UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
INSTITUTO DE INGENIERÍA
MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA

"SISTEMA DE IDENTIFICACIÓN DE LOCUTOR EN EL CONTEXTO DEL ESPAÑOL MEXICANO"

TESIS QUE PARA OBTENER EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA
JOSÉ MARTÍN OLGUÍN ESPINOZA

DIRECTOR
PEDRO MAYORGA ORTIZ

Mexicali, B. C., Enero de 2013

Resumen

Un elemento indispensable en la construcción de un sistema de Reconocimiento Automático de Locutor es el corpus de voz, el cual es necesario para la experimentación y evaluación. Las condiciones que debe cumplir un corpus para la tarea de reconocimiento es que contenga grabaciones de frases fonéticamente equilibradas de varios locutores usando distintos medios de grabación y a lo largo de varias sesiones. En este trabajo de tesis se presenta la metodología y resultados de la construcción del primer un corpus de voz apegado a la distribución fonética del Español Mexicano y orientado al reconocimiento de locutor. También se presenta la construcción de Vectores Acústicos Cuantílicos, una técnica novedosa en la obtención de vectores acústicos aplicada a la Identificación Automática de Locutor.

Abstract

Voice corpus is an essential element for Automatic Speaker Recognition systems. In order for a corpus to be useful in recognition tasks, it must contain recordings from several speakers pronouncing phonetically balanced utterances; recorded through several sessions using different recording media. This thesis shows the methodology, development and evaluation of the first Mexican Spanish Corpus wich is aimed to support research on speaker recognition. It also presents the construction of Quantile Acoustic Vectors, a novel technique in obtaining acoustic vectors applied to Automatic Speaker Identification.

Dedicatoria

A Héctor Manuel y a Mónica Livier, somos un equipo y vamos juntos en todo.

A mis padres, porque de uno admiro su sentido del humor a pesar de las adversidades y del otro la búsqueda constante de la perfección.

A Eva Cotero, por su ejemplo de fuerza y tenacidad.

A dos grandes ausentes, Pancho Venegas† por ser la influencia fundamental en mi carrera profesional; y a Nacho Ascencio† por sus lecciones prácticas sobre el significado del compañerismo y la amistad, su partida repentina me impidió agradecérselo personalmente.

Agradecimientos

Este trabajo de tesis hubiera sido imposible sin el apoyo de una gran cantidad de personas, a las cuales les estoy profundamente agradecido.

A mi director de tesis Dr. Pedro Mayorga, por la gran cantidad de horas invertidas y haberme "desatorado" cuando el cerebro ya no respondía.

A los integrantes del comité revisor, Dr. Hugo Hidalgo, Dr. Miguel Bravo, Dra. Larysa Burtseva y Dr. Angel Andrade, quienes semestre tras semestre dedicaron su tiempo para evaluar los avances y resultados.

Al M.C. Hugo González, por su valioso apoyo en la realización de los experimentos y del documento.

Al Dr. César Amaro, quien amablemente compartió el cuarto de grabación que armó con sus propias manos.

A la M.C. Brenda Flores, por sus acertadas opiniones sobre el documento y su invaluable apoyo como guía en las turbulentas aguas de la tramitología académica.

Al Ing. Alberto Mexia, por los recursos aportados al proyecto.

A las coordinadoras de la carrera de Ingeniero en Computación, "las jefas", M.C. Gloria Chávez y M.C. Aglay González-Pacheco, por las facilidades que me otorgaron para poder dedicarle tiempo a esta tesis.

A los estudiantes que "voluntariamente" participaron en las sesiones de grabación.

A la UABC, ya que a través de la 11va. Convocatoria Interna de Proyectos de Investigación financió parte de este trabajo.

Contenido

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

1 Introducción

Debido a que la voz es uno de los instrumentos de comunicación de los seres humanos, el procesamiento de las señales de voz por medio de la computadora ha sido una tarea de interés para la ciencia. El término *Tecnologías del Habla* es utilizado para designar al conjunto de aplicaciones que aprovechan el tratamiento digital de las señales de voz. Las Tecnologías del Habla por lo tanto, comprenden un amplio espectro de aplicaciones dentro de las cuales se ubica el *Reconocimiento Automático de Locutor (RAL)*. El objetivo principal del RAL consiste en identificar a un individuo con base en las características de su voz; formalmente, a dichos individuos se les denomina *locutores*. Un componente básico para la construcción de los sistemas de RAL es la base de datos de señales de voz pertenecientes a las personas que deberán ser reconocidas, dicha base de datos se denomina *corpus de voz*.

