmente se utilizan para el reconocimiento del habla, es decir a más de 12. Con el objetivo de tener más información para la tarea de discriminación, fijamos el número a 16 coeficientes. Aplicamos una ventana de 20 milisegundos, muy utilizada por el estado del arte para la segmentación de fonemas. Sin traslapes entre ventanas. Se tomaron muestras de 7, 30 y 50 segundos de señal de voz; 50 hablantes para cada idioma a identificar, teniendo 9 idiomas.

Este proceso se automatizó por medio de un programa hecho en la herramienta PRAAT [62], que incluye la fragmentación de la muestra y la obtención de las características acústicas de la señal de voz aplicando a cada una de esas muestras el modelado de señales de análisis local en la frecuencia por medio del método MFCC. Finalmente el programa obtiene una matriz y una tabla de valores reales de las características acústicas de la señal de voz, para su posterior proceso.

Sin utilizar ningún método de reducción de dimensionalidad se probó para 7 segundos de señal de voz. El clasificador utilizado fue el de Naïve Bayes usando clasificación con validación cruzada de 10 pliegues, vista en la sección 5.1.3. Los resultados fueron buenos, superando en algunos casos al estado del arte, en específico a Cummins et al [15] y Rouas et al [22].

	Alemán	Español	Japonés	Mandarín
Inglés	**90** (52)	**94** (62)	**92** (57)	**60** (58)
Alemán	-	**60** (51)	**67** (58)	**73** (65)
Español	-	-	66 (66)	**73** (47)
Japonés	-	-	-	**74** (60)

Tabla 5.2 Porcentajes de discriminación con muestras de señal de voz de 7 segundos, entre paréntesis el porcentaje obtenido por Cummins et al [15].

En la tabla 5.2 se muestran los resultados contra Cummins y en la tabla 5.3 se muestran los resultados contra Rouas. Hay que notar que tanto Cummins como Rouas utilizaron tamaños de muestras de señal de voz más grandes, lo cual representa tener más información para caracterizar la señal de voz.

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**90** (60)	**94** (68)	60 (75)	**70** (68)	**92** (68)	70 (79)	**94 (**77)	**90** (76)
Alemán	_	**60** (59)	**73** (62)	65 (66)	**67** (66)	63 (71)	56 (70)	61 (72)
Español	-	-	73 (81)	**75** (62)	**66** (63)	67 (76)	**69** (65)	65 (67)
Mandarín	-	-	-	**66** (50)	**74** (51)	54 (74)	**78** (74)	72 (76)
Vietnamita	-	-	-	-	65 (69)	**64** (56)	69 (71)	63 (67)
Japonés	-	-	-	-	-	**67** (66)	**66** (59)	63 (67)
Coreano	-	-	-	-	-	-	**65** (62)	71 (75)
Tamil	-	-	-	-	-	-	-	67 (70)

Tabla 5.3 Porcentajes de discriminación con muestras de señal de voz de 7 segundos, entre paréntesis el porcentaje obtenido por Rouas et al [22].

5.2.1 REDUCIENDO DIMENSIONALIDAD

Uno de los principales problemas es el tamaño de los vectores de características y este aumenta cuando utilizamos muestras de señal de voz más grandes, sobretodo con muestras de 30 y 50 segundos. Se utilizó un método de reducción de dimensionalidad: ganancia de información (vista en la sección 5.1.4). El umbral fue de 0.0, y se aplicó la ganancia de información antes de realizar la clasificación con validación cruzada de 10 pliegues. El clasificador utilizado fue el de Naïve Bayes. La cantidad de atributos por cada par de lenguajes está entre los 50 y 200 atributos. Los resultados mejoraron como se muestran en las tablas 5.4 y 5.5.

	Alemán	Español	Japonés	Mandarín
Inglés	**93** (52)	**94** (62)	**94** (57)	**89** (58)
Alemán	-	**91** (51)	**89** (58)	**82** (65)
Español	-	-	**91** (66)	**84** (47)
Japonés	-	-	-	**81** (60)

Tabla 5.4 Porcentajes de discriminación con muestras de señal de voz de 7 segundos, utilizando ganancia de información; entre paréntesis el porcentaje obtenido por Cummins et al [15].

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**93** (60)	**94** (69)	**89** (75)	**87** (68)	**94** (68)	**91** (79)	**97** (77)	**92** (76)
Alemán		**91** (59)	**82** (62)	**84** (66)	**89** (66)	**74** (71)	**87** (70)	**89** (72)
Español			**84** (81)	**83** (62)	**91** (63)	**77** (76)	**92** (65)	**92** (67)
Mandarín				**80** (50)	**81** (51)	**89** (74)	**85** (74)	**82** (76)
Vietnamita					**86** (69)	**79** (56)	**77** (71)	**92** (67)
Japonés					-	**77** (66)	**95** (59)	**90** (67)
Coreano					-	-	**76** (62)	**79** (75)
Tamil					-	-		**89** (70)

Tabla 5.5 Porcentajes de discriminación con muestras de señal de voz de 7 segundos, utilizando ganancia de información; entre paréntesis el porcentaje obtenido por Rouas et al [22].

En la tabla 5.6 se muestra los coeficientes utilizados como atributos para la clasificación después de aplicar la ganancia de información. Estos atributos son los más relevantes en la clasificación. Por espacio, solo se muestran los coeficientes después del 12, ya que nuestra propuesta consistió en utilizar más coeficientes cepstrales que los comúnmente utilizados en el estado del arte. La tabla 5.6 nos indica que si son relevantes los coeficientes 13, 14,15 y 16. De los 36 clasificadores binarios, en 27 se utilizó el coeficiente 13 y 15, en 19 el coeficiente 16 y el coeficiente 14 se utilizó en 31. Esto nos indica que el coeficiente más relevante fue el 14.

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	cf13 cf14 cf15 cf16	cf13 cf14 cf15 cf16	cf13 cf15 cf16	cf14 cf15 cf16	cf13 cf14 cf15	cf13 cf15 cf16	cf13 cf14 cf15 cf16	cf13 cf14 cf15 cf16
Alemán	---	cf13 cf14 cf15 cf16	cf13 cf14 cf15	cf13 cf14 cf15	cf13 cf14 cf15 cf16	cf14 cf16	cf13 cf14 cf15	cf14 cf15 cf16
Español	---	---	cf13 cf14 cf15	cf13 cf14 cf16	cf13 cf14 cf15	cf14 cf15	cf14 cf15 cf16	cf14
Mandarín	---	---	---	cf13 cf14 cf16	cf13 cf14	cf13 cf16	cf13 cf14 cf16	cf14 cf15
Vietnamita	---	---	---	---	cf13	cf13 cf14 cf15	cf14 cf15 cf16	cf13 cf14 cf15 cf16
Japonés	---	---	---	---	---	cf13 cf14 cf15	cf13 cf15 cf16	cf13 cf14 cf15
Coreano	---	---	---	---	---	---	cf13 cf14	cf14 cf15
Tamil	---	---	---	---	---	---	---	cf13 cf14 cf15

Tabla 5.6 Atributos después de aplicar ganancia de información a muestras de señal de voz de 7seg.

5.2.2 DISCUSIÓN

En general, con estos experimentos nos dimos cuenta que al aumentar el número de coeficientes a 16, la discriminación de los pares de lenguajes mejoró y en ocasiones superó el estado del arte. Con estos resultados podemos confirmar que el uso de más coeficientes es de importancia en la identificación del lenguaje hablado. Ahora el siguiente paso es la inclusión de otras características que capturen los fenómenos suprasegmentales. Por ello, es que decidimos experimentar con un nuevo conjunto de características acústicas resumiendo la información que tenemos en los coeficientes a través de sus variaciones. Dichas variaciones son capturadas por medio de los cambios (deltas) de los coeficientes a diferentes intervalos, lo cual se describe en la siguiente sección.

5.3 CARACTERIZANDO LOS CAMBIOS DE LA SEÑAL

El objetivo de este experimento fue proponer una nueva caracterización que capturara información suprasegmental. Esta nueva caracterización también redujo el conjunto de características acústicas haciéndolo más manejable, al ser más pequeño. Sabemos que los cambios temporales en el espectro juegan un papel muy importante en la percepción humana. Una de las maneras de capturar esta información es el uso de los cambios o deltas Δ en los coeficientes cepstrales. Con ellos es posible describir el cambio de cada coeficiente en el tiempo, capturando así, los cambios de la señal de voz, cambios posiblemente entre los fonemas, las sílabas y quizás hasta palabras, buscando de esta forma capturar parte de la información suprasegmental (la prosodia, el ritmo, la entonación, la duración, etc.). Por lo tanto, con este método pretendemos capturar las diferencias suprasegmentales que hay en el habla y que son propios de los idiomas. Inclusive en algunos idiomas son base distintiva entre el significado de una palabra u otra; como en las lenguas tonales, por ejemplo el chino que usa la entonación como marcador léxico.

El proceso de extracción de características está basado en el proceso descrito en la sección anterior, es decir, se obtienen 16 coeficientes cepstrales de frecuencia Mel (MFCC) para cada ventana de 20 milisegundos en las muestras de señal de voz digitalizada, obteniendo como resultado una matriz de características. Posteriormente a cada uno de los 16 coeficientes se les obtiene sus deltas Δ. Los deltas se definen de la siguiente manera:

- Tenemos 16 coeficientes en cada ventana, es decir, *i=1* hasta 16.
- Tenemos *N* ventanas de 20 milisegundos en toda la muestra de señal de voz digitalizada. Recordemos que *N* depende del tamaño de muestra de señal de voz, que para nuestros experimentos fue de 7, 30 y 50 segundos.
- Definimos delta 1 : $\Delta_1 c_i = c_{ik} - c_{i(k-1)}$, es decir, la diferencia entre el coeficiente *i* de la ventana *k* menos el coeficiente *i* de la ventana *k-1*.

- Para el delta 2 : $\Delta_2 c_i = c_{ik} - c_{i(k-2)}$, es decir, la diferencia entre el coeficiente *i* de la ventana *k* menos el coeficiente *i* de la ventana *k-2*.
- Para el delta 3 : $\Delta_3 c_i = c_{ik} - c_{i(k-3)}$, es decir, la diferencia entre el coeficiente *i* de la ventana *k* menos el coeficiente *i* de la ventana *k-3*.

En la figura 5.1 se muestra este proceso, note que el oscilograma es sólo para dar una idea de cómo es la extracción de estas características. El tamaño de la ventana es de 20 ms, la cual no está representada en forma real en esta figura.

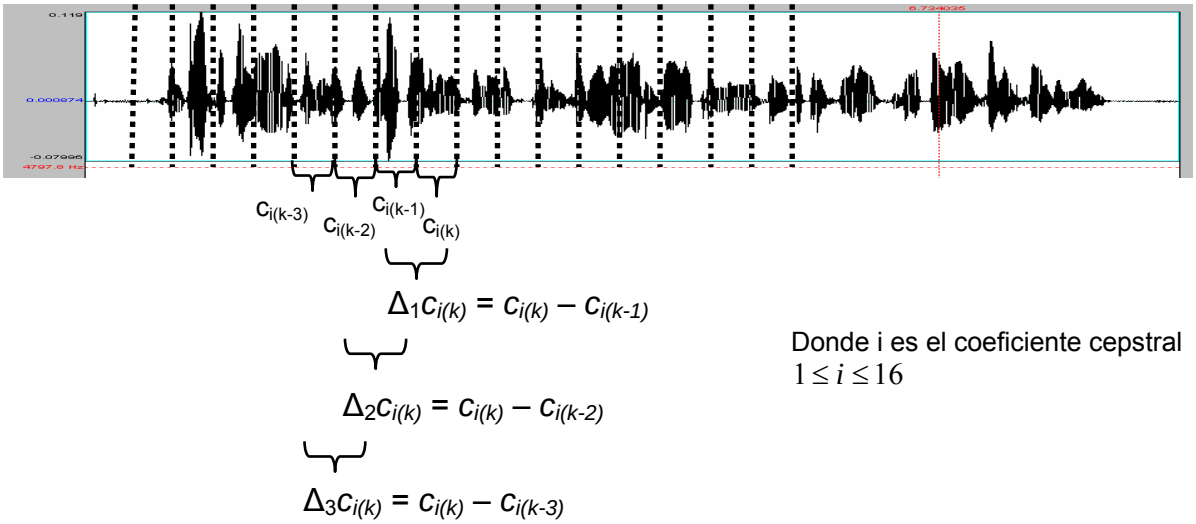

Figura 5.1 Oscilograma segmentado en ventanas de 20ms para la representación de la obtención de los deltas.

Como podemos observar, tenemos algo similar al método de referencia. Pero agregamos el cálculo de los deltas así como también sus promedios, máximos y mínimos, con lo que al final tenemos 192 atributos para cada muestra de señal de voz. En la tabla 5.6 se muestran las formulas utilizadas para la extracción de características.

Descripción	Formulas	Cantidad de atributos
Promedio de los coeficientes	$\widetilde{c}_i = \frac{1}{N}\sum_{k=1}^{N} c_{ik}$	16
Máximo valor de los coeficientes	$\hat{c}_i = \max_{k=1}^{N}\left(c_{ik}\right)$	16
Mínimo valor de los coeficientes	$\breve{c}_i = \min_{k=1}^{N}\left(c_{ik}\right)$	16
Promedio de los cambios en los coeficientes	$\widetilde{\Delta}_1 c_i = \frac{1}{N-1}\sum_{k=2}^{N} c_{ik} - c_{i(k-1)}$ $\widetilde{\Delta}_2 c_i = \frac{1}{N-2}\sum_{k=3}^{N} c_{ik} - c_{i(k-2)}$ $\widetilde{\Delta}_3 c_i = \frac{1}{N-3}\sum_{k=4}^{N} c_{ik} - c_{i(k-3)}$	48
Máximo valor de los cambios de los coeficientes	$\hat{\Delta}_1 c_i = \max_{k=2}^{N}\left(c_{ik} - c_{i(k-1)}\right)$ $\hat{\Delta}_2 c_i = \max_{k=3}^{N}\left(c_{ik} - c_{i(k-2)}\right)$ $\hat{\Delta}_3 c_i = \max_{k=4}^{N}\left(c_{ik} - c_{i(k-3)}\right)$	48
Mínimo valor de los cambios de los coeficientes	$\breve{\Delta}_1 c_i = \min_{k=2}^{N}\left(c_{ik} - c_{i(k-1)}\right)$ $\breve{\Delta}_2 c_i = \min_{k=3}^{N}\left(c_{ik} - c_{i(k-2)}\right)$ $\breve{\Delta}_3 c_i = \min_{k=4}^{N}\left(c_{ik} - c_{i(k-3)}\right)$	48

Tabla 5.7 Tabla de extracción de características independientes del tiempo.

Con estos deltas obtenemos un nuevo vector de características acústicas perceptivas. A partir de estos vectores calculamos los promedios, así como sus máximos y mínimos alcanzando una nueva caracterización que resume en un sólo vector las variaciones de toda una grabación.

En resumen este nuevo vector queda conformado de la siguiente manera para cada uno de los coeficientes cepstrales:

$$X_i = \begin{pmatrix} \text{Promedio-}c_i \\ \text{Máximo-}c_i \\ \text{Mínimo-}c_i \\ \text{Promedio } \Delta_1 c_i \\ \text{Promedio } \Delta_2 c_i \\ \text{Promedio } \Delta_3 c_i \\ \text{Máximo } \Delta_1 c_i \\ \text{Máximo } \Delta_2 c_i \\ \text{Máximo } \Delta_3 c_i \\ \text{Mínimo } \Delta_1 c_i \\ \text{Mínimo } \Delta_2 c_i \\ \text{Mínimo } \Delta_3 c_i \end{pmatrix} \qquad 1 \le i \le 16 .$$

Con esta nueva caracterización se reducen el número de atributos únicamente a 192. En el método de referencia se tenían un conjunto de 5584 características para muestras de 10 segundos, de 23984 para muestras de 30 segundos y de 39984 para muestras de 50 segundos. Los resultados alcanzados con esta nueva caracterización se muestran en la siguiente sección.

5.3.1 RESULTADOS

Sin utilizar ningún método de reducción de dimensionalidad, sólo con base en el nuevo conjunto de características de 192 atributos, se obtuvieron, en algunos casos, mejores resultados que los obtenidos por Cummins et al [15] y Rouas et al [22]; los resultados comparándonos contra Cummins se muestran en la tabla 5.7 para muestras de 7 segundos, en la tabla 5.8 para muestras de 30 segundos y en la tabla 5.9 para muestras de 50 segundos. Se realizaron pruebas con los cuatro clasificadores indicados en la

sección 5.1.2, por efecto de espacio solo se muestran los resultados con el clasificador Naïve Bayes, pero posteriormente se muestra una grafica comparativa de los resultados con los cuatro clasificadores en la sección 5.3.3.

	Alemán	Español	Japonés	Mandarín
Inglés	**86** (52)	**78** (62)	**68** (57)	**59** (58)
Alemán	-	**58** (51)	**72** (58)	**71** (65)
Español	-	-	62 (66)	**72** (47)
Japonés	-	-	-	**66** (60)

Tabla 5.8 Porcentajes de discriminación con muestras de señal de voz de 7 segundos, sin utilizar ganancia de información, entre paréntesis el porcentaje obtenido por Cummins et al [15].

	Alemán	Español	Japonés	Mandarín
Inglés	**61** (52)	**73** (62)	**62** (57)	**70** (58)
Alemán	-	**54** (51)	**61** (58)	58 (65)
Español	-	-	57 (66)	**74** (47)
Japonés	-	-	-	**68** (60)

Tabla 5.9 Porcentajes de discriminación con muestras de señal de voz de 30 segundos, sin utilizar ganancia de información, entre paréntesis el porcentaje obtenido por Cummins et al [15].

	Alemán	Español	Japonés	Mandarín
Inglés	**71** (52)	**88** (62)	**74** (57)	**62** (58)
Alemán	-	**52** (51)	**64** (58)	**69** (65)
Español	-	-	63 (66)	**71** (47)
Japonés	-	-	-	**68** (60)

Tabla 5.10 Porcentajes de discriminación con muestras de señal de voz de 50 segundos, sin utilizar ganancia de información, entre paréntesis el porcentaje obtenido por Cummins et al [15].

Como podemos observar, no en todos los casos este nuevo conjunto de características acústicas obtiene mejores resultados que Cummins, nuestra única dificultad es la pareja de español y japonés, pero definitivamente la pareja inglés y alemán se discriminaron con un muy buen porcentaje. También podemos observar que la muestra de

señal de voz más pequeña, en este caso de 7 segundos, es la que obtuvo mejores resultados. Lo que nos indica que las variaciones capturadas por los deltas Δ_1, Δ_2 y Δ_3 tiendan a equilibrarse con muestras de señal de voz grandes, y al parecer, a acercarse entre los diferentes idiomas. Una posible explicación es la mayor presencia de silencios y pausas mientras más grande sea la muestra de señal de voz.

Los resultados comparándonos contra Rouas et al [22] se muestran en las tablas 5.10 para muestras de 7 segundos, tabla 5.11 para muestras de 30 segundos y tabla 5.12 para muestras de 50 segundos; donde observamos que para el inglés con los demás idiomas funcionó mejor que Rouas, lo que no se cumple para todos los casos.

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**86** (60)	**78** (68)	59 (75)	**72** (68)	**68** (68)	69 (79)	**83** (77)	**81** (76)
Alemán	-	58 (59)	**71** (62)	**68** (66)	**72** (66)	64 (71)	**71** (70)	67 (72)
Español	-	-	72 (81)	53 (62)	62 (63)	58 (76)	52 (65)	64 (67)
Mandarín	-	-	-	**66** (50)	**66** (51)	60 (73)	**78** (74)	66 (76)
Vietnamita	-	-	-	-	61 (69)	53 (56)	66 (71)	58 (67)
Japonés	-	-	-	-	-	62 (66)	**62** (59)	63 (67)
Coreano	-	-	-	-	-	-	**66** (62)	65 (75)
Tamil	-	-	-	-	-	-	-	57 (70)

Tabla 5.11 Porcentajes de discriminación con muestras de señal de voz de 7 segundos, sin utilizar ganancia de información, entre paréntesis el porcentaje obtenido por Rouas [22].

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**61** (60)	**73** (68)	70 (75)	**73** (68)	62 (68)	59 (79)	**83** (77)	66 (76)
Alemán	-	54 (59)	58 (62)	**68** (66)	61 (66)	50 (71)	65 (70)	55 (72)
Español	-	-	74 (81)	**66** (62)	57 (63)	59 (76)	54 (65)	55 (67)
Mandarín	-	-	-	**68** (50)	**68** (50)	56 (74)	**74** (74)	64 (76)
Vietnamita	-	-	-	-	63 (69)	60 (56)	67 (71)	**66** (67)
Japonés	-	-	-	-	-	**66** (66)	**73** (59)	63 (67)
Coreano	-	-	-	-	-	-	**64** (62)	63 (75)
Tamil	-	-	-	-	-	-	-	72 (70)

Tabla 5.12 Porcentajes de discriminación con muestras de señal de voz de 30 segundos, sin utilizar ganancia de información, entre paréntesis el porcentaje obtenido por Rouas [22].

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**71** (60)	**88** (68)	62 (75)	**82** (68)	**74** (68)	75 (79)	**91** (77)	**77** (76)
Alemán	-	52 (59)	**69** (62)	**68** (66)	64 (66)	57 (71)	**73** (70)	60 (72)
Español	-	-	71 (81)	**63** (62)	**63** (63)	61 (76)	63 (65)	60 (67)
Mandarín	-	-	-	**57** (50)	**68** (51)	57 (74)	**77** (74)	**79** (76)
Vietnamita	-	-	-	-	57 (69)	**58** (56)	66 (71)	**75** (67)
Japonés	-	-	-	-	-	65 (66)	**61** (59)	65 (67)
Coreano	-	-	-	-	-	-	**70** (62)	69 (75)
Tamil	-	-	-	-	-	-	-	69 (70)

Tabla 5.13 Porcentajes de discriminación con muestras de señal de voz de 50 segundos, sin utilizar ganancia de información, entre paréntesis el porcentaje obtenido por Rouas [22].

Utilizando ganancia de información sobre el nuevo conjunto de características de 192 atributos independientes del tiempo, se obtuvieron mejores resultados que sin el uso de esta técnica de reducción de dimensionalidad. En este caso los resultados son mejores que los obtenidos por Cummins et al [15]. En las tablas 5.13, 5.14 y 5.15, se muestran los resultados contra Cummins y una gráfica a manera de resumen se muestra en la figura 5.2.

	Alemán	Español	Japonés	Mandarín
Inglés	**85** (52)	**83** (62)	**79** (57)	**67** (58)
Alemán	-	**71** (51)	**69** (58)	**83** (65)
Español	-	-	**70** (66)	**83** (47)
Japonés	-	-	-	**73** (60)

Tabla 5.14 Porcentajes de discriminación con muestras de señal de voz de 7 segundos, utilizando ganancia de información, entre paréntesis el porcentaje obtenido por Cummins[15].

	Alemán	Español	Japonés	Mandarín
Inglés	**79** (52)	**85** (62)	**72** (57)	**68** (58)
Alemán	-	**69** (51)	**62** (58)	**71** (65)
Español	-	-	**75** (66)	**81** (47)
Japonés	-	-	-	**80** (60)

Tabla 5.15 Porcentajes de discriminación con muestras de señal de voz de 30 segundos, utilizando ganancia de información, entre paréntesis el porcentaje obtenido por Cummins [15].

	Alemán	Español	Japonés	Mandarín
Inglés	**78** (52)	**85** (62)	**77** (57)	**75** (58)
Alemán	-	**71** (51)	**66** (58)	**79** (65)
Español	-	-	**69** (66)	**74** (47)
Japonés	-	-	-	**77** (60)

Tabla 5.16 Porcentajes de discriminación con muestras de señal de voz de 50 segundos, utilizando ganancia de información, entre paréntesis el porcentaje obtenido por Cummins [15].

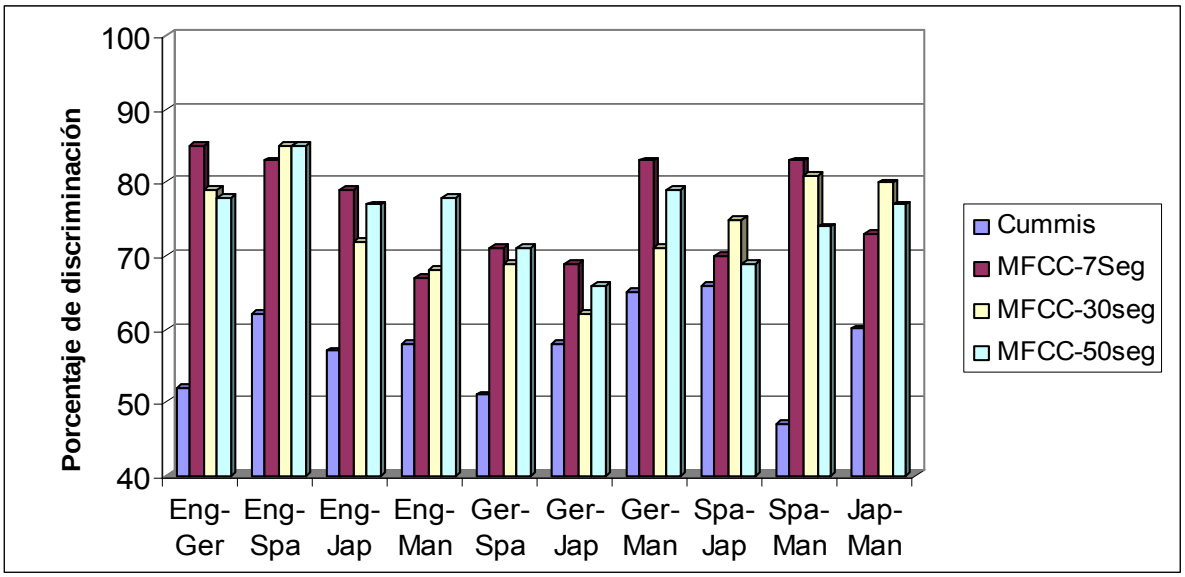

Figura 5.2 Comparativo de los porcentajes de discriminación con muestras de señal de voz de 7, 30 y 50 segundos, utilizando ganancia de información, contra el porcentaje obtenido por Cummins.

Los resultados comparándonos contra Rouas et al [22] se muestran en las tablas 5.16, 5.17 y 5.18, y de la misma manera que mostramos en la figura 5.3 un comparativo entre los diferentes tamaños de muestras y los resultados de Rouas.

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**85** (60)	**83** (68)	**67** (75)	**81** (68)	**79** (68)	76 (79)	**85** (77)	**86** (76)
Alemán	-	**71** (59)	**83** (62)	**85** (66)	**69** (66)	70 (71)	**77** (70)	68 (72)
Español	-	-	**83** (81)	**71** (62)	**70** (63)	63 (76)	54 (65)	64 (67)
Mandarín	-	-	-	**80** (50)	**73** (51)	65 (74)	**79** (74)	**75** (76)
Vietnamita	-	-	-	-	**72** (69)	**68** (56)	63 (71)	**72** (67)
Japonés	-	-	-	-	-	62 (66)	**67** (59)	61 (67)
Coreano	-	-	-	-	-	-	**66** (62)	65 (75)
Tamil	-	-	-	-	-	-	-	66 (70)

Tabla 5.17 Porcentajes de discriminación con muestras de señal de voz de 7 segundos, utilizando ganancia de información, entre paréntesis el porcentaje obtenido por Rouas [22].

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**79** (60)	**85** (68)	68 (75)	**81** (68)	**72** (68)	74 (79)	**82** (77)	**77** (76)
Alemán	-	**69** (59)	**71** (62)	**68** (66)	62 (66)	54 (71)	68 (70)	65 (72)
Español	-	-	**81** (81)	**70** (62)	**75** (63)	74 (76)	56 (65)	62 (67)
Mandarín	-	-	-	**80** (50)	**80** (51)	68 (74)	**84** (74)	**73** (76)
Vietnamita	-	-	-	-	**69** (69)	**72** (56)	70 (71)	**78** (67)
Japonés	-	-	-	-	-	61 (66)	**67** (59)	**71** (67)
Coreano	-	-	-	-	-	-	**64** (62)	54 (75)
Tamil	-	-	-	-	-	-	-	**74** (70)

Tabla 5.18 Porcentajes de discriminación con muestras de señal de voz de 30 segundos, utilizando ganancia de información, entre paréntesis el porcentaje obtenido por Rouas [22].

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**78** (60)	**85** (68)	**75** (75)	**76** (68)	**77** (68)	74 (79)	**86** (77)	**81** (76)
Alemán	-	**71** (59)	**79** (62)	**75** (66)	**66** (66)	67 (71)	**77** (70)	57 (72)
Español	-	-	74 (81)	**70** (62)	**69** (63)	74 (76)	**66** (65)	62 (67)
Mandarín	-	-	-	**72** (50)	**77** (51)	66 (74)	**84** (74)	**75** (76)
Vietnamita	-	-	-	-	**70** (69)	**66** (56)	68 (71)	**77** (67)
Japonés	-	-	-	-	-	**68** (66)	**66** (59)	**72** (67)
Coreano	-	-	-	-	-	-	**75** (62)	65 (75)
Tamil	-	-	-	-	-	-	-	**75** (70)

Tabla 5.19 Porcentajes de discriminación con muestras de señal de voz de 50 segundos, utilizando ganancia de información, entre paréntesis el porcentaje obtenido por Rouas [22].

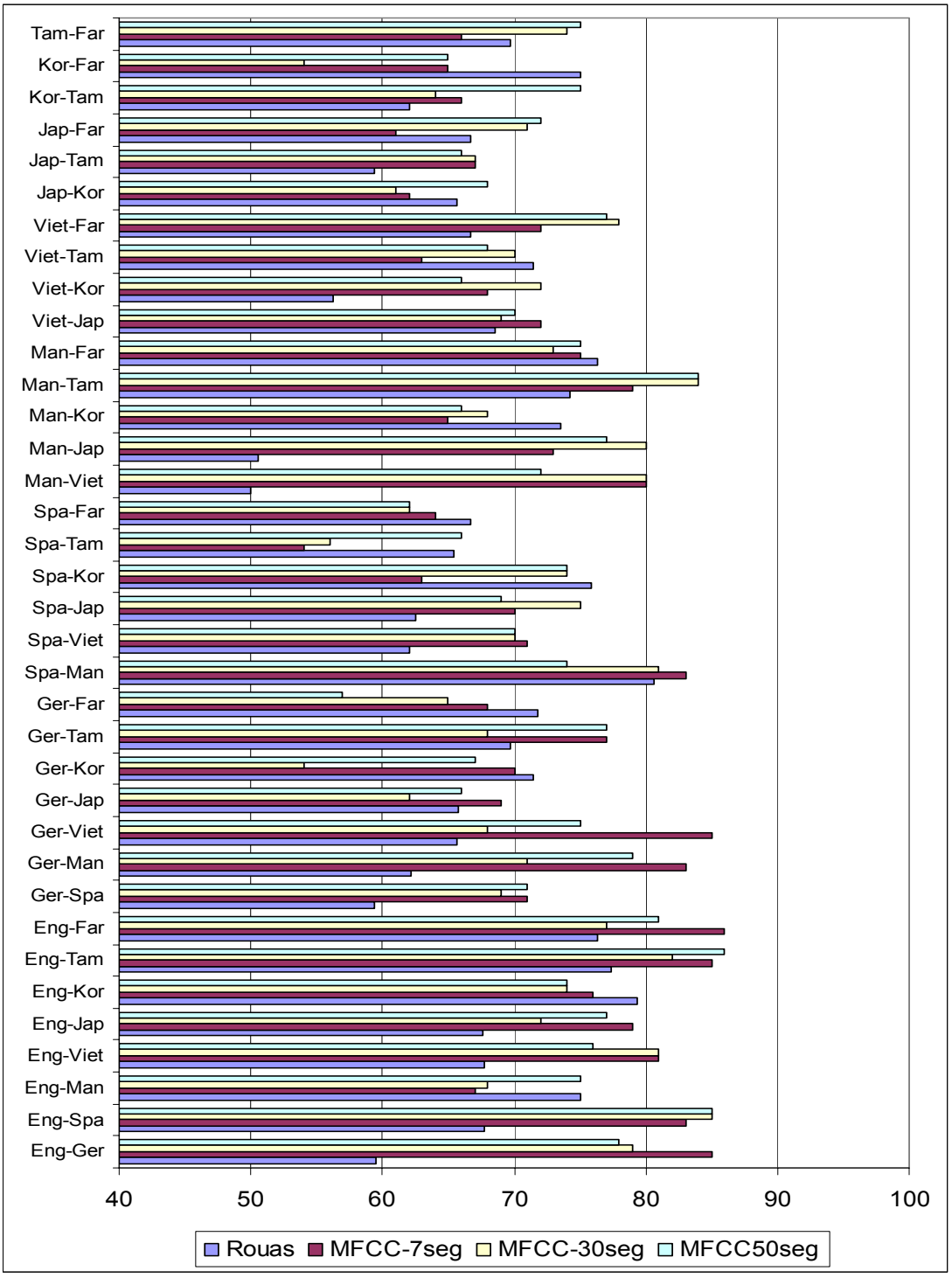

Figura 5.3 Comparativo de los porcentajes de discriminación con muestras de señal de voz de 7,30 y 50 segundos, utilizando ganancia de información, contra el porcentaje obtenido por Rouas[22].

En este caso, los resultados mejoran y superan en la mayoría de los casos a los reportados por Rouas et al [22]. De los 36 clasificadores binarios en 27 casos se tienen mejores resultados, siendo el coreano el idioma con más dificultades para su correcta identificación.

5.3.2 ANÁLISIS DE RESULTADOS

A modo de resumen, en las siguientes gráficas (figuras 5.4 y 5.5) se muestran los resultados al obtener el promedio de cada uno de los porcentajes que se obtuvieron por parejas de idiomas; tomando como base un idioma y obteniendo el promedio de sus porcentajes con los otros idiomas en pares. Aunque no es válido un promedio para nuestra tarea, los promedios nos permiten tener una idea global de los resultados obtenidos por los nuevos métodos de caracterización de la señal de voz.

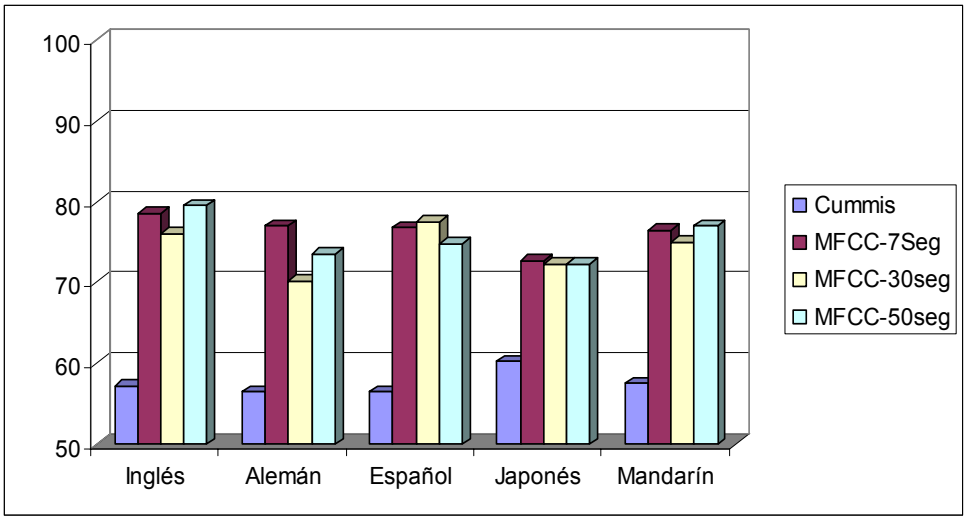

Figura 5.4 Promedio de cada uno de los idiomas utilizando nueva caracterización de los cambios de la señal (deltas y promedios de 16MFCC), con ganancia de información.

De lo que podemos concluir que con el último método de extracción de características superamos claramente a Cummins et al [15]. Sin embargo, el método propuesto no logra superar en todos los casos a los alcanzados por Rouas et al [22].

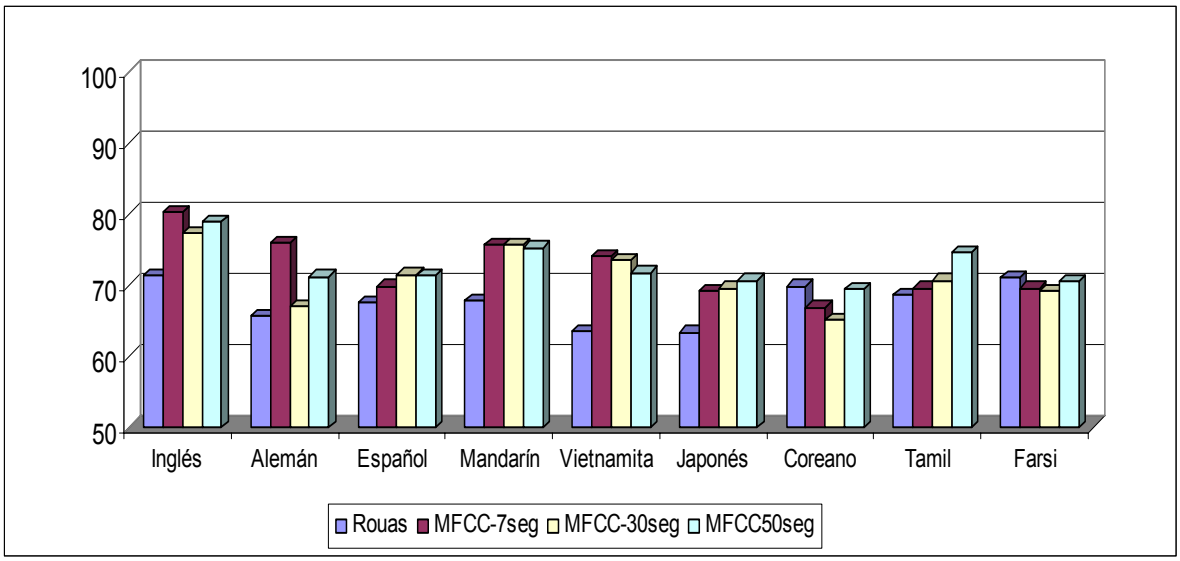

Figura 5.5 Promedio de cada uno de los idiomas utilizando nueva caracterización de los cambios de la señal (deltas y promedios de 16MFCC), con ganancia de información.

La diferencia entre el método propuesto (extraer las características suprasegmentales del habla) y Cummins et al [15], es que en el método de Cummins solo emplearon la frecuencia fundamental F0, tratando de capturar las características de la prosodia. En nuestro caso al aumentar los coeficientes cepstrales e introducir las variaciones de los coeficientes cepstrales por medio de los deltas, ampliamos la base para capturar la información suprasegmental del habla. Por otro lado, con el método propuesto, no tenemos el problema que Rouas et al [22] tienen con los idiomas rítmicamente semejantes. En nuestro caso se logra una mejor discriminación entre idiomas como el inglés-alemán, español-alemán, inglés-español. Así mismo, con los idiomas como el chino,

el vietnamita y el japonés nuestros porcentajes de discriminación son más altos. Sin embargo, con los únicos que no pudimos mejorar fue con el coreano y el farsi.

5.3.3 COMPARATIVO CON DIFERENTES CLASIFICADORES

Con el objetivo, de demostrar la pertinencia de la nueva caracterización de la señal de voz; se realizaron experimentos usando cuatro diferentes clasificadores. Principalmente, tratamos de probar que podríamos tener resultados similares usando diferentes técnicas de clasificación, con el mismo conjunto de datos obtenidos de nuestro proceso de caracterización de la señal de voz. Los clasificadores utilizados fueron:

- Vecinos más cercanos (KNN): NNge (nearest-neighbor general)
- El clasificador Naïve-Bayes
- Máquinas de vectores de soporte (SVM): SMO
- Árboles de decisión: C4.5

La figura 5.6 muestra el promedio de exactitud de cada clasificador para cada uno de los lenguajes, utilizando muestras de señal de voz de 7 segundos. Y la figura 5.7 muestra los resultados para muestras de 50 segundos de señal de voz. De lo cual, podemos observar que Naïve Bayes y la máquina de vectores de soporte SMO, obtuvieron los mejores resultados. Y por el contrario, vecinos más cercanos NNge y árboles de decisión C4.5 obtuvieron (en la mayoría de los casos) resultados bajos. Sin embargo, podemos observar que los cuatro clasificadores son relativamente consistentes. Con esto, podemos confirmar que los resultados obtenidos son consecuencia de la caracterización de la señal de voz y no solamente un resultado de la selección de un algoritmo de clasificación en específico.