La tarea de RAL a su vez se subdivide en dos especializaciones, la *Verificación Automática de Locutor (VAL)* y la *Identificación Automática de Locutor (IAL)*. La primera consiste en autenticar una identidad proclamada, es decir, aceptar o negar la solicitud de identidad de un locutor. Para llevar a cabo esto tarea, en VAL es necesario contar con la señal de voz y la identidad proclamada (*i.e.* un dato vinculado con un locutor en particular). Los posibles resultados que arroja un sistema de VAL son dos: Aceptado o Rechazado.

Por otra parte, la IAL tiene como objetivo determinar, a partir de una señal de voz desconocida dada como entrada al sistema, a cual locutor pertenece dicha señal. Por lo tanto, el resultado de un sistema de IAL es una etiqueta que identifica de manera única a un locutor conocido por el sistema (IAL en conjunto cerrado) o si se toma en cuenta la posibilidad de que la señal no

pertenezca a ninguno de los locutores conocidos, el resultado será una etiqueta que identifica a un locutor impostor (IAL en conjunto abierto).

En términos generales se puede decir que en un sistema de RAL las señales de voz representan la *entrada*, mientras que la decisión representa la *salida*.

Las aplicaciones implementadas alrededor de los conceptos de RAL, VAL e IAL son vastas [1]; por ejemplo: accesos controlados por voz[1], segmentación de locutor sobre archivos de audio [2, 3], identificación de personas a través de conversaciones telefónicas[2] [4], identificación de participantes en videojuegos [5] o reconocimiento de grupos selectos [6], sólo por mencionar algunas.

1.1 Definición del Problema

La construcción de un sistema eficiente de RAL tiene que afrontar una serie de retos, tales como: el ruido en las señales de voz, el medio utilizado para capturar dichas señales, o la variabilidad que existe entre señales emitidas por el mismo locutor [7]. Además de los retos anteriores, existen niveles de información del habla en los cuales pueden enfocarse los sistemas de reconocimiento para caracterizar a cada locutor, en [8] se presenta una clasificación de dichos niveles como: espectral, prosódico, fonético, sintáctico, dialógico y semántico. Por otra parte, en [9-11] se reportan resultados de reconocimiento obtenidos mediante la fusión de sistemas enfocados en diferentes niveles, concluyendo que incluir información de diferentes niveles aumenta el desempeño en el reconocimiento.

Si bien existen avances importantes en los sistemas de RAL y es un área de investigación activa, en lo que respecta a estudios o aplicaciones específicas para el Español Mexicano no se

[1] Tecnología de voz a texto de la compañia Nuance: [http://www.nuance.com/for-business/by-product/nuance-voice-control/index.htm]

[2] Adaptación de soluciones en autenticación de voz de la compañía VoiceTrust (Alemania) [http://www.voicetrust.de/en/voice-biometrics-solutions-2.html]

encontraron trabajos previos en la literatura consultada. De aquí que, en el presente trabajo se propone como objetivo general el desarrollo de un sistema de identificación de locutor en el contexto del Español Mexicano, como una forma de generar una primera contribución al tema del reconocimiento de locutor para esa modalidad del idioma español.

Una primera pregunta que se tiene que plantear sobre el objetivo propuesto es la siguiente: ¿Es relevante para un sistema de IAL el idioma de los locutores? Si la respuesta fuera negativa, hacer un esfuerzo por desarrollar un sistema como el planteado no tendría interés científico alguno, ya que el proceso se reduciría a seguir lo establecido en los trabajos previos. Sin embargo, una respuesta afirmativa nos genera un cúmulo de nuevas preguntas, como por ejemplo: ¿Cuáles características del español mexicano son relevantes para hacer reconocimiento automático de locutor? ¿Qué características deben tener las señales de voz de estos locutores? Por supuesto que además de las preguntas anteriores se generarán otras más, lo importante a destacar es que tanto estas como las que pudieran surgir en el futuro, constituyen un punto de partida para una línea de investigación de IAL específica para el Español Mexicano, la cual, hasta donde cubrió el presente trabajo, no existe.