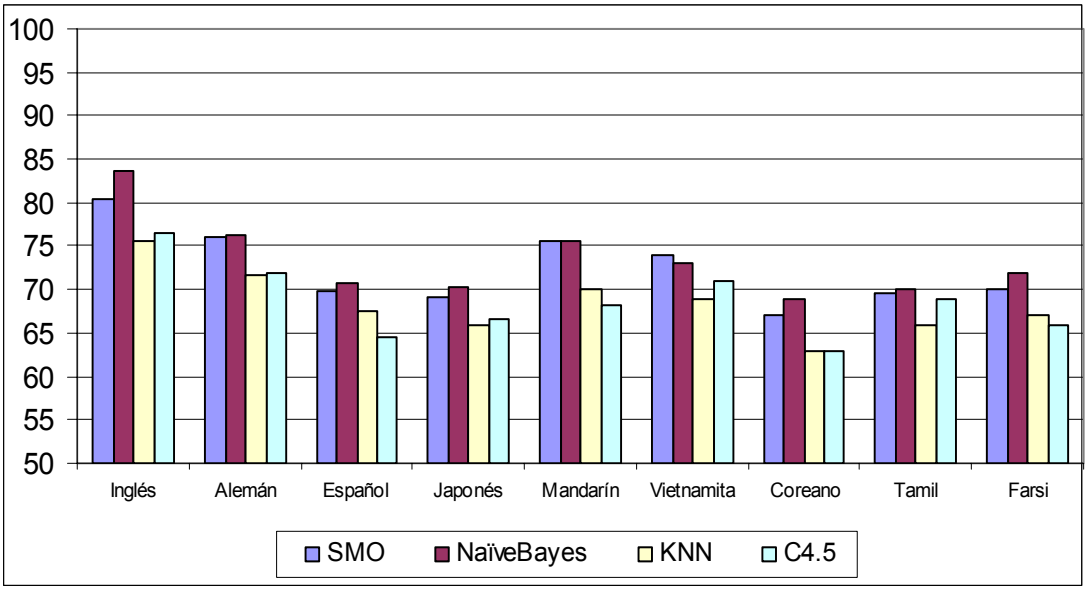

Figura 5.6 Promedio por lenguaje usando diferentes clasificadores.
Con muestras de señal de voz de 7 segundos.

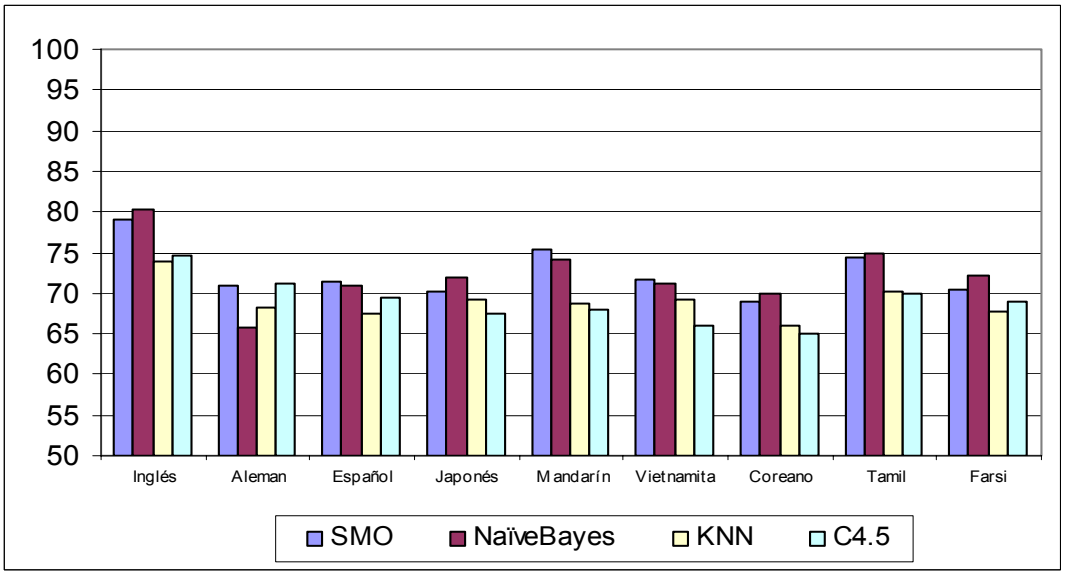

Figura 5.7 Promedio por lenguaje usando diferentes clasificadores.
Con muestras de señal de voz de 50 segundos.

5.5 CONCLUSIONES

De los resultados alcanzados por nuestro método podemos concluir que el nuevo conjunto de características acústicas es una alternativa apropiada en la tarea de identificación automática de idiomas. La caracterización propuesta supera lo obtenido por Cummins et al [15], cuyo trabajo usa la frecuencia F0 para capturar la información suprasegmental. En nuestro caso al incluir el cálculo de deltas Δ_1, Δ_2 y Δ_3 –los cambios de los coeficientes– se capturan los cambios que se realizan en la pronunciación en cada idioma.

Para el caso de Rouas et al [22] superamos la mayoría sus resultados (en 27 de los 36 clasificadores binarios comparados), ellos tuvieron problemas en la identificación de la pareja de inglés y alemán, que son idiomas con ritmos semejantes (pertenecen al grupo de *stress-timed*). En nuestro caso los resultados entre esos dos lenguajes (inglés y alemán) están por encima de los reportados por Rouas. Cabe mencionar que el método de Rouas, que consiste en el uso de las características de entonación –relación entre intervalos vocálicos y consonánticos– obtuvo buenos resultados, como ellos lo mencionan, en la discriminación entre lenguajes *stress-timed* contra los que utilizan la entonación como un marcador léxico (el chino-mandarín y vietnamita). En nuestro caso se tienen resultados similares al discriminar entre los tres grupos rítmicos: *stress-timed, syllable-timed* y *mora-timed.* Por ejemplo, para el japonés –un lenguaje *mora-timed*– fue identificado tanto para lenguajes *stress-timed* como lenguajes tonales (aquellos que usan la entonación como marcador léxico). Por otro lado, cabe resaltar los bajos resultados alcanzados en el caso del coreano, en este caso inferiores a los alcanzados por Rouas et al [22].

Una segunda observación a discutir del método propuesto, es la mínima o nula ganancia al utilizar muestras grandes de señal de voz. Intuitivamente mientras mayor información –más segundos de grabación– se esperaba incrementar las exactitudes de los clasificadores. Desafortunadamente, este es un punto en contra de esta caracterización pues las variaciones capturadas por los deltas Δ_1, Δ_2 y Δ_3 tiendan a equilibrarse con muestras de señal de voz grandes, y al parecer, a acercarse entre los diferentes idiomas.

Una posible explicación es la mayor presencia de silencios y pausas mientras más grande sea la muestra de señal de voz.

Otra observación más es respecto al número de coeficientes cepstrales. Al utilizar la ganancia de información como método de reducción de dimensionalidad pudimos observar que entre los atributos seleccionados encontramos a los coeficientes 13, 14, 15 y 16, con ello se tiene evidencia de que el aumentar el número de coeficientes es pertinente para esta tarea. Sin embargo, es precisamente esta circunstancia la que nos lleva a reflexionar sobre las posibilidades de la caracterización propuesta. En este sentido la caracterización puede ser enriquecida con muchos otros atributos. Para empezar puede ampliarse el número de coeficientes a 24 o más. También es posible estudiar la posibilidad de usar un Δ_4 o un Δ_5, así como, incluir variaciones de los deltas (es decir, los dobles Δ). O considerar la energía y su derivada, etc. En fin existen aún muchas otras variantes sobre el método base propuesto. No obstante todas estas nuevas variantes estarían basadas en la transformada de Fourier y más específicamente de los coeficientes cepstrales de frecuencia Mel (MFCC). Esta transformada nos obliga a mantener pequeñas ventanas afectando la adecuada extracción de las frecuencias bajas. Como se vio en capítulos anteriores, las características suprasegmentales se encuentran en las frecuencias bajas de la señal de voz. Es por ello que decidimos explorar otro tipo de extracción de características: las wavelet.

CAPÍTULO 6

UNA NUEVA CARACTERIZACIÓN ORIENTADA AL RITMO

Recordemos que Cummins, Samouelian y Rouas, han trabajado con la frecuencia fundamental F0, ligándola con la prosodia y el ritmo, es decir, con las características suprasegmentales del habla. Como hemos visto en el capítulo 3, existe evidencia de que la prosodia y el ritmo son importantes en la identificación de los idiomas. La frecuencia fundamental F0 es la más baja de todas las frecuencias que componen la señal [23], de ahí la importancia de extraer adecuadamente las bajas frecuencias que es donde se encuentran las características suprasegmentales. Entonces necesitamos capturar de la señal de voz, las bajas frecuencias con una muy buena resolución. Para ello proponemos el uso de las wavelet motivados por su capacidad de representar señales adecuadamente en los dominios del tiempo y frecuencia [27]; Además, la transformada wavelet hace una separación entre las altas y bajas frecuencias, permitiendo una buena resolución en las bajas frecuencias que es donde suponemos están la prosodia y el ritmo. Este método es completamente diferente a los usados anteriormente basados en la transformada de Fourier. Cabe mencionar que las wavelet han sido utilizadas en el reconocimiento del habla [18][19], pero hasta ahora no se había reportado ningún trabajo aprovechando el uso de wavelet en la identificación de idiomas.

De acuerdo con lo visto en el capítulo 2, la STFT tiene un defecto al generar una buena resolución tanto en tiempo como en frecuencia de manera instantánea ya que el ancho de la ventana es fijo. Vimos que el ancho de ventana constituye un parámetro de gran importancia ya que a través de éste podemos establecer el grado de resolución tanto de tiempo como de frecuencia. Si la ventana es muy pequeña tenemos una buena resolución en tiempo pero una mala resolución en frecuencia y por el contrario, si la ventana es muy grande tendremos una buena resolución en frecuencia pero una mala resolución en tiempo. Dicho problema es solucionado por medio de las wavelet, ya que con ellas tenemos una función ventana que tiene la capacidad de cambiar su ancho en forma automática dependiendo del contenido espectral del segmento de la señal analizada, porque una situación ideal del análisis sería tener una buena resolución en tiempo para frecuencias altas y una buena resolución en frecuencia para el contenido de la señal de voz en frecuencias bajas. Situación que es importante para nosotros si queremos capturar la información suprasegmental del habla, ya que la voz es una señal cuya amplitud varía en forma rápida y abrupta en el tiempo, además su contenido de frecuencias es variable de un instante de tiempo a otro, es decir, es una señal no estacionaria y cuya información acerca de la prosodia, el ritmo, la duración y la entonación están ligadas a las frecuencias bajas.

La transformada wavelet descompone la señal de voz en lo que se conoce como multiresolución [27], lo que nos permite obtener un detalle de las frecuencias bajas en el tiempo. Dicha multiresolución nos da una localización tiempo-frecuencia instantánea de la señal, representación que puede explicarse como un pentagrama musical, donde la localización y forma de la figura musical nos dice cuando ocurre el tono y cuál es su frecuencia. Esto significa que la mayor parte de la energía de la señal es bien representada por unos pocos coeficientes. Mientras que un coeficiente de Fourier representa un componte en un intervalo de tiempo, un coeficiente wavelet es bien localizado en el tiempo.

Por otro lado, sabemos que las wavelet son robustas para el procesamiento de señales con ruido, esto nos ayuda cuando la señal de voz es espontánea. La transformada wavelet acentúa la información más sobresaliente acerca de la señal hablada y la hace más robusta. En nuestro trabajo utilizamos la transformada wavelet de Daubechies, ver

capítulo 2. A continuación detallamos el proceso de extracción de características usando las wavelet para la identificación del lenguaje hablado.

6.1 EL USO DE LA TRANSFORMADA DAUBECHIES

La transformada Wavelet descompone la señal en niveles sucesivos de baja-alta frecuencias (multiresolución). Esta caracterización permite obtener en detalle la descripción de las señales, y hace más clara la distinción entre bajas y altas frecuencias. Esto es de especial importancia para nuestra aplicación, porque las bajas frecuencias capturan algunos fenómenos acústicos tales como el ritmo, los cuales son fundamentales para la identificación de los idiomas. Por otro lado, las Wavelet no requieren dividir la muestra de señal de voz en pequeños segmentos para obtener una descripción global. Esta propiedad hace la diferencia contra los otros enfoques usados en el reconocimiento del habla. En las figuras 6.1 y 6.2 se muestran las descomposiciones de la señal de voz por medio de la transformada wavelet Daubechies, para un hablante japonés y otro de español. De las figuras se puede ver que la descomposición fue en 17 niveles para cada muestra. El primer nivel representa la banda de más alta frecuencia y las ventanas más cortas. Y a la inversa el último nivel de descomposición corresponde con la banda de frecuencia más baja y las ventanas más anchas. Si los coeficientes wavelet obtenidos para un nivel en concreto poseen magnitudes pequeñas (valores próximos a ceros), se espera que esos coeficientes wavelet estén en los siguientes niveles de descomposición. En las figuras hemos representado la magnitud de los coeficientes con diferentes niveles de gris. El color blanco indica que la señal de voz está totalmente representada tal y como es la señal, en caso contrario el color negro indica que la señal no está representada. Eso significa que la señal en ese tiempo no tiene ese tipo de frecuencia.

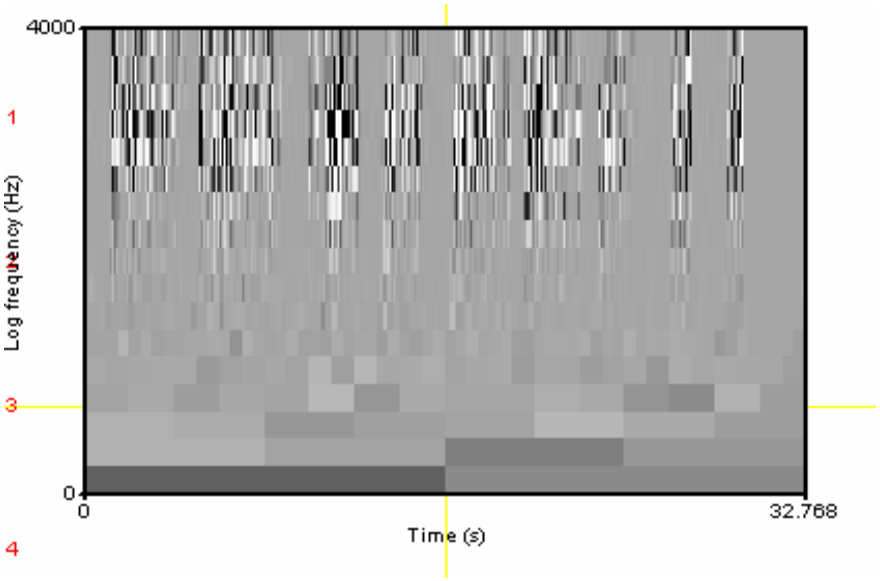

Figura 6.1 Descomposición de la señal de voz por medio de wavelet de un hablante japonés, muestra de 30 seg.

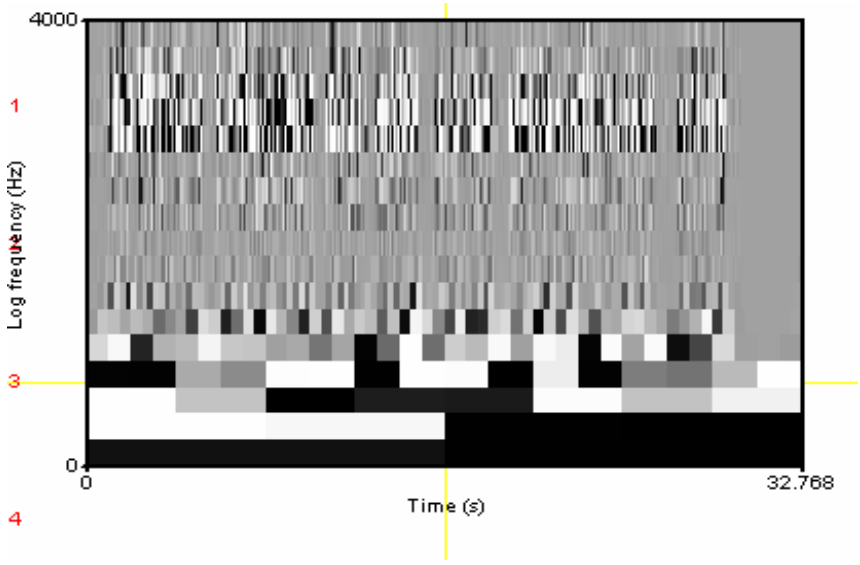

Figura 6.2 Descomposición de la señal de voz por medio de wavelet de un hablante de español, muestra de 30 seg.

Al igual que la transformada rápida de Fourier, la transformada discreta wavelet (DWT) es una operación rápida y lineal que opera sobre un vector de datos, transformándolo en un vector diferente numéricamente pero de la misma longitud. La transformada wavelet es invertible y ortogonal. La transformada inversa cuando es vista como una gran matriz, es simplemente la transpuesta de la transformada. Como se mencionó anteriormente, a diferencia de las funciones de seno y coseno que definen la transformada única de Fourier; en la DWT no hay un solo conjunto único de funciones wavelet. Un conjunto particular de wavelet son las Daubechies. Que en nuestro trabajo utilizamos la db4, con el conjunto de coeficientes del filtro dado por la tabla 2.1 del capítulo 2. Por lo tanto tenemos 4 coeficientes, $c_0,\ldots, c_3$. Y considere la matriz de transformación siguiente, que actúa sobre una columna de un vector de datos.

$$\begin{pmatrix}
c_0 & c_1 & c_2 & c_3 & & & & & & & & \\
c_3 & -c_2 & c_1 & -c_0 & & & & & & & & \\
 & & c_0 & c_1 & c_2 & c_3 & & & & & & \\
 & & c_3 & -c_2 & c_1 & -c_0 & & & & & & \\
 & & & & & & \ddots & & & & & \\
 & & & & & & & c_0 & c_1 & c_2 & c_3 & \\
 & & & & & & & c_3 & -c_2 & c_1 & -c_0 & \\
c_2 & c_3 & & & & & & & & & c_0 & c_1 \\
c_1 & -c_0 & & & & & & & & & c_3 & -c_2
\end{pmatrix} \tag{6.1}$$

Las entradas en blanco significan ceros. El primer renglón genera un componente de datos convolucionados con los coeficientes wavelet daubechies $c_0,\ldots,c_3$. Igualmente el tercero, quinto y los otros renglones nones. Si los renglones pares siguen este patrón, desplazados en una posición, entonces la matriz sería circular, una ordinaria convolución. Los renglones pares ejecutan una convolución diferente, con coeficientes c_3, $-c_2$, c_1, $-c_0$. La

acción de la matriz, sobretodo, es así ejecutar 2 convoluciones relacionadas. La operación de multiplicar a esta matriz con la señal de entrada es igual a la operación que realiza la ecuación 2.2 del capítulo 2. Para el caso de daubechies db4, los coeficientes $c_0,\ldots, c_3$, son los coeficientes del filtro h, pasa-bajas. Y por el signo de menos los coeficientes $c_3, -c_2, c_1, -c_0$ son los coeficientes del filtro g, pasa-altas.

La ecuación 2.2 de la DWT es sólo para una escala. Para obtener otra escala se tiene que hacer las mismas operaciones, pero cambiando la señal *x(n)* de entrada, por la señal $x_0(n)$. De esta forma se continúa para obtener la DWT a diferentes escalas (multiresolución). En nuestro caso a 17 diferentes niveles. En otras palabras, la DWT consiste en aplicar los coeficientes wavelet de la matriz 6.1 al vector de entrada original, el cual lo representamos por $y_0,\ldots,y_n$. Para obtener las siguientes escalas realizamos la misma operación con la matriz 6.1 pero tomando el vector de salida del filtro pasa-bajas pero de longitud N/2 (quita la mitad de los valores), e interpagina las mitades restantes. Para el siguiente paso tomamos el vector del filtro pasa-bajas de longitud N/4, repitiendo el mismo procedimiento hasta llegar a los dos últimos componentes. En la figura 6.3 se muestra esta secuencia, con un vector de entrada de longitud N=12.

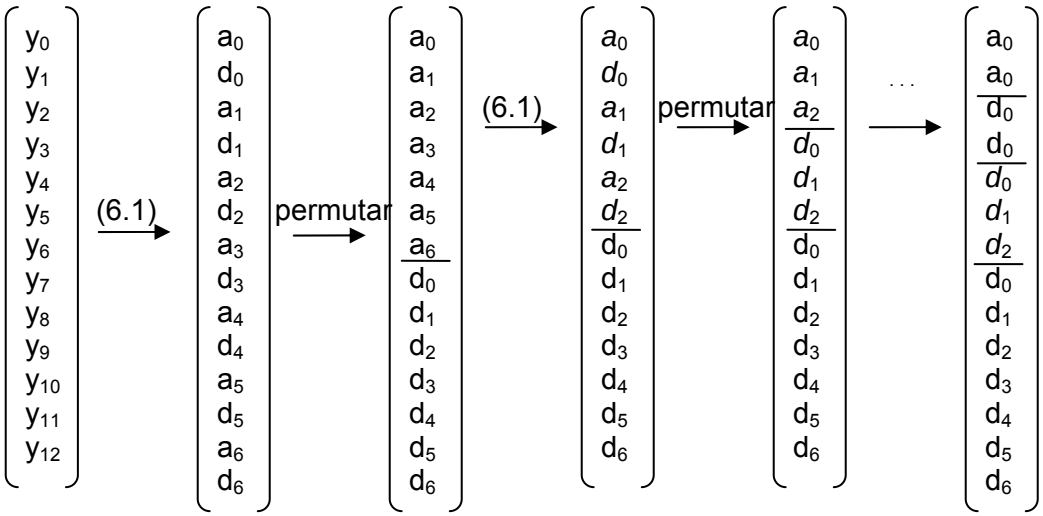

Figura 6.3 Secuencia de la aplicación de la matriz 6.1 con el vector de entrada.

Para que tal caracterización sea útil, debe ser posible reconstruir el vector de datos original de longitud N desde sus N/2 componentes “a” (aproximación) del filtro pasa-bajas y sus N/2 componentes “d” (detalle) del filtro pasa-altas. Para tal efecto requerimos que la matriz 6.1 sea ortogonal, y esto se consigue con la transpuesta de la matriz.

Este proceso se automatizó por medio de un programa hecho en la herramienta de proceso acústico PRAAT versión 4.0.5 [64], que incluye la fragmentación del tamaño de la muestra de 10, 30 y 50 segundos; y la obtención de las características acústicas de la señal de voz aplicando a cada una de esas muestras la trasformada Daubechies db4, es decir con cuatro coeficientes. Los coeficientes Daubechies se muestran en la tabla 2.1 del capítulo 2. De aplicar la transformada discreta Daubechies db4 obtenemos un conjunto de coeficientes wavelet, de la manera mencionada anteriormente. El total de coeficientes wavelet resultantes es de acuerdo el tamaño de muestra, por lo tanto para muestras de 10 segundos de señal de voz obtenemos 131072 coeficientes wavelet, para muestras de 30 segundos obtenemos 262144 y para muestras de 50 segundos son 524288.

Una vez que hemos obtenido el conjunto de coeficientes para una señal, es necesario distinguir entre aquellos correspondientes a las altas y bajas frecuencias. Para ello nos basamos en la magnitud de los coeficientes. Los coeficientes con magnitudes grandes corresponden a una buena representación de la señal de voz, y los coeficientes con magnitudes pequeñas (cercanos a cero) corresponden a una mala representación de la señal o que no existe esa frecuencia en ese tiempo [27]. Para extraer únicamente aquellos coeficientes de alta magnitud, empleamos el método de truncado de aproximación. A continuación se describe dicho método.

6.1.1 TRUNCADO DE APROXIMACIÓN

El truncado de aproximación es muy utilizado en la compresión de imágenes, para reducir los datos a procesar, que regularmente para una imagen son datos redundantes. Generalmente, la energía de las imágenes se concentra en las frecuencias bajas. Una

imagen tiene un espectro que se reduce con el incremento de las frecuencias. Para obtener los píxeles más significativos utilizan el truncado de aproximación. De manera similar para nuestro caso, necesitamos obtener los coeficientes wavelet de las bajas frecuencias, por lo que proponemos el uso del truncado de aproximación para nuestra tarea.

Para ver como trabaja el truncado de aproximación [65], considere el ejemplo mostrado en la figura 6.4. La parte superior de la figura muestra una señal arbitraria de una función de prueba. Suave excepto por un pico, ejemplificado sobre un vector de longitud 2^{10}. La parte inferior de la figura muestra, en una escala logarítmica, el valor absoluto de los componentes del vector después de aplicar la transformada wavelet discreta Daubechies Db2, con cuatro coeficientes. De la gráfica se puede observar de derecha a izquierda los diferentes niveles de descomposición, 512-1023, 256-511, 128-255, etc. En cada nivel, los coeficientes wavelet tienen magnitudes grandes en las posiciones cercanas al pico, o muy cerca de los límites tanto izquierdo como derecho del rango de descomposición.

La curva punteada en la figura 6.4 (b) muestra las mismas amplitudes como de la curva sólida, pero en orden decreciente de la magnitud de los coeficientes wavelet. Se puede observar, por ejemplo, que los 130 coeficientes wavelet de magnitud alta tienen una amplitud menor que 10^{-5} de los coeficientes más grandes cuya magnitud es ~10. Entonces, la función ejemplo puede ser bien representada por sólo los 130 coeficientes wavelet, en vez de los 1024 coeficientes wavelet obtenidos en total por la transformada wavelet discreta. Note que este tipo de truncado se realiza en todo el espacio del vector, el cual no está ordenado como se ve en la gráfica. Sino por el contrario los coeficientes wavelet de magnitud alta se encuentran esparcidos por todo el vector que tiene en total 1024, y los restantes son cero o muy cercanos a cero. Esto es muy importante, ya que los vectores en el espacio wavelet son truncados de acuerdo a la amplitud de sus componentes, no a su posición en el vector. Cuando se comprime una función con wavelet, se tienen que tener presentes ambos valores y la posición de los coeficientes que no son ceros.

Por lo tanto, tenemos que escoger un porcentaje de truncado, si aplicamos el 0.1 como umbral de truncado, significa que estamos conservando el 10 % de los coeficientes wavelet de magnitud alta y poniendo a cero el resto de ellos. Simplemente estamos eliminando los coeficientes de magnitud más pequeña.

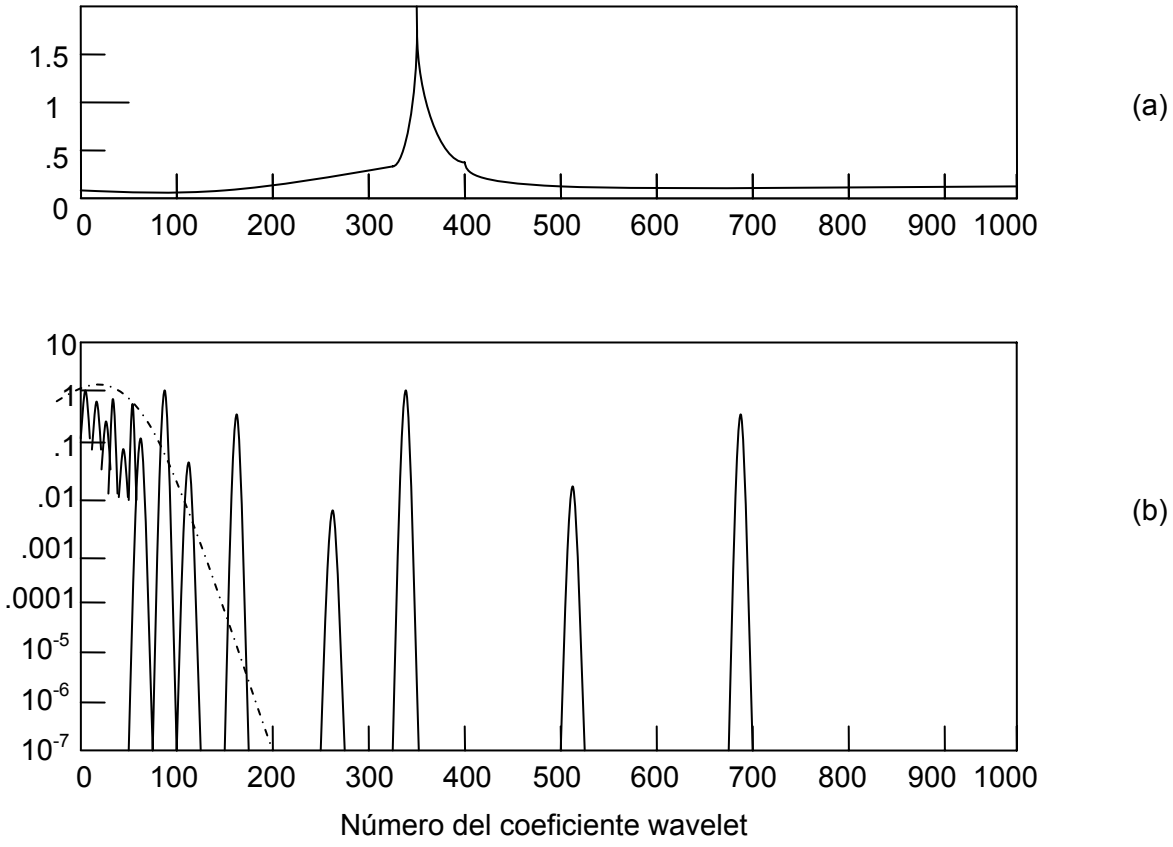

Figura 6.4 (a) Una función arbitraria, con pico, muestreada sobre un vector de longitud de 1024. (b) Los valores absolutos de los 1024 coeficientes wavelet producidos por la transformada wavelet discreta de la función de (a). Note la escala es logarítmica. La curva punteada es el resultado de ordenar las amplitudes en orden decreciente. De lo que podemos observar que solamente 130 de los 1024 coeficientes son más grandes que 10^{-4}. (Tomada de [65])

Entonces aplicando el truncado de aproximación son un umbral de 0.01, enfocaremos nuestro interés en sólo el 1% de los coeficientes wavelet. Es decir, para

muestras de 10 segundos reducimos de 131072 a 1312, de 262144 de 2623 (para muestras de señal de voz de 30 segundos) y de 524,288 a 5244 (para muestras de 50 segundos), enfocándonos en la información más relevante para nuestra tarea.

6.1.2 DETERMINANDO EL PORCENTAJE DE TRUNCADO

Para obtener el porcentaje de truncado más adecuado se realizaron experimentos con varios umbrales. Para el procesamiento de imágenes, de acuerdo a lo que se requiera, hay comprensión de imágenes con buenos resultados variando el truncado de 25% hasta de un 5% [65]. En nuestro caso, realizamos experimentos con diferentes umbrales entre el 10% y el 0.1% el porcentaje de truncado, y variando los tamaños de muestras de señal de voz de 10 y 50 segundos. Recordemos que los tamaños de los vectores de coeficientes wavelet aumentan de acuerdo al tamaño de la muestra de señal de voz. En la figura 6.5 se muestran los resultados con muestras de 10 segundos y variando el porcentaje de truncado de 0.1% a 10%.

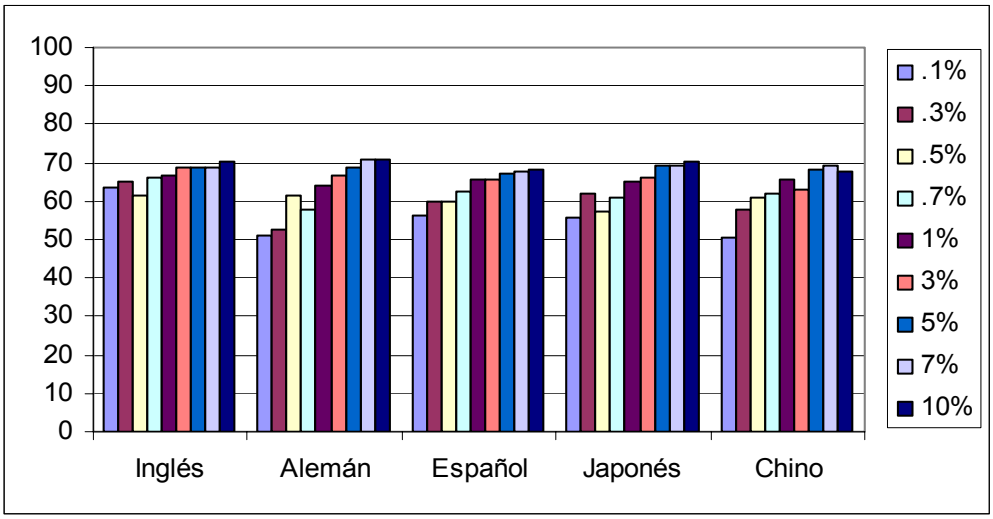

Figura 6.5 Comparativo de resultados variando el porcentaje de truncado para cinco idiomas con muestras de 10 segundos.

De la figura 6.5 podemos observar que los resultados muestran que con umbrales muy pequeños 0.1% empezamos a perder información y comienza a decaer nuestra exactitud. En el caso de umbrales de 1% a 10% la exactitud se mantiene pero la cantidad de coeficientes wavelet que conservamos es mayor mientras más grande es el umbral.

Para el caso de 10 segundos al aumentar el porcentaje de truncado el manejo de los coeficientes es posible, pero cuando aumentamos el tamaño de la muestra a 50 segundos el tamaño de los vectores de coeficientes wavelet aumenta, haciendo su manejo difícil para las herramientas de clasificación. Los resultados son similares al variar los porcentajes de truncado, ya que estos están aumentando en proporciones pequeñas, pero no así los tamaños de los vectores resultantes. En la gráfica 6.6 mostramos los resultados variando el porcentaje de truncado con muestras de 50 segundos. De la cual observamos que presenta el mismo comportamiento. Tomando en cuenta lo anterior y teniendo en contra la dimensionalidad de los vectores, se determino mantener un umbral del 1%.

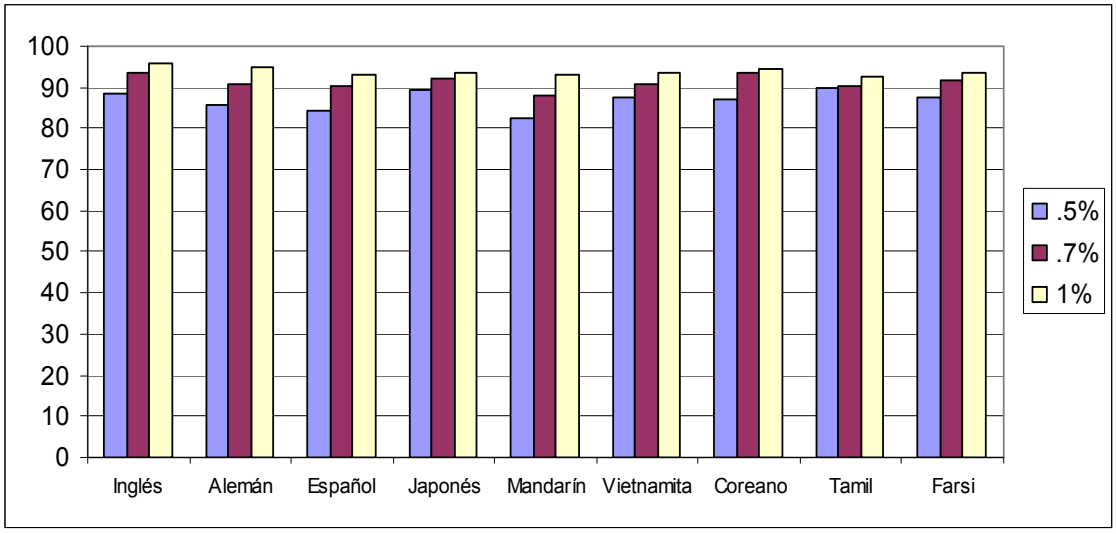

Figura 6.6 Comparativo de resultados variando el porcentaje de truncado para nueve idiomas con muestras 50 segundos.

6.1.3 RESULTADOS

Después de aplicar el truncado de aproximación, para obtener los coeficientes wavelet más representativos de la señal de voz, aún tenemos vectores de características grandes y éstos aumentan cuando utilizamos muestras de señal de voz más grandes, en específico con muestras de 30 y 50 segundos. Por lo tanto, se utilizó un método de reducción de dimensionalidad: ganancia de información (vista en la sección 5.1.4).

Utilizamos las mismas bases de experimentación mencionadas en el capítulo 5 sección 5.1. A cada una de esas muestras se les aplica la transformada wavelet de Daubechies, con 4 coeficientes y normalizada a [-1,1]. Las pruebas fueron realizadas con el corpus OGI_TS [21] con cinco idiomas como en [15] y nueve idiomas como en [22], discriminando entre pares de idiomas. Los resultados iniciales demuestran que la extracción de características acústicas de la señal de voz por medio de wavelet obtiene mejores porcentajes de discriminación que los reportados en el estado del arte, inclusive que los obtenidos con nuestro método de extracción de características basado en Fourier (MFCC) Los resultados se muestran en la tabla 6.1 para 10 segundos de señal de voz, comparándonos con Cummins et al [15], en la tabla 6.2 aumentando el tiempo de muestra de señal de voz a 30 segundos y finalmente se muestra en la tabla 6.3 con muestras de 50 segundos. Cabe mencionar que los resultados de Cummins son con muestras de 50 segundos, por lo que el comparativo es el mostrado en la tabla 6.3. Pero a modo de referencia también mostramos los resultados de Cummins en las tablas con 10 y 30 segundos. De lo cual observamos que en general se obtuvieron mejores resultados con el proceso acústico por medio de wavelet, superando lo obtenido por Cummins. Para las muestras de 50 segundos el porcentaje es más alto y no así para las muestras de 30 y 10 segundos, lo que implica que para una buena discriminación, las muestras de señal de voz deben ser mayores a 30 segundos. Sin descartar que los resultados de 30 y 10 segundos son mejores que los obtenidos por Cummins.

	Alemán	Español	Japonés	Mandarín
Inglés	**63** (52)	**64** (62)	**68** (57)	**72** (58)
Alemán	-	**62** (51)	**62** (58)	**66** (65)
Español	-	-	**70** (66)	**66** (47)
Japonés	-	-	-	**65** (60)

Tabla 6.1 Porcentaje de discriminación obtenido utilizando ganancia de información con muestras de señal de voz de 10 segundos. Entre paréntesis el resultado de Cummins [15].

	Alemán	Español	Japonés	Mandarín
Inglés	**64** (52)	**80** (62)	**63** (57)	**78** (58)
Alemán	-	**67** (51)	**60** (58)	63 (65)
Español	-	-	64 (66)	**64** (47)
Japonés	-	-	-	**61** (60)

Tabla 6.2 Porcentaje de discriminación obtenido utilizando ganancia de información con muestras de señal de voz de 30 segundos. Entre paréntesis el resultado de Cummins [15].

	Alemán	Español	Japonés	Mandarín
Inglés	**75** (52)	**86** (62)	**71** (57)	**78** (58)
Alemán	-	**75** (51)	**76** (58)	**72** (65)
Español	-	-	**70** (66)	**71** (47)
Japonés	-	-	-	**79** (60)

Tabla 6.3 Porcentaje de discriminación obtenido utilizando ganancia de información con muestras de señal de voz de 50 segundos. Entre paréntesis el resultado obtenido por Cummins [15].

En la figura 6.7 se muestra un comparativo de los pares de lenguajes con diferentes tamaños de muestra de señal de voz y los resultados de Cummins. Los resultados superan a Cummins y mejoran mientras más grande es la muestra de señal de voz.