1.2 Alcance de la tesis

Partiendo del objetivo general planteado en la sección anterior: Desarrollo de un Sistema de Identificación de Locutor en el contexto del Español Mexicano, se establecen los siguientes objetivos particulares:

1) Identificar las características dependientes del idioma que son relevantes para caracterizar a un locutor.
2) Construir un corpus de voz que contenga características específicas del Español Mexicano.

Para el logro de los objetivos establecidos anteriormente, el trabajo de investigación se dividió en las siguientes etapas:

- Análisis del estado del arte en RAL.
- Integración de una plataforma de cómputo para IAL.
- Construcción y evaluación de un corpus de voz en Español Mexicano.

Como resultado de las actividades desarrolladas en estas etapas se obtuvieron resultados que constituyen las principales contribuciones del trabajo de tesis. La primera contribución es la validación de la distribución fonética del español mexicano utilizando textos periodísticos de la Web. La segunda contribución es el primer corpus de voz en Español Mexicano orientado para el reconocimiento de locutor. La tercera contribución es la aplicación de cuantiles estadísticos en la IAL.

1.3 Organización del documento

En el Capítulo 2 se hace una descripción del campo de las Tecnologías del Habla y un estudio del estado del arte del RAL y de la IAL.

Lo relacionado con la arquitectura y las herramientas computacionales para construir un sistema de IAL, están descritas en el Capítulo 3, en el cual se presenta una descripción de bibliotecas de software utilizadas y la integración de una plataforma de identificación de locutor basada en software libre.

En el Capítulo 4 se exponen las actividades realizadas para obtener la distribución fonética del Español Mexicano a partir de documentos recolectados automáticamente de la Web y su comparación con estudios previos.

En el Capítulo 5 se presenta la metodología, construcción y validación de *VoCMex*, el corpus de voz en Español Mexicano orientado al reconocimiento de locutor.

En el Capítulo 6 se presenta la teoría de cuantiles estadísticos, su aplicación a la construcción de vectores acústicos y los resultados de su uso en experimentos de IAL.

Por último, en el Capítulo 7 se presentan las conclusiones generales de la investigación y las perspectivas de trabajo a futuro.

Capítulo 2

Reconocimiento Automático de Locutor

2 Reconocimiento Automático de Locutor

En este capítulo, se explican cuáles son las áreas relacionadas con el tratamiento de voz, su relación entre ellas y sus particularidades. Primeramente, se establece una taxonomía para ubicar el tema central de esta tesis: el reconocimiento de locutor y las tareas en las que se subdivide, con énfasis especial en la IAL. Asimismo, se analizan los distintos niveles de aplicación en una plataforma dedicada al RAL, enfatizando cuál es la participación de este trabajo.

2.1 Taxonomía del Tratamiento de la Voz

Es común hoy en día ver en cine y televisión escenas de sistemas computarizados que identifican a una persona por medio de su voz, o que convierten a texto una conversación (transcripción automática). ¿Realmente las computadoras pueden hacer esto? Aunque parece una pregunta sencilla, la respuesta no lo es. Con la finalidad de diferenciar la ficción de la realidad, se hace una revisión de las *Tecnologías del Habla* (TH), la cual es la denominación del conjunto de técnicas especializadas en este tipo de tareas.

La materia prima de las tecnologías del habla es la *señal de voz*, nombre dado a la representación computacional de la elocución vocálica de una persona. Es decir, una señal de voz son datos numéricos, resultado del proceso de conversión analógico-digital.

En la Figura 2.1 se pueden ver las diferentes áreas en las que se subdividen las Tecnologías del Habla. Cada área se especializa en una tarea particular del tratamiento de las señales de voz, y su complejidad y varía entre ellas y en base a los métodos utilizados para abordarse [7, 12].