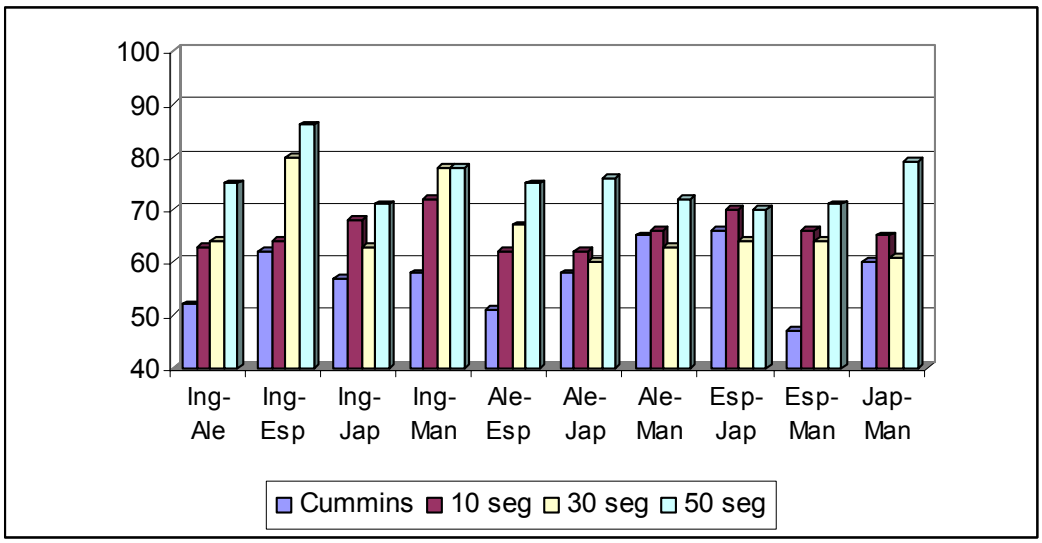

Figura 6.7 Comparativo de los porcentajes de discriminación con muestras de señal de voz de 10, 30 y 50 segundos, utilizando ganancia de información, contra el porcentaje obtenido por Cummins[15].

Para el segundo grupo con nueve idiomas similar a los utilizados por Rouas et al [22], los resultados se presentan en la tabla 6.4, para 50 segundos de muestra de señal de voz y en las tablas 6.5 y 6.6 se muestran los resultados con 30 y 10 segundos de señal de voz respectivamente. Aun con muestras más pequeñas a las usadas por Rouas los resultados demuestran un buen porcentaje de discriminación. Entre paréntesis se muestra el porcentaje obtenido por Rouas et al [22]. Como puede observarse, los resultados no fueron buenos para el lenguaje tamil y para el coreano, pero hemos de considerar que para la mayoría de los lenguajes se obtuvieron mejores porcentajes de discriminación

En este segundo grupo de idiomas al igual que en el primero, al aumentar el tamaño de muestra de 30 a 50 segundos los porcentajes de discriminación aumentan. Indicándonos que este tipo de procesamiento acústico por medio de wavelet debe ser con muestras grandes de señal de voz.

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**75** (60)	**86** (68)	**78** (75)	**72** (68)	**71** (68)	76 (79)	60 (77)	**78** (76)
Alemán	-	**75** (59)	**72** (62)	**73** (66)	**76** (66)	**72** (71)	64 (70)	64 (72)
Español	-	-	71 (81)	**81** (62)	**70** (62)	72 (76)	**70** (65)	**76** (67)
Mandarín	-	-	-	**66** (50)	**79** (51)	**75** (74)	73 (74)	62 (76)
Vietnamita	-	-	-	-	**79** (69)	**76** (56)	68 (71)	**71** (67)
Japonés	-	-	-	-	-	**78** (66)	**77** (59)	**79** (67)
Coreano	-	-	-	-	-	-	62 (62)	73 (75)
Tamil	-	-	-	-	-	-	-	**73** (70)

Tabla 6.4 Porcentaje de discriminación obtenido utilizando ganancia de información con muestras de señal de voz de 50 segundos. Entre paréntesis el resultado obtenido por Rouas [22].

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**64** (60)	**80** (68)	**78** (75)	**75** (68)	63 (68)	73 (79)	65 (77)	**76** (76)
Alemán		**67** (59)	**63** (62)	**72** (66)	60 (66)	65 (71)	63 (70)	**73** (72)
Español			64 (81)	**67** (62)	**63** (63)	55 (76)	**73** (65)	51 (67)
Mandarín				**66** (50)	**61** (51)	71 (74)	66 (74)	60 (76)
Vietnamita					57 (69)	**61** (56)	66 (71)	56 (67)
Japonés					-	**68** (66)	**63** (59)	65 (67)
Coreano					-	-	**62** (62)	69 (75)
Tamil					-	-		**81** (70)

Tabla 6.5 Porcentaje de discriminación obtenido utilizando ganancia de información con muestras de señal de voz de 30 segundos. Entre paréntesis el resultado obtenido por Rouas[22].

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**63** (60)	64 (68)	72 (75)	66 (68)	**68** (68)	60 (79)	58 (77)	68 (76)
Alemán		**62** (59)	**66** (62)	63 (66)	62 (66)	60 (71)	68 (70)	65 (72)
Español			66 (81)	**67** (62)	**70**(63)	59 (76)	**66** (65)	**71** (67)
Mandarín				**62** (50)	**65** (51)	64 (74)	69 (74)	58 (76)
Vietnamita					58 (69)	**71** (56)	65 (71)	60 (67)
Japonés					-	**70** (66)	**65** (59)	**68** (67)
Coreano					-	-	**68** (62)	62 (75)
Tamil					-	-		58 (70)

Tabla 6.6 Porcentaje de discriminación obtenido utilizando ganancia de información con muestras de señal de voz de 10 segundos. Entre paréntesis el resultado obtenido por Rouas [22].

6.1.4 DISCUSIÓN

A modo de resumen, en las siguientes gráficas (figuras 6.8 y 6.9) se muestran los resultados al obtener el promedio de cada uno de los porcentajes que se obtuvieron por parejas de idiomas; tomando como base un idioma y obteniendo el promedio de sus porcentajes con los otros idiomas en pares. Se obtuvo un promedio para tener una idea global de los resultados obtenidos por los nuevos métodos de caracterización de la señal de voz.

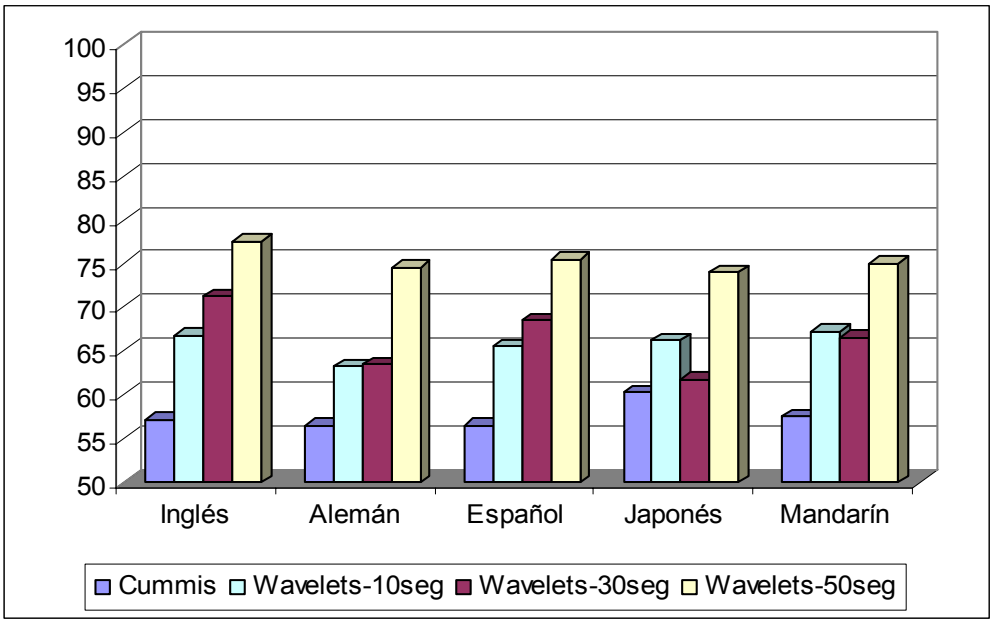

Figura 6.8 Promedio de cada uno de los idiomas utilizando wavelet con ganancia de información contra Cummins et al [15].

En general, podemos decir que el uso de la transformada wavelet para caracterizar la señal de voz en la tarea de identificación del idioma alcanza mejores resultados. Éste es otro tipo de caracterización del ritmo diferente a los usados por Cummins y Rouas, comprobando que existe información importante en las frecuencias bajas para la

discriminación de los idiomas. Sin embargo en este experimento el truncado de aproximación ha eliminado los coeficientes que son ceros dejando un vector comprimido sin respetar los lugares de dichos coeficientes en el vector. De acuerdo a lo visto en la sección 6.1.1, cuando se comprime una señal con el método de truncado de aproximación es importante conservar las posiciones de los ceros en el vector de los coeficientes wavelet si es que deseamos reconstruir la señal de voz (que en nuestro caso no es necesario), pero la idea de eliminar todos los ceros y comprimir el vector no es del todo satisfactoria. Por lo tanto, decidimos experimentar respetando las posiciones de los coeficientes, de tal forma que en algunos casos no se eliminarán todos los ceros del vector de coeficientes wavelet. Es decir, sólo se eliminarán aquellas posiciones que presentan ceros en todas las instancias de los dos idiomas a identificar. De esta manera conservamos información relevante para la discriminación de las lenguas en cuestión. La siguiente sección describe a detalle el método anterior.

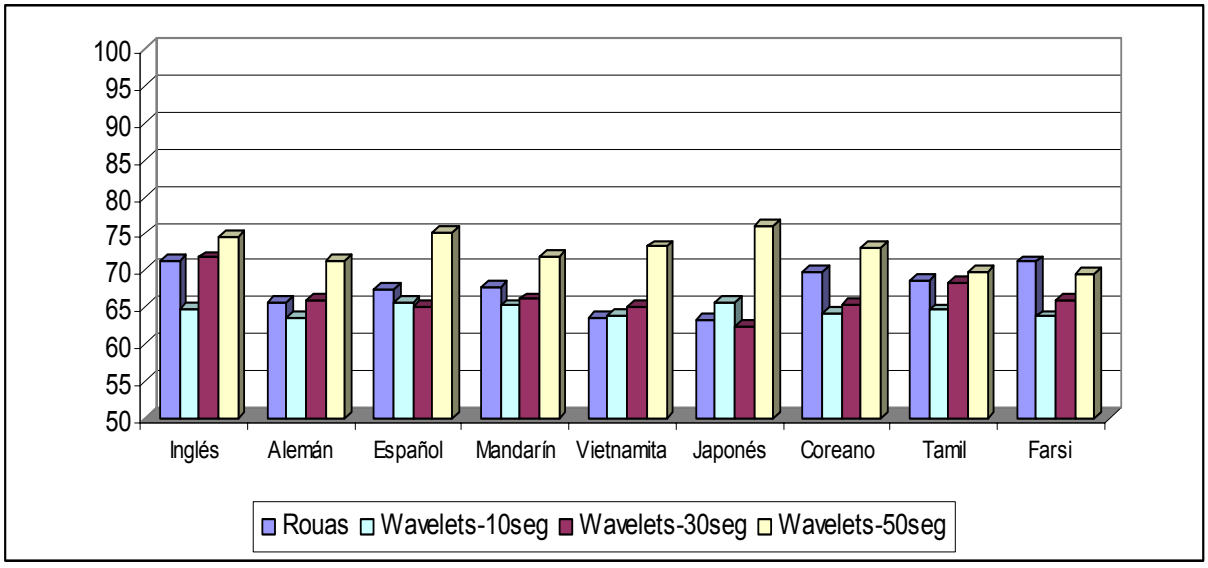

Figura 6.9 Promedio de cada uno de los idiomas utilizando wavelet con ganancia de información contra Rouas et al [22].

6.2 SELECCIONANDO ATRIBUTOS POR PARES DE LENGUAJES

En esta nueva caracterización de la señal de voz utilizamos nuevamente la transformada Wavelet Daubechies db2 [20], con cuatro coeficientes y normalizada a [-1,1]. Utilizando el truncado de aproximación para distinguir los coeficientes wavelets que representan bien la señal de voz (visto en la sección 6.1.1). Por lo tanto, se filtran los coeficientes que tengan magnitud alta extrayendo el 1% de ellos.

En este caso, antes de construir los clasificadores –uno por cada par de lenguajes- es necesario aplicar un proceso para reducir la dimensionalidad. Este proceso consiste en dos grandes pasos. El primero elimina todos los coeficientes iguales a cero posterior al método de truncado de aproximación. Esto es, eliminar los coeficientes cuya magnitud es cero para todas las instancias de ambos lenguajes, evitando comprimir los vectores wavelet para cada idioma de forma independiente. Con ello se obtienen representaciones para cada muestra de voz con todos aquellos coeficientes con información para la tarea de identificación. El segundo paso, consiste en aplicar ganancia de información para identificar los coeficientes con mayor poder discriminativo entre el par de lenguajes en cuestión. Por ejemplo, cuando se construyó el clasificador para el Inglés-Alemán, se calcularon los coeficientes Wavelet para cada muestra de señal de voz, obteniendo 131,072 coeficientes. Entonces, se realiza el filtrado, aplicando el método de truncado de aproximación, y respetando aquellos coeficientes en los que al menos en una instancia es diferente de cero. En este caso, se reduce el número de coeficientes a 37,875. Finalmente, se aplica ganancia de información limitándose a 641 coeficientes. Obviamente, este proceso fue realizado para cada par de lenguajes. En la figura 6.10 se muestra el proceso.

Para los experimentos se consideraron 50 diferentes hablantes para cada lengua y seleccionamos muestras de 10 y 50 segundos por cada hablante. En total usamos 450 diferentes hablantes. El protocolo de experimentación es el mismo que el usado en la sección 5.1.

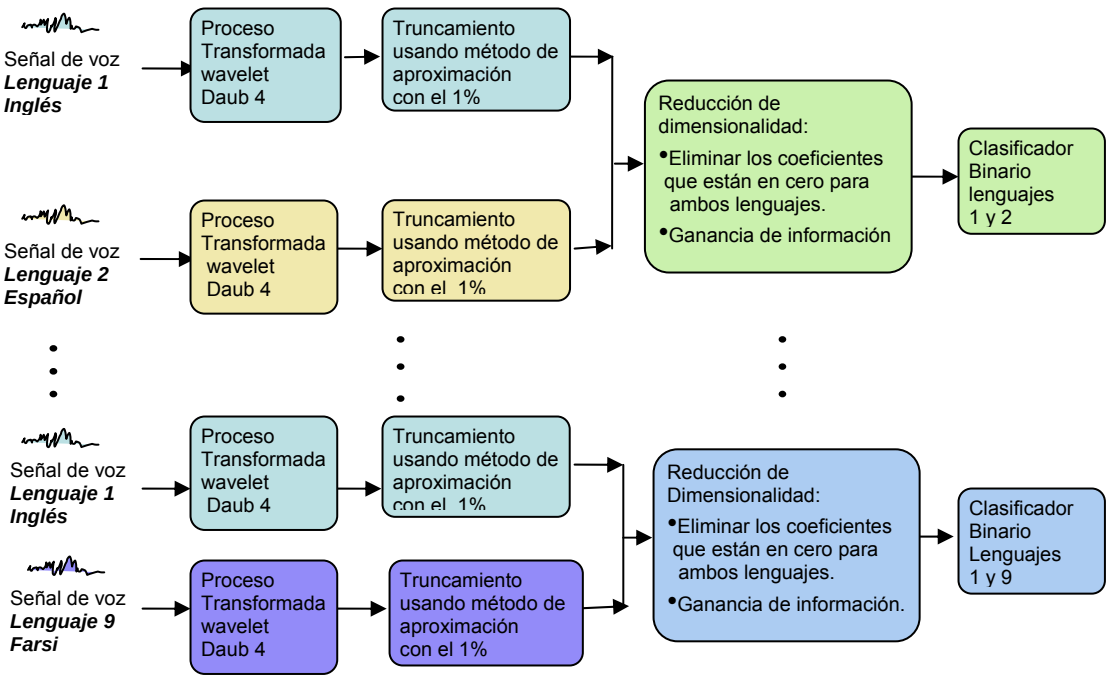

Figura 6.10 Proceso de extracción de características rítmicas usando wavelet

Las tablas 6.7 y 6.8 muestran los resultados usando muestras de señal de voz de 10 y 50 segundos respectivamente. En las tablas también están los resultados de Cummins (indicado por paréntesis) para efectos de comparación. Recordemos que Cummins et al [15] usaron la frecuencia fundamental de la señal como características principal y una red neuronal (LSTM) como método de clasificación.

	Alemán	Español	Japonés	Mandarín
Inglés	**94** (52)	**96** (62)	**94** (57)	**85** (58)
Alemán	-	**80** (51)	**84** (58)	**83** (65)
Español	-	-	**86** (66)	**90** (47)
Japonés	-	-	-	**89** (60)

Tabla 6.7 Comparativo de los porcentajes de discriminación con muestras de señal de voz de 10 segundos.

	Alemán	Español	Japonés	Mandarín
Inglés	**97** (52)	**97** (62)	**96** (57)	**93** (58)
Alemán	-	**93** (51)	**98** (58)	**94** (65)
Español	-	-	**92** (66)	**91** (47)
Japonés	-	-	-	**95** (60)

Tabla 6.8 Comparativo de los porcentajes de discriminación con muestras de señal de voz de 50 segundos.

En todos los casos nuestro enfoque supera los resultados de Cummins, confirmando que la transformada wavelet obtiene una buena resolución de las frecuencias y por lo tanto la extracción de un conjunto de características pertinentes. Estos resultados muestran también, que mientras más grande la muestra de señal de voz mejores los porcentajes de discriminación.

En la tabla 6.9 se muestran los resultados correspondientes a los nueve lenguajes con muestras de señal de voz de 50 segundos. Los resultados fueron obtenidos utilizando Naïve Bayes. En esta tabla se muestra claramente que nuestros resultados están por encima de los reportados por Rouas et al [22] (indicado en paréntesis), el cual utilizó las unidades de ritmo de la señal de voz (por ejemplo, la relación entre intervalos vocálicos y consonánticos) como una de sus principales características para la identificación, y los modelos de mezclas gaussianas (GMM) como técnica de clasificación.

La tabla 6.10 muestra los resultados obtenidos usando muestras de señal de voz de 10 segundos.

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**97** (60)	**97** (68)	**93** (75)	**94** (68)	**96** (68)	**95** (79)	**99** (77)	**96** (76)
Alemán	-	**93** (59)	**94** (62)	**93** (66)	**98** (66)	**98** (71)	**94** (70)	**91** (72)
Español	-	-	**91** (81)	**86** (62)	**92** (63)	**98** (76)	**91** (65)	**94** (67)
Mandarín	-	-	-	**95** (50)	**95** (51)	**93** (74)	**89** (74)	**94** (76)
Vietnamita	-	-	-	-	**93** (69)	**96** (56)	**95** (71)	**95** (67)
Japonés	-	-	-	-	-	**93** (66)	**89** (59)	**94** (67)
Coreano	-	-	-	-	-	-	**95** (62)	**91** (75)
Tamil	-	-	-	-	-	-	-	**90** (70)

Tabla 6.9 Comparativo de los porcentajes de discriminación con muestras de señal de voz de 50 segundos contra Rouas et al [22].

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	**94** (60)	**96** (68)	**85** (75)	**88** (68)	**94** (68)	**83** (79)	**98** (77)	**83** (76)
Alemán	-	**80** (59)	**83** (62)	**87** (66)	**84** (66)	**83** (71)	**80** (70)	**82** (72)
Español	-	-	**90** (81)	**84** (62)	**86** (63)	**88** (76)	**87** (65)	**79** (67)
Mandarín	-	-	-	**85** (50)	**89** (51)	**83** (74)	**85** (74)	**94** (76)
Vietnamita	-	-	-	-	**85** (69)	**84** (56)	**83** (71)	**86** (67)
Japonés	-	-	-	-	-	**83** (66)	**75** (59)	**89** (67)
Coreano	-	-	-	-	-	-	**86** (62)	**87** (75)
Tamil	-	-	-	-	-	-	-	**86** (70)

Tabla 6.10 Comparativo de los porcentajes de discriminación con muestras de señal de voz de 10 segundos contra Rouas et al [22].

6.2.1 COMPARATIVO DE RESULTADOS

A modo de resumen, en las siguientes gráficas (figuras 6.11 y 6.12) se muestran los resultados al obtener el promedio de cada uno de los porcentajes que se obtuvieron por

parejas de idiomas; tomando como base un idioma y obteniendo el promedio de sus porcentajes con los otros idiomas en pares. Como es de recordar, para simplificar y tener una idea global de los resultados obtenidos por los nuevos métodos de caracterización de la señal de voz. Se obtuvo un promedio por idioma.

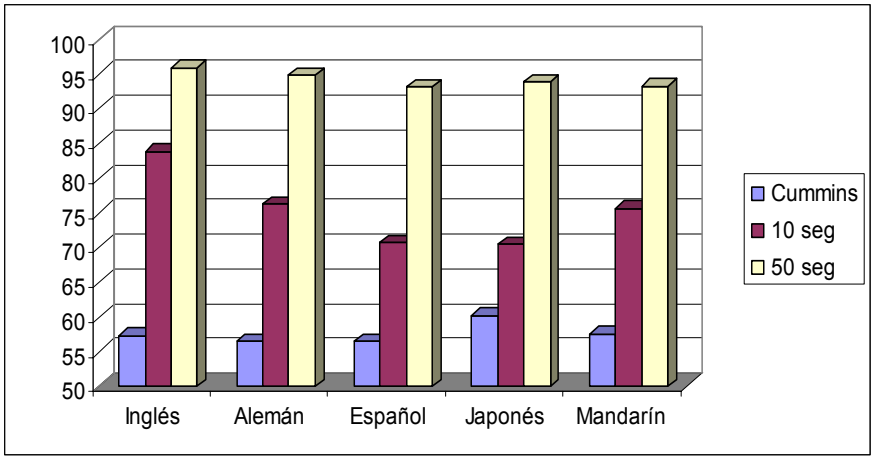

Figura 6.11 Comparativo del promedio de cada uno de los idiomas utilizando wavelet con ganancia de información y matrices de pares de lenguajes contra Cummins et al [15].

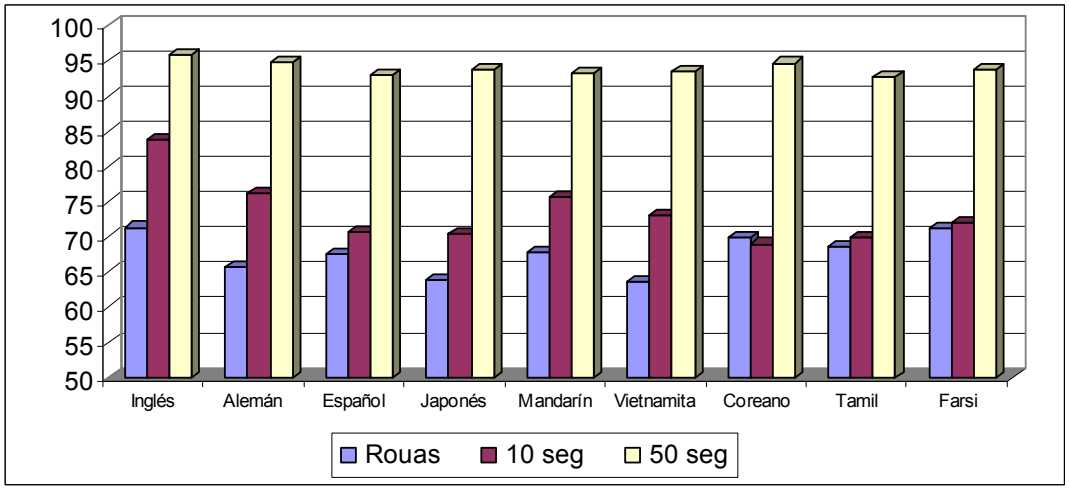

Figura 6.12 Comparativo del promedio de cada uno de los idiomas utilizando wavelet con ganancia de información y matrices de pares de lenguajes contra Rouas et al [22].

6.2.3 COMPARATIVO CON DIFERENTES CLASIFICADORES

Se realizaron experimentos usando cuatro diferentes clasificadores. Con el objetivo, de demostrar la pertinencia de la nueva caracterización de la señal de voz. Principalmente, tratamos de probar que podríamos tener resultados similares usando diferentes técnicas de clasificación, con el mismo conjunto de datos obtenidos de nuestro proceso de caracterización de la señal de voz. Los clasificadores utilizados fueron:

- Vecinos más cercanos (KNN): NNge (nearest-neighbor general)
- El clasificador Naïve-Bayes
- Máquinas de vectores de soporte (SVM): SMO
- Árboles de decisión: C4.5

La figura 6.13 muestra el promedio de exactitud de cada clasificador para cada uno de los lenguajes, utilizando muestras de señal de voz de 10 segundos. Y la figura 6.14 muestra los resultados para muestras de 50 segundos de señal de voz. Las figuras indican que Naïve Bayes, Máquina de Vectores de Soporte (SVM), y vecinos más cercanos KNN obtiene los mejores resultados. Y por el contrario, C4.5 obtuvo los resultados más bajos. Sin embargo, dado el comportamiento consistente de los cuatro clasificadores, entonces podemos asegurar la pertinencia de la caracterización. Esto es, confirmamos que los resultados obtenidos son consecuencia de la caracterización de la señal de voz y no un resultado de la selección del algoritmo de clasificación.

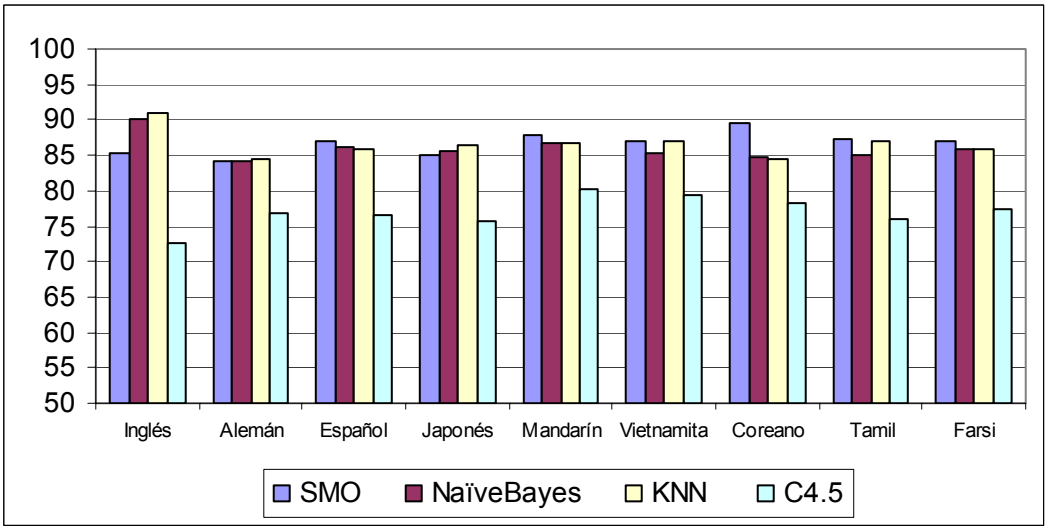

Figura 6.13 Comparativo de promedios por lenguajes usando muestras de 10 segundos.

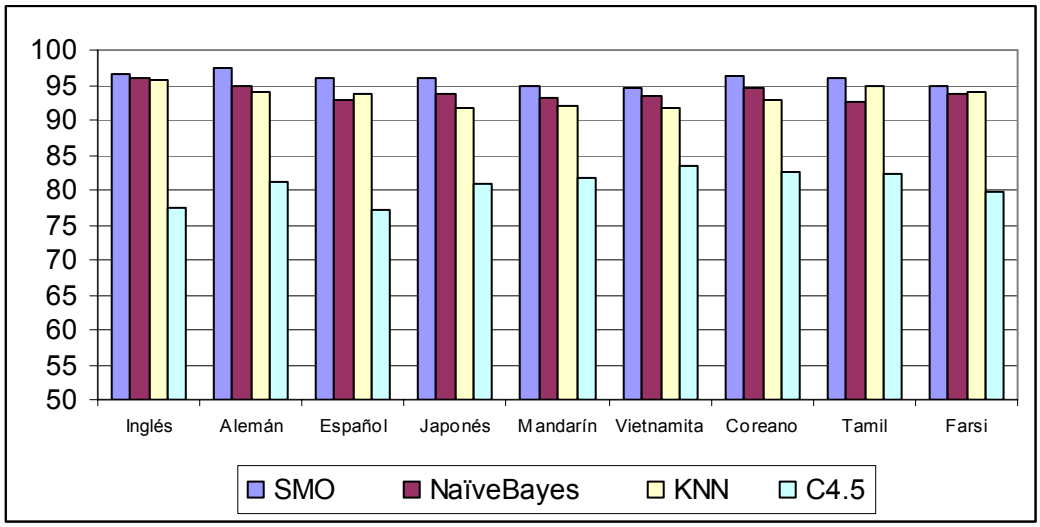

Figura 6.14 Comparativo de promedios por lenguajes usando muestras de 50 segundos

6.3 COMPARATIVO ENTRE LOS DOS MÉTODOS PROPUESTOS

En la tabla 6.11 se muestran los porcentajes de discriminación entre pares de lenguajes utilizando los dos métodos propuestos en este trabajo de investigación, con muestras de señal de voz de 50 segundos. En el lado derecho de cada binario se encuentran los resultados con el método de inclusión de características suprasegmentales utilizando MFCC y los resultados remarcados en negro (de lado izquierdo) son con el método de una nueva caracterización del ritmo utilizando Wavelet. Podemos observar que los mejores resultados son cuando extraemos las características del ritmo por medio de Wavelet. Recordemos que éste método es sensible al tamaño de muestra y cuanto más grande sean los tamaños de muestras, los resultados mejoran. En cambio, el método de inclusión de características suprasegmentales por medio de MFCC tiene el efecto contrario, es decir, cuando las muestras son pequeñas los resultados mejoran. Por ello, en la tabla 6.12 podemos observar que de los 36 clasificadores binarios uno de ellos, inglés-tamil, obtuvo mejor resultado el método utilizando MFCC que el de Wavelet. Cabe mencionar que los tamaños de muestras en la tabla 6.12 son diferentes, pero como en este caso, para el método de inclusión de características suprasegmentales al tener tamaño de muestras más pequeños mejora el resultado, no afecta que la muestra no sea de 10 segundos, al contrario los resultados son mejores. Y aún así, el método de una nueva caracterización del ritmo utilizando Wavelet lo supera. En todos los 36 clasificadores binarios los resultados son mejores a excepción del par inglés-farsi.

	Alemán		Español		Mandarín		Vietnamita		Japonés		Coreano		Tamil		Farsi	
Inglés	**97**	78	**97**	85	**93**	75	**94**	76	**96**	77	**95**	74	**99**	86	**96**	81
Alemán	-		**93**	71	**94**	79	**93**	75	**98**	66	**98**	67	**94**	77	**91**	57
Español	-		-		**91**	74	**86**	70	**92**	69	**98**	74	**91**	66	**94**	62
Mandarín	-		-		-		**95**	72	**95**	77	**93**	66	**89**	84	**94**	75
Vietnamita	-		-		-		-		**93**	70	**96**	66	**95**	68	**95**	77
Japonés	-		-		-		-		-		**93**	68	**89**	66	**94**	72
Coreano	-		-		-		-		-		-		**95**	75	**91**	65
Tamil	-		-		-		-		-		-		-		**90**	75

Tabla 6.11 Comparativo de los porcentajes de discriminación de los dos métodos propuestos, Wavelet y MFCC (del lado derecho), con muestras de señal de voz de 50 segundos.

	Alemán		Español		Mandarín		Vietnamita		Japonés		Coreano		Tamil		Farsi	
Inglés	**94**	85	**96**	**83**	**85**	67	**88**	81	**94**	79	**83**	76	**98**	85	83	**86**
Alemán	-		**80**	**71**	**83**	83	**87**	85	**84**	69	**83**	70	**80**	77	**82**	68
Español	-		-		**90**	83	**84**	71	**86**	70	**88**	63	**87**	54	**79**	64
Mandarín	-		-		-		**85**	80	**89**	73	**83**	65	**85**	79	**94**	75
Vietnamita	-		-		-		-		**85**	72	**84**	68	**83**	63	**86**	72
Japonés	-		-		-		-		-		**83**	62	**75**	67	**89**	61
Coreano	-		-		-		-		-		-		**86**	66	**87**	65
Tamil	-		-		-		-		-		-		-		**86**	66

Tabla 6.12 Comparativo de los porcentajes de discriminación de los dos métodos propuestos, Wavelet y MFCC (de lado derecho), con muestras de señal de voz de 10 y 7 segundos respectivamente.

En la gráfica 6.15 se muestran los promedios por idioma, a modo de tener una idea global del comportamiento de cada uno de los métodos propuestos. Se muestra también los resultados de Cummins et al [15] por idioma como método a comparar del estado del arte y los dos métodos propuestos en este trabajo de investigación. En general de la gráfica 6.15 podemos observar que el método de una nueva caracterización del ritmo utilizando Wavelet obtiene mejores resultados que el método de inclusión de características suprasegmentales utilizando MFCC. Reafirmando que los dos métodos son sensibles al tamaño de la muestra de señal de voz. Los resultados con el método MFCC y 50 segundos de muestra, superan a Cummins y a su vez, el método Wavelet con muestras de 50 segundos supera a MFCC. Pero cuando la muestra es pequeña, de 7 y 10 segundos, el método de inclusión de características suprasegmentales utilizando MFCC supera al método del ritmo basado en Wavelet para el idioma español. Y los resultados de los dos métodos con muestras de 7 y 10 segundos son muy similares.

En la gráfica 6.16 se muestra un comparativo de los dos métodos propuestos y Rouas et al [22] de los promedios por idioma.

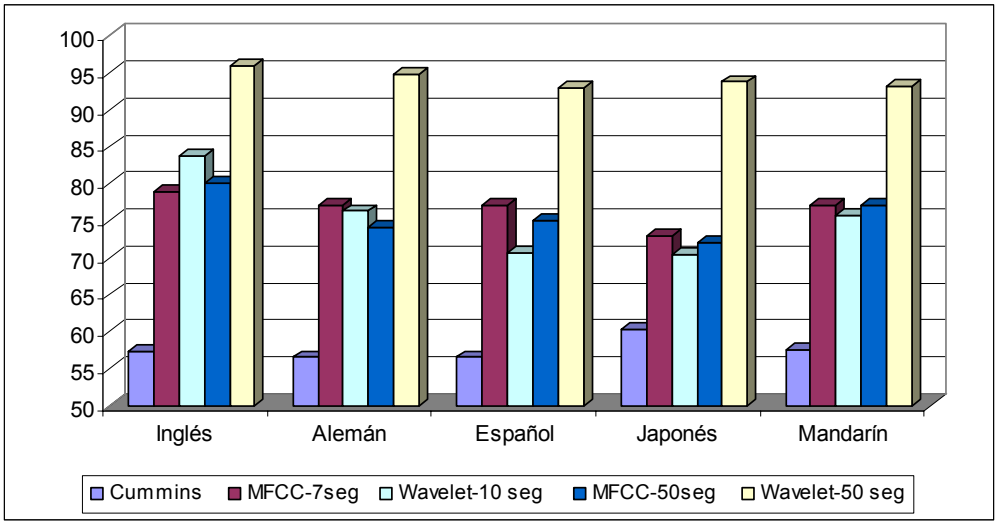

Figura 6.15 Comparativo del promedio de cada uno de los idiomas utilizando los dos métodos propuestos contra Cummins et al [15].

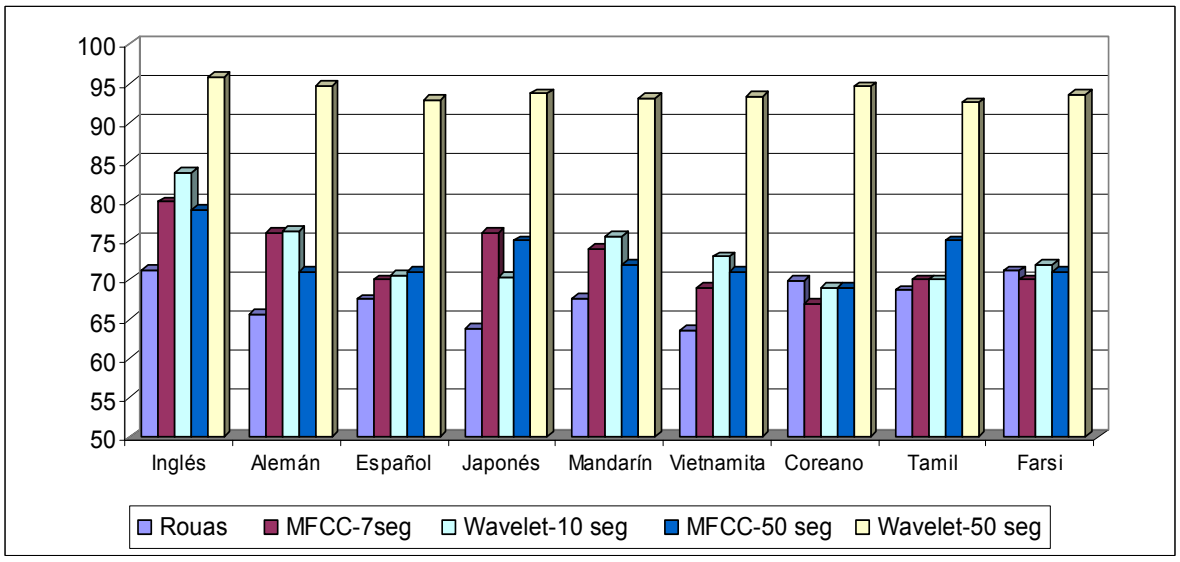

Figura 6.16 Comparativo del promedio de cada uno de los idiomas utilizando los dos métodos propuestos contra Rouas et al [15].

6.4 CONCLUSIONES

Los métodos propuestos en este capítulo demuestran los beneficios del uso de la transformada wavelet en la extracción de características acústicas de la señal de voz en la tarea de identificación del lenguaje hablado. El método propuesto permite el tratamiento de todos los lenguajes, incluyendo los lenguajes marginados, sin depender de ningún tipo de información lingüística. Los resultados obtenidos son muy buenos porque están arriba de otros métodos que no utilizan información lingüística, es decir, técnicas de extracción que sólo usan las características acústicas para identificar los idiomas.

Adicionalmente, los experimentos fueron hechos con diferentes técnicas de clasificación demostrando la pertinencia de la caracterización de la señal de voz, porque ellos confirman que los resultados son consecuencia de esta caracterización y no sólo por el efecto de la selección de un clasificador en especial.

Tenemos las siguientes observaciones importantes a notar. Primero, respecto al umbral del truncado de aproximación, éste se obtuvo al procesar todos los idiomas, pero se podría determinar por pares de idiomas, es decir, de acuerdo a cada par de idiomas se podría determinar un umbral específico. Segundo, el método propuesto es sensible al tamaño de la muestra, para muestras de señal de voz de tamaño grande, los porcentajes de discriminación mejoran. Esta observación indica que es necesario mejorar, aún más, la caracterización de la señal de voz para trabajar con muestras de señal de voz pequeñas.

En general podemos concluir, que el uso de las wavelet para el procesamiento acústico de la señal de voz alcanza resultados excelentes en la identificación del lenguaje, ya que la transformada wavelet permite enfocarnos en las bajas frecuencias que es donde se encuentran el ritmo. Con estos resultados nos estamos acercando a los porcentajes de discriminación que obtienen los sistemas que utilizan representación fonética, que como hemos dicho anteriormente son los que mejores resultados han obtenido hasta ahora.