La *Síntesis de Voz* trata lo concerniente a la reproducción artificial de la voz, esto es, que la computadora sea capaz de "hablar", poniendo especial cuidado en que la voz generada artificialmente sea lo más parecida a la voz humana. En esta área se tiene especial interés en

considerar los elementos que permitan expresar estados de ánimo, sentimientos y demás factores que eviten detectar que la señal generada no es humana. En este tema hay un gran avance, las aplicaciones comerciales son numerosas, un ejemplo común es la generación automática de respuestas de voz implementadas en conmutadores telefónicos. Los laboratorios de investigación de la empresa AT&T tienen a disposición del público una aplicación web para hacer la conversión de texto a voz (TTS, *text-to-speech*) en diferentes idiomas (disponible en http://www.research.att.com/~ttsweb/tts/demo.php).

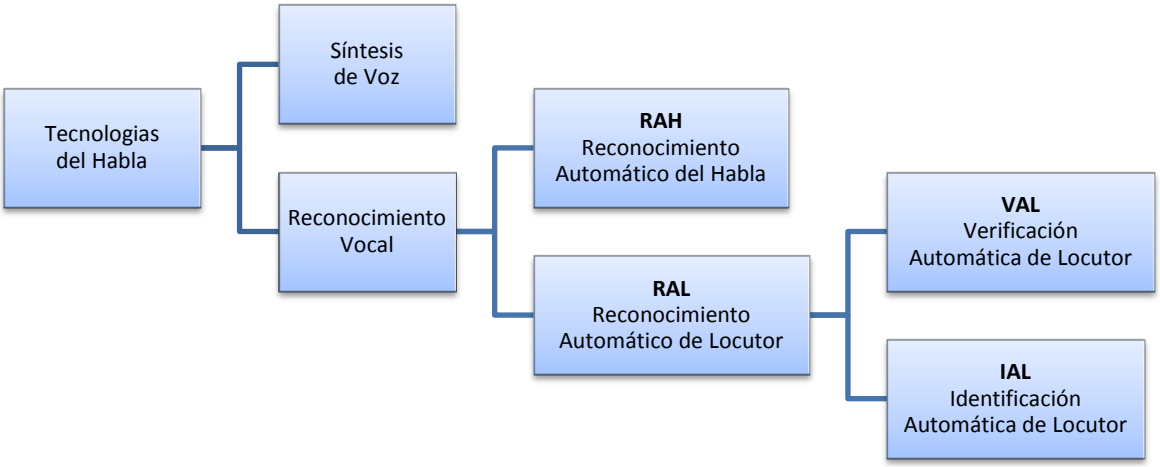

Figura 2.1 – Áreas de las Tecnologías del Habla.

El *Reconocimiento Vocal,* trata sobre la capacidad de la computadora para analizar una señal de voz, a fin de caracterizar tanto a la persona que habló como lo que dijo. Para esto se subdivide en dos tareas: *Reconocimiento Automático del Habla* y *Reconocimiento Automático de Locutor*.

2.2 Reconocimiento Automático del Habla y de Locutor

El *Reconocimiento Automático del Habla* (RAH) consiste en analizar una señal de voz y convertirla a texto. Uno de los productos más exitosos es el software *Dragon Naturally Speaking*® de la empresa Nuance, el cual permite que un usuario dicte a una computadora por medio de un micrófono para convertir automáticamente su habla a texto. En este campo existe un

gran avance, siendo el software anteriormente referido uno de los mejores ejemplos de un producto comercial. En RAH, mucho del trabajo pendiente está centrado en la construcción de modelos de lenguaje para diferentes idiomas, ya que hasta ahora el mayor esfuerzo se ha centrado en los idiomas inglés, japonés, alemán, francés y otros idiomas asiáticos. En el caso del español, predominan los esfuerzos sobre el español Ibérico (también llamado coloquialmente *español de España*). Desafortunadamente, hace falta profundizar en las demás variantes de este idioma, tales como el español de México o el español de Argentina (por citar algunos ejemplos).