CAPÍTULO 7

CONCLUSIONES GENERALES Y TRABAJO FUTURO

En general la identificación automática del lenguaje hablado es el proceso por el cual el lenguaje de una muestra de señal de voz digitalizada es reconocido por una computadora, sin considerar al hablante y lo que está diciendo. Hemos visto la utilidad de la identificación del lenguaje hablado en diferentes aplicaciones, agrupadas en dos grandes categorías: pre-procesamiento para sistemas, por ejemplo, en un sistema de traducción multilingüe; y pre-procesamiento para humanos, por ejemplo, en el caso de asistencia legal o médica de personas monolingües. Este último de gran importancia en México.

Como mencionamos anteriormente existen muchos trabajos en la identificación automática del lenguaje hablado, que resumimos en dos grandes grupos: los que se basan en la representación fonética de la señal de voz, es decir, los que segmentan la señal de voz en fonemas y emplean los modelos de lenguajes particulares de cada idioma, para poder distinguir entre los idiomas; y los que no utilizan representación fonética, los cuales se basan en las características suprasegmentales de la señal de voz, tales como la

prosodia, el ritmo, etc. Vimos también que los problemas del primer enfoque, son la segmentación de la señal de voz en fonemas, para lo que se necesitan grandes cantidades datos (grabaciones) previamente etiquetados. Por otro lado, para construir el modelo de lenguaje es necesario recopilar grandes cantidades de texto y voz. Con estos datos se calculan las probabilidades de las diversas secuencias de los fonemas. A pesar de que este tipo de sistemas alcanzan los mejores resultados, este enfoque no es de utilidad para lenguas que no cuentan con trascripción a texto (ni trascripción fonética), es decir las lenguas marginadas. Como es el caso de muchas de las lenguas indígenas de México.

Por lo anterior, el presente trabajo de investigación fue orientado a la identificación del lenguaje hablado sin representación fonética, es decir, sin depender de ningún recurso lingüístico. Para ello nos basamos en las características suprasegmentales de la señal de voz, tales como la prosodia, la entonación y el ritmo. Las características suprasegmentales son particulares de cada idioma y se encuentran en la señal de voz. Dichas características, de acuerdo a los lingüistas, no se deben separar a nivel de fonemas, porque tanto el tono, el ritmo y la duración van ligadas a más de un fonema. Cuando un fonema se une a otro para formar una sílaba, generan características muy distintivas entre los idiomas. Por lo tanto, se debe tomar la voz como un todo, y evitar su segmentación en fonemas.

El problema radica en cómo extraer las características suprasegmentales, que de acuerdo al estado del arte, la frecuencia fundamental F0, ha sido la más usada. Por ejemplo los trabajos de: Itahashi et al [13][58] en 1994 y 1995, después, Thyme-Gobbel y Hutchins [14] en 1996, así como Cummins et al [15] en 1999. En conclusión estos trabajos probaron que los parámetros prosódicos, en especifico la frecuencia fundamental F0, pueden ser utilizados en la discriminación entre una lengua y otra.

Rouas et al en 2003 [22] y 2005 [36] obtuvo otro conjunto de características, basándose en los trabajos previos de Ramus et al [33], el cual consiste en identificar los lenguajes en base a su entonación y ritmo, por medio de la caracterización del ritmo en función de intervalos vocálicos y consonánticos. Los lingüistas agrupan los lenguajes de acuerdo a su ritmo, en tres grandes grupos: *stress-timed, syllable-timed* y *mora-timed.* El inglés y alemán pertenecen al primer grupo y el español al segundo. El japonés pertenece al *mora-timed*. La entonación separa al chino-mandarín y vietnamita del resto, ya que

utilizan la entonación para diferenciar entre palabras con diferente significado. Su modelo parte de segmentar la señal de voz en intervalos formados por vocales e intervalos formados por consonantes, para obtener los parámetros del ritmo.

7.1 CONCLUSIONES

En este trabajo de investigación doctoral, se obtuvieron dos nuevos métodos de extracción de características específicos para la identificación del lenguaje hablado, sin utilizar la representación fonética de la señal de voz, los dos están basados en las características suprasegmentales, distintivas entre los idiomas.

La idea central del *primer método* fue la *inclusión de información suprasegmental*, obteniendo una nueva caracterización basándonos en la existencia de los elementos suprasegmentales de los fonemas, como la prosodia, la entonación y la duración. Para ello nos basamos en el procesamiento de la señal de voz por medio de la transformada de Fourier, específicamente los coeficientes cepstrales de frecuencia Mel (MFCC). Todos los trabajos anteriores han utilizado sólo la frecuencia fundamental F0. Nosotros propusimos el uso de los cepstrales MFCC, con 16 coeficientes, capturando además de la frecuencia fundamental F0, frecuencias secundarias que pueden ser importantes en la tarea de discriminar lenguajes. Además propusimos el uso de los deltas Δ_1, Δ_2 y Δ_3, los cuales pretenden capturar el cambio de un coeficiente cepstral entre una ventana y su adyacente. Con estos deltas, nosotros buscamos capturar las diferencias que hay entre los fonemas y posiblemente entre sílabas. Obteniendo un nuevo método de caracterización de la señal de voz, con 192 atributos. Los resultados son comparables con el estado del arte en sistemas que no utilizan representación fonética. Superando a los tres trabajos, vistos en el estado del arte, Itahashi et al [13][58] en 1994 y 1995, Thyme-Gobbel y Hutchins [14] en 1996, que sólo utilizan la frecuencia fundamental F0, y en específico al trabajo de Cummins et al [15] reportado en 1999. De la comparación con el trabajo de Rouas reportado en 2003 [22] y 2005 [36], podemos concluir que superamos la mayoría de sus

resultados, pues ellos tuvieron problemas en la identificación de la pareja de inglés y alemán, que son idiomas con ritmos semejantes (pertenecen al grupo de stress-timed). En nuestro caso los resultados entre esos dos lenguajes (inglés y alemán) están por encima de los reportados por Rouas. Cabe mencionar que el método de Rouas, que consiste en el uso de las características de entonación y cantidad de intervalos vocálicos y consonánticos, obtuvo buenos resultados, como ellos lo mencionan, en la discriminación entre grupos de lenguajes stress-timed contra los que utilizan la entonación como un marcador léxico (el chino-mandarín y vietnamita). En nuestro caso es interesante notar los bajos resultados en el caso del coreano. Para el japonés que es un lenguaje *mora-timed*, pudimos discriminarlo bien contra los *stress-timed*.

Otra observación que tenemos es que con muestras pequeñas de señal de voz se obtuvieron mejores resultados. Desafortunadamente, este es un punto en contra de esta caracterización pues como pudo observarse los promedios tienden a estabilizarse con muestras de señal de voz más grandes. Probablemente por la mayor aparición de pausas y silencios.

La idea central del *segundo método*, fue obtener una nueva *caracterización orientada a las bajas frecuencias*. Recordemos que la frecuencia fundamental es la más baja de todas y fue el parámetro más utilizado por los métodos del estado del arte para representar la prosodia. Por lo tanto podemos asumir que en las frecuencias bajas hay información relevante para la identificación del lenguaje hablado; las cuales podrían representar al ritmo, la entonación y la duración –en general las características suprasegmentales que usamos al hablar –. Por lo que propusimos el uso de la transformada wavelet para el procesamiento de la señal de voz; ya que la transformada wavelet tiene una muy buena resolución en las bajas frecuencias. Haciendo una separación entre las altas y bajas frecuencias. Este método es completamente diferente a los usados anteriormente basados en la transformada de Fourier.

La diferencia principal contra el primer método propuesto es que el ventaneo de la STFT es fijo en contra del ventaneo de la wavelet que es variable. Recordando lo visto, tenemos que el ancho de ventana constituye un parámetro de gran importancia ya que a través de éste podemos establecer el grado de resolución tanto de tiempo como de

frecuencia. Si la ventana es muy pequeña tenemos una buena resolución en tiempo pero una mala resolución en frecuencia y por el contrario, si la ventana es muy grande tendremos una buena resolución en frecuencia pero una mala resolución en tiempo. En otras palabras, para tener una buena resolución de las altas frecuencias necesitamos ventanas pequeñas y para una buena resolución de las bajas frecuencias necesitamos ventanas grandes. Esto es importante, ya que deseamos capturar la información suprasegmental del habla. La voz es una señal cuya amplitud varía en forma rápida y abrupta en el tiempo, además su contenido de frecuencias es variable de un instante de tiempo a otro, es decir, es una señal no estacionaria y cuya información acerca de la prosodia, el ritmo, la duración y la entonación están ligados a las frecuencias bajas.

Los resultados superaron a los trabajos reportados por Cummins et al [15] en 1999 y a Rouas et al en 2003 [22] y 2005 [36], demostrando los beneficios del uso de la transformada wavelet en la extracción de características acústicas de la señal de voz en la tarea de identificación del lenguaje hablado.

Los métodos permiten el tratamiento de cualquier lenguaje, incluyendo los lenguajes marginados, ya que dichos métodos no dependen de ningún tipo de información lingüística. Lo que nos abre las puertas hacia un método de identificación automática para las lenguas indígenas de México (el apéndice A muestra un primer intento en este sentido).

Con estos resultados nos estamos acercando a los porcentajes de discriminación que obtienen los sistemas que utilizan representación fonética, que como hemos dicho anteriormente son los que mejores resultados han obtenido hasta ahora.

Los dos nuevos métodos fueron probados con diferentes técnicas de clasificación demostrando la pertinencia de la caracterización de la señal de voz, confirmando que los resultados son consecuencia de esta caracterización y no solo por el efecto de la selección de un clasificador en especial.

De los dos métodos propuestos podemos observar que los mejores resultados son cuando extraemos las características del ritmo por medio de Wavelet. Las diferencias a considerar es que los dos métodos son sensibles al tamaño de muestra de la señal de voz,

para el caso de la extracción por medio de wavelet necesitamos muestras grandes alrededor de 50 segundos para obtener muy buenos resultados. No así para el caso de la extracción de características suprasegmentales por medio de MFCC que entre más pequeñas las muestras los resultados son mejores.

7.2 APORTACIONES

El trabajo de tesis doctoral aportó:

1. Dos métodos de identificación del lenguaje hablado sin utilizar reconocimiento fonético, extrayendo las características suprasegmentales del habla:
    1. Uno basado en el uso de los coeficientes cepstrales de frecuencia Mel, llamado inclusión de características suprasegmentales.
    2. El segundo basado en el uso de la transformada wavelet; extrayendo las frecuencias bajas con muy buena resolución, que es donde se encuentra el ritmo.
2. Abrir las puertas hacia un método de identificación automática para las lenguas indígenas de México.

7.3 TRABAJO FUTURO

Como trabajo futuro, planeamos extender nuestro método para trabajar con clasificadores multiclase (recuerde que los resultados reportados corresponden a conjuntos de clasificadores de pares de idiomas –binarios–). Estas modificaciones podrían permitirnos comparar nuestra técnica con otros métodos, incluyendo aquellos que están basados en el uso de la información fonotáctica de los lenguajes. Este es el caso del trabajo de Caseiro et al [5], el cual reporta un 79.6% de exactitud para la discriminación de 6 lenguajes usando muestras de señal de voz de 10 segundos.

Otro trabajo interesante, sería mezclar diferentes extracciones de características acústicas de la señal de voz, tales como el ritmo como en [22][36], con los coeficientes wavelet y generar características del habla híbridas, que nos ayuden a capturar de una forma más completa las características suprasegmentales del habla.

Por el lado de los clasificadores, sería interesante utilizar los modelos de mezclas gaussianas (GMM), muy usadas en la identificación del locutor. Aplicadas sobre los dos métodos de caracterización propuestos en este trabajo. Por otro lado, el uso de este tipo de clasificador nos permitiría probar nuestras ideas en la tarea de verificación del idioma. Con ello estaríamos en posición de compararnos con los trabajos reportados en el NIST (The National Institute of Standards and Technology).

Otro trabajo sería la aplicación del método usando wavelet para la identificación de acentos regionales. Un primer experimento podría realizarse con el corpus TIMIT, el cual divide al inglés americano en cinco regiones. Posteriormente, dependiendo de los resultados alcanzados se podría intentar construir un corpus para el español.

También sería interesante comprobar el alcance de los métodos para la tarea de identificación del habla no nativa, con un corpus en donde la persona no hable su lengua materna, por ejemplo, una persona cuya lengua materna sea el español, realice una grabación hablando inglés, y a partir de esas muestras de señal de voz identificar el idioma hablado. Recordemos lo dicho por Abercombrie [12]: el “ritmo” es probablemente el rasgo de la *base articulatoria* de una lengua cuya adquisición o dominio resulta más difícil al estudiante adulto de un idioma extranjero y, aunque la inteligibilidad depende en gran parte de su correcta emisión, a éste no se le presta la atención debida en la enseñanza de idiomas extranjeros. Esto es, cuando un adulto aprende un idioma extranjero, es necesario invertir un enorme esfuerzo para adquirir el ritmo de la lengua. Aunque pronunciemos correctamente una palabra, las pautas rítmicas que emitimos al pronunciarla no corresponden al idioma. El problema para este tipo de pruebas, es que no existen corpus para ello. Por lo que un trabajo futuro sería construir ese tipo de corpus.

LISTA DE PUBLICACIONES

- A.L. Reyes, C. Reyes: “Estado del arte en el procesamiento de la voz”, IV Simposium Internacional en Tecnologías Inteligentes, Instituto Tecnológico de Apizaco, Pue. Páginas 67-76, (2004).

- A.L. Reyes, L. Villaseñor: “Identificación Automática del Lenguaje Hablado”, Sexto encuentro de investigación, INAOE, Tonantzintla, Pue. Páginas 303-306, (2005)

- A.L. Reyes, L. Villaseñor, M. Montes: “A Straightforward Method for Automatic Identification of Marginalized Languages”, FinTAL 2006, Turku, Finlandia. LNAI 4139, pp. 68-75, Springer-Verlag (2006).

- A.L. Reyes, L. Villaseñor, M. Montes: “Automatic Language Identification using Wavelets”, The Ninth International Conference on Spoken Language Processing, Interspeech 2006, pp. 401-404, Pittsburgh, USA. (2006).

- A.L. Reyes, L. Villaseñor: “Hacia la Identificación Automática de Lenguas Indígenas de México: Náhuatl y Zoque de Oaxaca”, Séptimo encuentro de investigación, INAOE, Tonantzintla, Pue. Páginas 285-288, (2006)

PREMIO OBTENIDO

Se obtuvo el primer lugar en la segunda Convención Nacional de Investigación Aplicada y Desarrollo Tecnológico 2006. Con el trabajo “Identificación Automática de Lenguas sin Transcripción Fonética: Náhuatl y Zoque de México”, concursando en el área de Informática Nivel C: Postgrados. El evento se celebró en la Ciudad de Puebla en la Universidad Tecnológica de Puebla. Este evento fue organizado por el Consejo de Ciencia y Tecnología del Estado de Puebla, la Secretaría de Educación Pública Federal a través de la Subsecretaría de Educación Superior, la Universidad Tecnológica de Puebla, el Instituto Mexicano de la Propiedad Industrial, la Benemérita Universidad Autónoma de Puebla, el Instituto Tecnológico de Puebla, la Universidad Politécnica de Puebla, la Universidad Juárez Autónoma de Tabasco y el Foro Consultivo Poblano para la innovación y el Desarrollo Tecnológico.

REFERENCIAS

[1] Ventura Morales Santiago (2000). “Condena injusta por prejuicio cultural”. Boletín mensual del FIOB, El Tequio, enero del 2000, existe también página web: http://www.laneta.apc.org/fiob/condena.html

[2] De León Lourdes (1999) “<<Si>> means <<Yes>> Lenguaje y poder en el juicio de un mixteco en EUA”, en Gabriela Vargas Cetina, ¿Mirando…hacia fuera?: experiencias de investigación, México, CIESAS, 1999.

[3] Davis Alex, (2002) “Unusual Woddburn office helps Indigenous Mexicans”. Associated Press Newswires, 28 de octubre de 2002.

[4] Yan Y. (1995):”Language identification based on language-dependent phone recognition”, PhD. Thesis, Oregon graduate institute of science and technology, USA.

[5] Casseiro D., Trancoso I., (1998): “Language Identification Using Minimum Linguistic Information”, in 10th Portuguese Conference on Pattern Recognition (RECPAD’98), Lisbon Portugal.

[6] Andersen O., Dalsgaard P. (1997): “Language Identification based on Cross-Language Acoustic models and Optimized Information Combination”, In EUROSPEECH-1997, 67-70.

[7] Rodríguez C. (2003):”Proyecto multidisciplinario de asistencia monolingüe”. www.iling.unam.mx/langid/.

[8] Torres-Carrasquillo, P. A., Reynolds, D. A., and Deller J. R. (2002): “Language Identification using Gaussian Mixture Model Tokenization. In Proc. International Conference on Acoustics, Speech, and Signal Processing in Orlando, FL, IEEE, pp. I: 757-760, 13-17 May 2002.

[9] Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., Deller J. R. (2002): “Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features. In Proc. International Conference on Spoken Language Processing in Denver, CO, ISCA, pp. 33-36, 82-92 September 2002.

[10] Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D.A. (2003): “Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Recognition”. In Proc. Eurospeech in Geneva, Switzerland, ISCA, pp. 1345-1348, 1-4 September 2003.

[11] Alcaraz-Varó E., Martínez-Linares M.A., (1997): *Diccionario de Lingüística Moderna*. Editorial Ariel, S.A. Córcega, 270-08008 Barcelona, España. ISBN: 84-344-0510-5.

[12] Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh University Press.

[13] Itahashi, S. Du, L. (1995) "Language identification based on speech fundamental frequency". In: Eurospeech, Vol. 2, pp. 1359-1362.

[14] Thyme-Gobbel A.E., Hutchins S.E. (1996). "On using prosodic cues in automatic language identification". In: International Conference on Spoken Language Processing, vol. 3, pp. 1768-1772.

[15] Cummins F., Gers F., Schmidhuber J., (1999): "Language Identification from Prosody without explicit Features", Proc. EUROSPEECH'99, Budapest, Hungary, 1, pp. 371-374.

[16] Muthusamy Y.K., Barnard E., Cole R.A., (1994). "Reviewing automatic language identification". IEEE Signal Process. Mag. 11 (4), pp. 33-41.

[17] Huang X., Acero A., Hon H-W., (2001): "Spoken Language Processing", Microsoft Research, Prentice Hall.

[18] Modic R., Lindberg B., Petek B. (2003): "Comparative wavelet and MFCC speech recognition experiments on the Slovenian and English speechDat2", in NOLISP-2003, Le Croisic, France, paper 016.

[19] Gupta M., Gilbert A., (2001): "Robust speech recognition using wavelet coefficient features", in IEEE Automatic Speech Recognition and Understanding Workshop, USA, pp. 445-448.

[20] Daubechies I., (1992): *Ten lectures on Wavelets,* Vol. 61, SIAM Press, Philadelphia, PA USA, 1992.

[21] Muthusamy Y.K., Cole R., Oshika B., (1992):"The OGI multi-language telephone speech corpus". International Conference on Spoken Language Processing (ICSLP'92), volume 2, pp.895-998, Alberta, Canada.

[22] Rouas J-L., Farinas J., Pellegrino F., André-Obrecht R., (2003): "Modeling prosody for language identification on read and spontaneous speech" in Proc. IEEE ICASSP 2003, vol 1, pp. 40-43.

[23] Stein J.Y. (2000): *Digital Signal Processing: A computer Science Perspective*. Wiley Series in Telecommunications and Signal Processing. John G. Proakis, Series Editor. A Wiley-Interscience Publication. ISBN 0-471-29546-9.

[24] Jurasfsky D., Martin J.H., (2000): *Speech and Language Processing: An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artificial Intelligence. S. Russell and P. Norving, Editors. ISBN 0-131-095069-6.

[25] Burrus C.S., Gopinath R.A., and Guo H. (1998): *Introduction to wavelet and wavelet transform: A primer*. Simon & Schuster, A Viacom Company. Prentice Hall Upper Saddle River, New Jersey. ISBN 0-13-489600-9.

[26] Vetterli M. and Kovacevic J.: (1995): *Wavelets and subband coding*. Prentice Hall Signal Processing Series, New Jersey Prentice-Hall Inc. ISBN 0-13-097080-8.

[27] Mallat S. (1999): *A Wavelet Tour of Signal Processing*. Academic Press. Second edition. ISBN 0-12-466606-X.