La otra parte del Reconocimiento del Habla es el *Reconocimiento Automático de Locutor (RAL)*, cuyo objetivo es que la computadora identifique de manera automática a una persona por medio de su voz, ya sea en tiempo real o a través de una grabación. De acuerdo con Juang [13], el RAL se subdivide en **Verificación Automática de Locutor** (VAL) e **Identificación Automática de Locutor** (IAL). Mientras que en VAL un locutor proclama su identidad, en IAL no se provee ninguna información adicional además de la voz, por lo que el sistema debe identificar en su banco de datos al locutor a quien le corresponde esa señal, o en su caso establecer que no corresponde a ninguno de los locutores que tiene registrados.

En el dominio de RAL faltan muchos progresos por hacer, actualmente no existe software comercial de aplicación general que implemente RAL. No obstante, existen herramientas básicas (bibliotecas de software) para construir sistemas RAL de aplicación específica; sin embargo, su uso demanda un gran conocimiento tanto en programación de computadoras como en procesamiento digital de señales. Además, alcanzar el 100% de precisión en el reconocimiento es aún un reto por vencer. La gran cantidad de factores que influyen en la generación de la voz, así como los medios donde se genera o transmite, hacen difícil la automatización del reconocimiento, siendo este un campo de investigación muy activo.

2.3 Retos Asociados al RAL

Corinne Fredouille [7, 14] presenta una clasificación de los elementos que intervienen en las tareas de reconocimiento de locutor, los cuales seccionó en **variabilidad interlocutor**, **variabilidad intralocutor** y **variabilidad debida al material**.

2.3.1 Variabilidad Interlocutor

Esta característica es la que se desea aprovechar en RAL. La variabilidad intralocutor está dada por las diferencias que existen entre fonemas o frases al ser pronunciados por personas distintas. Además, en un contexto biológico, la fisionomía de un individuo impacta directamente en sus características vocales. Algunas características de la fisionomía humana que influyen en la voz incluyen: tracto vocal, tejido vivo, cadencia al hablar, género o edad. Identificar las diferencias antes mencionadas es lo que permite caracterizar a un locutor en particular.

2.3.2 Variabilidad Intralocutor

Si bien la variabilidad interlocutor permite distinguir a una persona de otra, la variabilidad intralocutor es un tanto confusa, ya que un mismo fonema o frase pronunciada por la misma persona es diferente en distintos momentos y bajo distintas circunstancias. En la Figura 2.2 se puede apreciar el mismo fonema pronunciado por la misma persona; además, se puede ver que la forma de la señal no es idéntica.

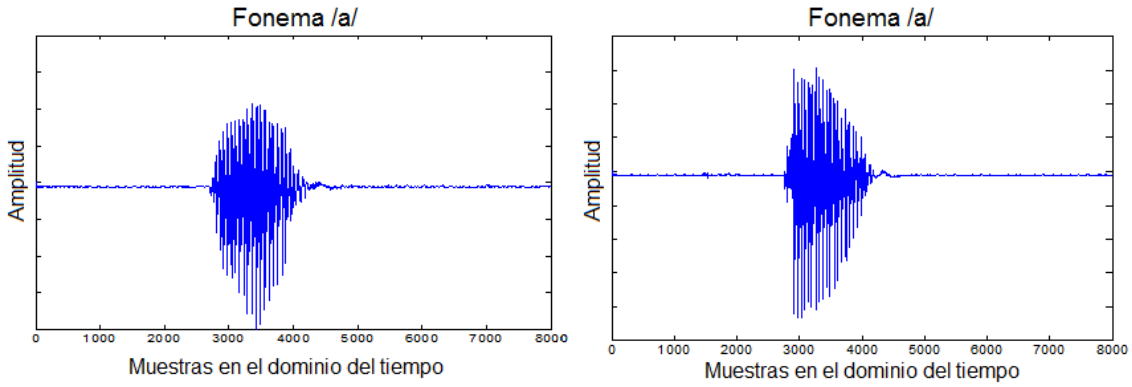

Figura 2.2 – Representación tiempo vs amplitud del fonema /a/ producido por la misma persona en dos ocasiones.

2.3.3 Variabilidad debida al Material

Asumiendo que es posible aprovechar la variabilidad interlocutor y que el efecto de la variabilidad intralocutor es despreciable, todavía existe un obstáculo más por salvar: la variabilidad debida al dispositivo con el que se captura la señal. Debido a que la conversión de señal analógica (las ondas sonoras) a señal digital (representación en la computadora) también puede presentar una fuente de distorsión debida al tipo de dispositivo utilizado, la señal digital será diferente entre un micrófono de alta definición y un teléfono celular, o entre un teléfono inalámbrico y uno convencional (alámbrico).