[28] Quilis, A. (1992). *Tratado de fonología y fonética españolas*. Madrid: Gredos.

[29] Ladefoged, P.A. (1971). *Preliminaries to Linguistic Phonetics*. The University of Chicago Press.

[30] Herrero, A. (1988). *Semiótica y creatividad. La lógica abductiva*. Madrid: Palas Atenea.

[31] Cummins, F. (2002): "Classifying languages based on speech rhythm". In Artificial Intelligence and Cognitive Science: Proceedings of the 13th Irish International Conference (AICS 2002), volume 2464 of Lecture Notes in Computer Science. Springer Verlag.

[32] Nazzi T., Bertoncini J. and Mehler J. (1998):"Language discrimination by newborns; towards an understanding of the role of rhythm". Journal of Experimental Psychology: Human Perception and Performance, 24:756-766. APA (American Psychological Association).

[33] Ramus F., Nespor M., Mehler J., (1999):"Correlates of linguistic rhythm in the speech signal". Cognition, 73(3), pp. 265-293. Elsevier.

[34] Barry W.J., Andreeva B., Russo M., DImitrova S. and Kostadinova T. (2003): "DO rhythm measures tell us anything about language type?" In: Proceedings of the 15th ICPhS, Barcelona. 2639-2636. International Phonetic Alphabet (IPA).

[35] Dellwo V. (2003): "Rhytm & Speech Rate: A variation coefficient for deltaC". In: Proceedings of the 38 Linguistic Colloquium, Budapest. Elsevier.

[36] Rouas J.L., Farinas J., Pellegrino F. and André-Obrecht R., (2005): "Rhythmic unit extraction and modeling for automatic language identification". Journal Speech Communication, Volume 47, Issue 4, December 2005, Pages 436-456. Elsevier.

[37] Galves A., Garcia J., Duarte D. and Galves C. (2002): "Sonority as a basis for rhythmic class discrimination". Proceedings of Speech Prosody, Aix-en-Provence.

[38] Steiner I. (2003): “A refined acoustic analysis of speech rhythm”. Proceedings of the 38 Linguistic Colloquium, Budapest. Elsevier.

[39] Muthusamy Y.K., Jain N., Cole R., (1994): ”Perceptual benchmarks for automatic language identification”, in Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP’94), Adelaide, Australia. Vol. 1, pp. 333-336.

[40] Zissman M.A., Berkling K.M. (2001). “Automatic language identification”. Journal Speech Communication 35, pp. 115-124. Elsevier NH.

[41] Riek L., Mistretta W., Morgan D. (1991). “Experiments in language identification”, Technical Report SPCOT-91-002, Lockheed Sanders, Inc., Nashua, NH.

[42] Nakagawa S., Ueda Y., Seino T. (1992) “Speaker-independent, text-independent language identification by HMM”. In: International Conference on Spoken Language Processing, vol. 2, pp. 1011-1014.

[43] House, A.S., Neuburg, E.P. (1997) “Toward automatic identification of the language of an utterance. I. preliminary methodological considerations.” J. Acoust. Soc. AMER. 62 (3), 708-713.

[44] Savic M, Acosta E., Gupta S.K, (1991). “An Automatic language identification system”. In: International Conference on Acoustic, Speech, and Signal Processing, vol. 2, pp.817-820.

[45] Zissman M.A. (1993). “Automatic language identification using Gaussian mixture and hidden Markov models.” In: International Conference on Acoustic, Speech and Signal Processing, vol.2, pp. 399-402.

[46] Nakagawa S. Sieno T., Ueda Y. (1994). “Spoken language identification by ergodic HMMs and its state sequences”. Electron. Commune. Jpn. part 3 77 (6), 70-79.

[47] Lamel, L.F., Gauvain, J-L., (1993) “Cross-lingual experiments with phone recognition.” In: International Conference on Acoustic, Speech, and Signal Processing, Vol. 2, pp. 507-510.

[48] Muthusamy Y.K., et al (1993):”A comparison of approaches to automatic language identification using telephone speech”, in Proceedings of EUROSPEECH. Vol. 2, pp. 1307-1310.

[49] Damashek M, (1995):”Gauging similarity with n-grams: language-independent categorization of text. Science 267 (5199), pp. 843-848.

[50] SPEECHDAT (M): EU-project LRE-63314.

[51] Zissman M., Singer E., (1994): ”Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling”, in Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP’94), Vol. 1, pp. 305-308, Adelaide, Australia.

[52] Yan Y, Barnard E., (1995):”An approach to automatic language identification based on language-dependent phone recognition”. In International Conference on Acoustic, Speech and Signal Processing (ICASSP’95), Vol. 5, pp. 3511-3514.

[53] Campbell, W. M., Singer, E., Torres-Carrasquillo, P. A., Reynolds, D. A. (2004): “Language Recognition with Support Vector Machines. In Proc. Odyssey: The Speaker and Language Recognition Workshop in Toledo, Spain, ISCA, pp. 41-44, 31 May - 3 June 2004.

[54] Kadambe S, Hieronymus J. (1995). “Language Identification with phonological and lexical models”. International Conference on Acoustics, Speech and Signal Processing, Vol. 5, pp. 3507-3511.

[55] Mendoza S., et al (1996). “Automatic language identification using large vocabulary continuous speech recognition”. International Conference on Acoustics, Speech and Signal Processing, Vol. 2, pp. 785-788.

[56] Berkling, K.M., (1996):”Automatic language identification with sequences of language-independent phoneme clusters”, Ph.D. Thesis, Oregon graduate institute of science and technology. USA.

[57] Li, K-P., (1994) “Automatic language identification using syllabic spectral features.” In: International Conference on Acoustic, Speech, and Signal Processing, Vol. 1, pp. 297-300.

[58] Itahashi, S., Zhou, J., Tanaka, K. (1994). “Spoken language discrimination using speech fundamental frequency.” In: International Conference on Spoken Language Processing, vol 4, pp. 1899-1902.

[59] Samouelian A., (1996): “Automatic Language Identification using Inductive Inference”, in 4th International Conference on Spoken Language Processing (ICSLP 96), Philadelphia, USA.

[60] Platt J., (1998): "Fast Training of Support Vector Machines using Sequential Minimal Optimization". Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf, C. Burges, and A. Smola, eds., MIT Press.

[61] Witten I., Eibe F., (2000): *Weka (Waikato Environment for Knowledge Analysis).Data Mining, Practical Machine Learning Tools and Techniques with java Implementations*, Academic Press, USA. 2000.

[62] Boersma P., Weenink D., Praat v 4.2.08. (2004): “A system for doing phonetics by computer”. Institute of Phonetic Sciences of the University of Amsterdam. June 2004.

[63] Mitchell, T.M. (1997): *Machine Learning*. McGraw-Hill Companies, Inc. ISBN 0070428077.

[64] Boersma P., Weenink D.,(2002): "Praat v 4.0.5. A system for doing phonetics by computer". Institute of phonetic Sciences of the University of Amsterdam.

[65] Press W., Vetterling W., Teukolsky S., Flanery B. (2002): *Numerical recipes in C++: The Art of Scientific Computing*. Cambridge University Press. 2002. Second edition. ISBN 0-524-75033-4.

[66] INEGI. (2005): "Indicadores sociodemográficos de México 2004" http://cuentame.inegi.gob.mx/poblacion/

[67] AILLA. (2005): "Archivo de los idiomas indígenas de Latinoamérica", grabaciones de H. Johnson. http://www.ailla.utexas.org/site/welcome_sp.html

APÉNDICE A

LA IDENTIFICACIÓN AUTOMATICA DE LENGUAS SIN TRANSCRIPCION FONÉTICA: NÁHUATL Y ZOQUE DE MÉXICO.

Actualmente en México 6 de cada 100 habitantes (de 5 años y más) hablan alguna lengua indígena, lo que representa el 7.3% de la población total de México [66]. La mayoría de ellos son indígenas monolingües, los cuales tienen la necesidad de interactuar con médicos y autoridades tanto en México como en EE. UU, por casos de emergencias médicas o simplemente cuando soliciten información. Por otro lado, la gran mayoría de las aproximadamente 69 lenguas indígenas en México no tienen trascripción a texto. Por ejemplo, el Náhuatl tiene transcripción fonética a partir de idioma español. Y algunas lenguas incluso ya están en vías de extinción. Por los problemas que acarrea el hablar dicha lengua, los indígenas prefieren aprender y enseñarles a sus hijos el español y no su lengua materna.

El único sistema que existe hasta hoy, para identificar una lengua indígena mexicana, es “Que Lenguas Hablas” [http://cdi.gob.mx/ini/lenguahablas/]. En las figuras A.1 y A.2 se muestra las pantallas de presentación del sistema, el cual solo contiene 39 lenguas de las 69 que existen en todo México.

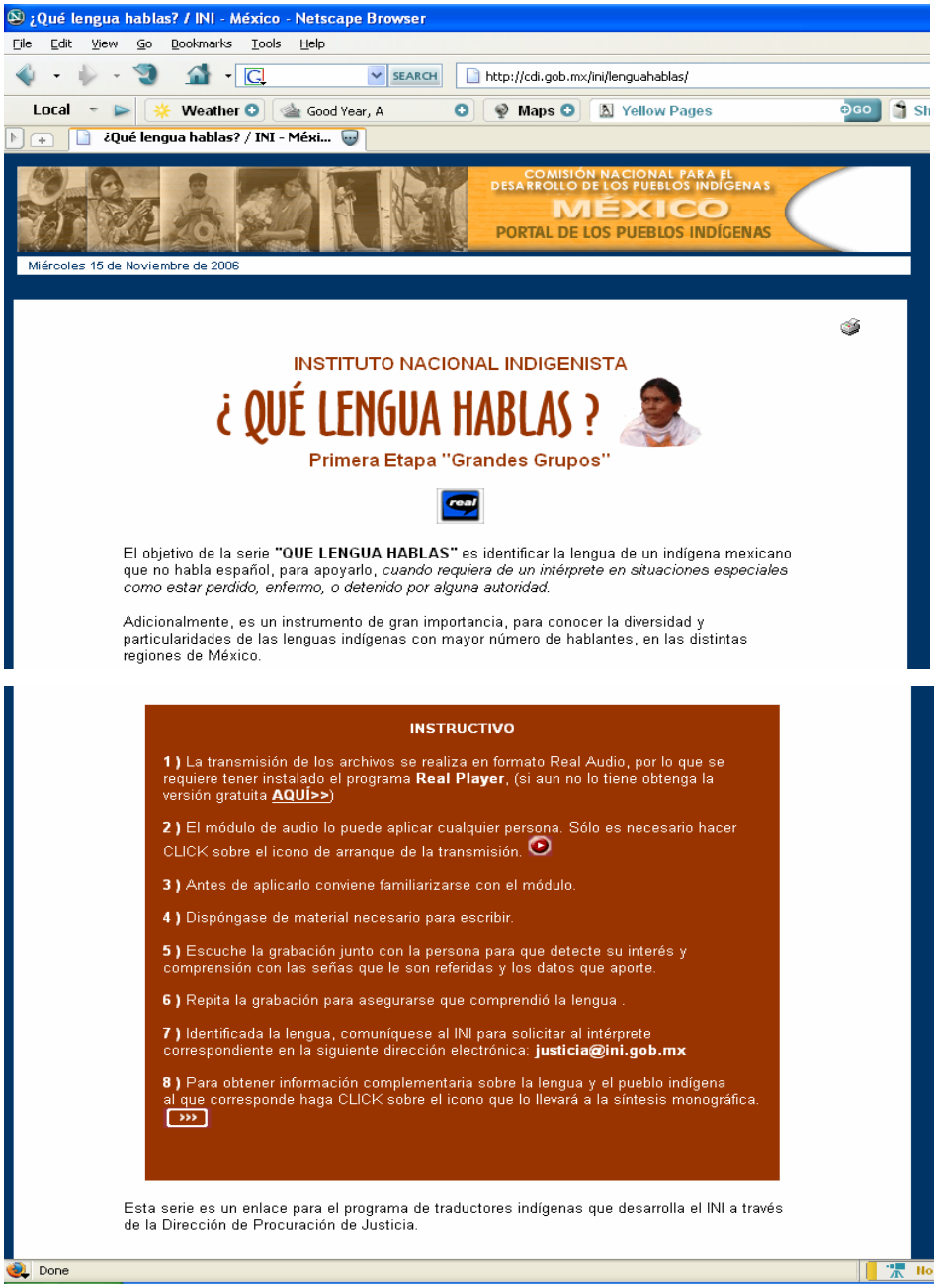

Figura A.1 Pantalla de presentación del sistema ¿Qué lengua hablas?

Figura A.2 Continuación de la pantalla de presentación del sistema ¿Qué lengua hablas?

El sistema “¿Que lengua Hablas?” identifica la lengua indígena en forma manual, es decir, se hace escuchar a la persona monolingüe todas las posibles grabaciones de lenguas que hay en el sistema, que son alrededor de 39 divididas en cinco grandes grupos: Mayas, Oaxaca, Nahuas, Región centro y Región Norte; y hasta que entienda la frase que se le dice, entonces tenemos identificado que lengua indígena habla. Este proceso es lento, cansado y en ocasiones inoperante en casos de emergencias. Por ello, existe la necesidad de crear un sistema automático de identificación de lenguas indígenas de México. Permitiendo su uso remoto mediante teléfono o un portal Internet. Dicho sistema permitirá a las autoridades la identificación automática de la lengua del hablante monolingüe, y proporcionará intérpretes adecuados y una lista de instrucciones y derechos fundamentales.

Como mencionamos anteriormente, los sistemas con mejores resultados, son los que basan la identificación del lenguaje hablado en el empleo de las características lingüísticas propias de cada lenguaje. Desafortunadamente, para las lenguas que no tienen trascripción a texto, como muchas de las lenguas indígenas de México; este enfoque no es de utilidad.

Uno de los objetivos de este trabajo de tesis, fue proponer métodos de identificación automática del lenguaje hablado sin representación fonética, con el fin de poder ser utilizados en lenguas que no tengan transcripción fonética. Como las lenguas indígenas de México. De este trabajo se obtuvieron dos nuevos métodos, los cuales fueron aplicados a dos lenguas indígenas de México: Náhuatl y Zoque de Oaxaca.

Se tomaron el Náhuatl y el Zoque de Oaxaca porque una de ellas es la más representativa de México. El Náhuatl con el 24% de hablantes de lenguas indígenas en todo el país y el Zoque de Oaxaca con el 1% de hablantes [66]. Además se tomó el idioma español, el cual es el idioma oficial en México. Las grabaciones de Zoque de Oaxaca y Náhuatl fueron tomadas del corpus AILLA (Archivo de los idiomas indígenas de Latinoamérica [67], y el español fue tomado del corpus OGI_TS [21]. Se tomaron de cada una de las lenguas indígenas y el idioma español 20 hablantes diferentes, sin restricciones en la forma de hablar y que decir; con co-articulación y pausas. Las grabaciones de las lenguas indígenas están hechas a 44kHz, por lo que fueron re-muestreadas a 8kHz, y las

del idioma español fueron hechas vía telefónica, es decir, a 8Khz; ya que el corpus fue hecho en base a llamadas telefónicas. En total se obtienen 60 muestras de señal de voz hablada, con diferentes duraciones 3, 7 y 10 segundos. A continuación se describen los resultados obtenidos.

A.1 CARACTERÍSTICAS SUPRASEGMENTALES

La caracterización de la señal y el proceso de aprendizaje fueron iguales a los descritos en la sección 5.3. Para la evaluación se utilizó la validación cruzada con 10 pliegues; también descrito en el capítulo cinco. El primer experimento se realizó con datos sin aplicar ninguna técnica de reducción de dimensionalidad, el segundo aplicando ganancia de información. Los dos experimentos fueron realizados con los 3 diferentes tamaños de muestras de señal de voz de 3, 7 y 10 segundos, generados por el procesamiento acústico de la señal de voz, previamente descrito. Los resultados se muestran en las tablas A.1 y A.2. Utilizando el clasificador Naïve-Bayes. Estos dos experimentos fueron hechos para pares de idiomas. Posteriormente se realizaron pruebas para los tres idiomas juntos, es decir, multiclase. Los resultados se muestran en la tabla A.3. De la cual podemos observar que el mejor resultado es para muestras de 7 segundos, con un 93% de exactitud en la identificación de los tres lenguajes.

	3 segundos		7 segundos		10 segundos	
	Náhuatl	Español	Náhuatl	Español	Náhuatl	Español
Zoque	85	95	79	93	84	95
Náhuatl	-	94	-	93	-	94

Tabla A.1 Porcentaje de discriminación entre las lenguas indígenas: Náhuatl y Zoque de Oaxaca y el español. Sin utilizar ganancia de información.

	3 segundos		7 segundos		10 segundos	
	Náhuatl	Español	Náhuatl	Español	Náhuatl	Español
Zoque	94	98	94	98	97	98
Náhuatl	-	97	-	97	-	94

Tabla A.2 Porcentaje de discriminación entre las lenguas indígenas: Náhuatl y Zoque de Oaxaca y el español. Utilizando ganancia de información.

Los tres idiomas	Sin ganancia de información	Con Ganancia de información
3 segundos	82	91
7 segundos	**87**	**93**
10 segundos	86	92

Tabla A.3 Comparativo de las tres lenguas de acuerdo a sus diferentes tamaños de muestras.

A.2 CARACTERÍSTICAS RITMICAS

Para este proceso, utilizamos la caracterización de la señal de voz utilizando la transformada wavelet, descrita en el capítulo seis. Por lo tanto, usamos Daubechies db2, con cuatro coeficientes. El porcentaje para el truncado de aproximación fue del 1%, es decir, el umbral fue .01. Con la construcción de matrices de pares de lenguajes. El corpus utilizado fue el mismo que en la sección anterior. Esto es, las lenguas de Náhuatl, Zoque de Oaxaca y español. Con 20 hablantes diferentes, para cada lengua. Teniendo un total de 60 hablantes diferentes. Fue habla espontánea. Las muestras de señal de voz fueron de 10, 30 y 50 segundos. Para el caso del uso de la transformada wavalet los resultados mejoran.

El clasificador utilizado fue Naïve Bayes, con el método de validación cruzada con 10 pliegues, utilizando ganancia de información para reducir la dimensionalidad. Los resultados se muestran en la tabla A.4, por pares de lenguajes, con diferentes tamaños de muestras de señal de voz. En la tabla A.5 se muestran los resultados de los tres idiomas, es decir, multiclase. Donde se muestra que para muestras de señal de voz más grandes el porcentaje de identificación es mayor, de forma similar a como lo que obtuvimos con los otros idiomas, del corpus OGI_TS [21].

	10 segundos		30 segundos		50 segundos	
	Náhuatl	Español	Náhuatl	Español	Náhuatl	Español
Zoque	90	95	87.5	95	92.5	100
Náhuatl	-	92.5	-	97.5	-	97.5

Tabla A.4 Porcentaje de discriminación entre las lenguas indígenas: Náhuatl, Zoque de Oaxaca y el español. Utilizado la transformada wavelet.

Los tres idiomas	Con Ganancia de información
10 segundos	85
30 segundos	94
50 segundos	**95**

Tabla A.5 Comparativo de las tres lenguas de acuerdo a sus diferentes tamaños de muestras.

A.3 CONCLUSIONES

Como puede observarse los resultados son muy alentadores y con esto podemos sentar las bases para un identificador automático de lenguas sin trascripción fonética para las lenguas indígenas de México, es claro que el corpus utilizado fue muy pequeño, y será necesario realizar experimentos más completos en cuanto contemos con corpus grandes de las lenguas indígenas de México. Los resultados son alentadores pero debemos

considerar las diferencias de las condiciones de grabación. Las grabaciones del corpus OGI_TS[21] son muy diferentes a las condiciones de grabación de corpus ALLIA [67] de las dos lenguas indígenas de México. En primer lugar el canal de grabación y las diferencias en el muestreo. Aún así, la discriminación entre las dos lenguas indígenas es buena. Habrá que realizar pruebas con otras lenguas, pero nos enfrentamos a que no existen corpus de lenguas indígenas de México. Esto podría generar un trabajo futuro muy interesante y en conjunto con otras áreas, en específico con los lingüistas, para generar primero un corpus de las lenguas indígenas de México.