Las preguntas que se hacen los científicos de RAL están relacionadas con estas fuentes de variabilidad: ¿cómo hacer para que se pueda obtener una "huella" de la voz de una persona sin que se vea afectada por esta variabilidad?

2.4 Niveles de Información para el RAL

Los tipos de variabilidades expuestos en las secciones anteriores están estrechamente asociadas a las características físicas de la señal, lo que conduce a preguntarse si existen otras fuentes de información que se puedan aprovechar además de lo contenido en el nivel básico de la señal acústica. Para responder esta pregunta, en la Figura 2.3 los diferentes niveles de información que se pueden utilizar para caracterizar a un locutor, este modelo fue propuesto por Faúndez-Zanuy en 2005 [8]. Este modelo proporciona una guía para clasificar las diferentes aportaciones científicas realizadas en el campo del RAL y de manera gráfica muestra el concepto de integración de información establecido por Doddington en 2001 [15]. Los siguientes apartados explican las características de cada nivel y los trabajos previos relacionados que fueron encontrados a lo largo del desarrollo de la tesis.

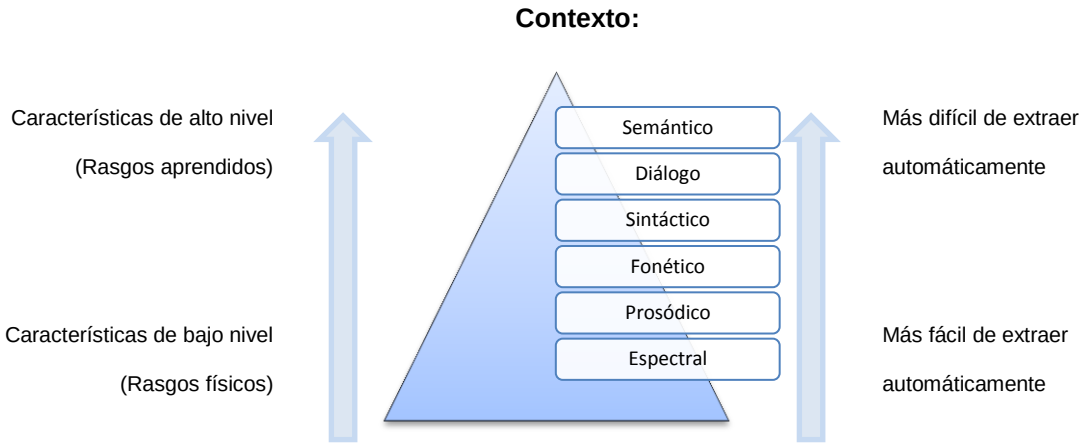

Figura 2.3 – Niveles de información para el reconocimiento automático de locutor, mostrados de manera jerárquica.

2.4.1 Nivel Espectral

En primer lugar se tiene el nivel espectral, el cual está relacionado directamente con la forma de la señal que produce el tracto vocal de un locutor. La anatomía de esta parte del cuerpo es diferente en cada persona, por lo que en este nivel se trata de modelar el tracto como un sistema físico [12]. Entre los métodos más encontrados en la literatura para este nivel están la Codificación Cepstral en Frecuencia de Mel (o MFCC, del inglés *Mel Frequency Cepstral Coding*), Codificación Lineal Predictiva (LPC, *Linear Predictive Coding*) o la Codificación Lineal Perceptual (PLP, *Perceptual Linear Coding*). Estos métodos son comúnmente utilizados para la obtención de vectores acústicos y posteriormente los locutores son modelados utilizando la técnica de Modelado de Mezclas Gaussianas (GMM, del inglés *Gaussian Mixture Models*) o Cuantización Vectorial (VQ, del inglés *Vector Quantization*) [16-20].

Una gran cantidad de trabajos de investigación se han desarrollado en relación a la obtención de información en este nivel. Magrin-Chagnolleau [21] propone el uso de Componentes Principales Tiempo-Frecuencia como una forma de obtener vectores acústicos; Kinnunen et al. [22] presentan una optimización al método de VQ; Hosseinzadeh y Krishan [23-25] presentan el

resultado de experimentos usando MFCC combinado con otros métodos menos comunes; Grimaldi [26] propone el uso de Frecuencias Instantáneas.

Si bien capturar señales en el contexto espectral es una tarea de relativa facilidad, el principal problema asociado al análisis a este nivel está directamente ligado a la variabilidad intralocutor y variabilidad de los materiales [27].

2.4.2 Nivel Prosódico

El nivel prosódico está centrado en la energía que el locutor produce al hablar; en este contexto, es posible hacer una distinción entre acento y entonación. Se puede distinguir a un locutor de otro identificando que tan "fuerte" o "suave" habla. Los sistemas basados en sólo información de este nivel no proveen resultados tan buenos como los que utilizan sólo información espectral; sin embargo, se ha demostrado que la combinación de ambos produce mejores resultados [28]. Una clasificación de los métodos de modelado prosódico pueden dividirse en: 1) Distribución de tono y energía, 2) Estadísticas de características prosódicas, y 3) Dinámica de contorno prosódico. En los métodos del primer tipo, usualmente se modela la distribución del tono con GMM y la energía a nivel de tramas de señal. Los del segundo tipo se enfocan en extraer y modelar varias características prosódicas de largo alcance (cubriendo más de una trama), tales como el estilo de entonación y pendientes de energía [29], características de duración y las características de regiones de extracción no uniformes (Non-uniform Extraction Regions Features, NERF), usando *n-gramas* a nivel de fonos [30]. En los métodos del tercer tipo se trata de aprender un modelo de lenguaje de gestos de contorno prosódico (mayores a una sílaba o segmento), mediante la conversión de las características prosódicas continuas en una secuencia de símbolos discretos, llamados *n-gramas* a nivel de sílabas [31].

2.4.3 Nivel Fonético

A nivel fonético se analizan las diferencias que presentan diversos locutores al pronunciar el mismo fonema; por ejemplo, el fonema /ch/ se pronuncia diferente en la región norte que en el centro de México. En este nivel es necesario introducir técnicas de reconocimiento del habla y modelos de lenguaje, porque es necesario, en primer lugar, determinar el contenido de la elocución para después tratar de analizar la manera en la que se pronuncia. Entre los trabajos encontrados para el reconocimiento aprovechando información en este nivel están [32, 33], que presentan el uso de Máquinas de Soporte Vectorial (SVM, *Support Vector Machine*s) para hacer verificación; Klusácek [34] y Leung [35] tratan el modelado de Pronunciación Condicional (CPM, *Conditional Pronunciation Modeling*), el cual consiste en reconocer al locutor modelando la relación entre lo que ha pronunciado (fonemas) contra cómo lo pronunció (fonos). Por otra parte, Hatch [36] presenta el uso de decodificación de retículas para mejorar el reconocimiento.

2.4.4 Nivel Sintáctico

El nivel sintáctico está relacionado con la frecuencia de palabras usadas por parte del locutor; un ejemplo es el uso de muletillas (“este”, “¿sí?”, “¿no?”). El trabajo pionero en este enfoque fue por parte de Doddington [15], el cual presenta el análisis de la frecuencia con la que un locutor pronuncia secuencias de palabras.

2.4.5 Niveles Dialógico y Semántico

En los niveles dialógico y semántico se analiza el contexto de una conversación para identificar frecuencia de frases del locutor bajo situaciones particulares con propósitos de segmentación, esto es, separar la señal de audio de manera que quede identificado cada segmento que corresponda a un locutor diferente; o con propósitos de agrupamiento, el cual consiste en reunir todos los segmentos que correspondan a un mismo locutor. Estos son los niveles más complejos

de analizar mediante la computadora ya que implica, al igual que los niveles fonético y sintáctico, aplicar simultáneamente métodos de reconocimiento del habla, porque es necesario primero conocer el contenido de la elocución para posteriormente aplicar el análisis de las características de los locutores.

Desde la década de los 90's, existen campañas de evaluación dedicadas a los sistemas RAL. Estas campañas son organizadas cada año por el Instituto de Estándares y Tecnologías NIST (*National Institute of Standars and Technologies*) desde 1996, orientadas a los laboratorios de investigación públicos y privados. Primeramente, dichas campañas estaban basadas en la evaluación de los sistemas VAL en un contexto conversacional y telefónico, pero después se extendieron a otro tipo de tareas [37-39].

El estado actual de la investigación en RAL está centrado en la obtención de información de los niveles anteriormente presentados y en la construcción de sistemas que integren eficientemente la información arrojada en cada nivel, para elevar la precisión en el reconocimiento [9, 10, 15]. Más recientemente, Baum [40] hace un estudio explorando las posibilidades de aplicar modelos de tópico (temas de conversación); su investigación considera dos tipos de modelado de conversación (*topic modeling*): El método de *implicidad*, el cual se basa en palabras clave pertenecientes a un locutor dado; y el método de *explicidad*, en el cual se crean modelos de conversación derivados automáticamente y se asignan a segmentos de una conversación. Esta investigación arrojó que sí existen segmentos de habla durante una conversación que reflejan información particular de un locutor.

Además del enfoque de integración de niveles de información, también existen avances en la combinación de audio y video para hacer reconocimiento, con la finalidad de aumentar la precisión de los sistemas [41].

Actualmente, con el desarrollo de la tecnología móvil, la adaptación de los sistemas RAL e IAL a condiciones de ruido, movimiento y calidad de la señal son motivo de estudio [42]. Desafortunadamente, el español, en particular en el contexto mexicano, se encuentra aún rezagado en este contexto, por lo cual resulta necesario hacer una revisión y una propuesta a ésta problemática.

2.5 Modelado de Locutor

Con la información que se obtiene de los niveles mencionados en la sección 2.4, se deben crear los modelos de cada locutor que serán los que conforman la base de locutores cliente que el sistema debe reconocer. En el trabajo de Kinnunen del 2001 [43], se presentan dos clasificaciones de modelos: en paramétricos y no paramétricos; o generativos y discriminativos. En los paramétricos, cada locutor es modelado como una fuente probabilística con una función de densidad de probabilidad desconocida pero fija. La fase de entrenamiento consiste en estimar los parámetros de la función de densidad de probabilidad a partir de una muestra de datos. El reconocimiento usualmente se hace evaluando la similitud de la señal de prueba con respecto al modelo. En esta clase de modelos tenemos dos predominantes, GMM para reconocimiento independiente del texto y Modelos Ocultos de Markov (HMM, *Hidden Markov Models*) para reconocimiento dependiente del texto.

En el caso de modelos no paramétricos, los vectores acústicos de entrenamiento y prueba se comparan directamente con la presunción de que uno es una réplica imperfecta del otro. El monto de la distorsión entre uno y otro representa el grado de similitud entre ellos. La Cuantización Vectorial (VQ) y la distorsión de tiempo dinámica (DTW, *Dynamic Time Warping*) son dos métodos representativos para reconocimiento independiente del texto y dependiente del texto, respectivamente.

En la otra clasificación, los modelos generativos como los GMM y VQ estiman la distribución interna de características de cada locutor. En contraste, los modelos discriminativos, tales como las Redes Neuronales Artificiales (ANN, del inglés *Artificial Neural Network*) y las máquinas de soporte vectorial (SVM), modelan los límites ente locutores.

2.6 La Identificación Automática de Locutor (IAL)

La Identificación Automática del Locutor consiste en determinar de entre una población de locutores conocidos, la persona a la que pertenece cierta señal de voz dada como entrada. En la identificación se proponen dos modos: identificación en *conjunto cerrado*, para el cual se asume que la señal de voz es pronunciada por un locutor conocido por el sistema. La salida del sistema de identificación en este modo será el identificador del locutor con la mayor similitud a la señal de voz de entrada. Por otro lado, en identificación en *conjunto abierto*, para la cual cabe la posibilidad de que el locutor pueda no pertenecer al conjunto de locutores conocidos por el sistema, es decir, la posibilidad que el locutor sea un *impostor*. En identificación en conjunto abierto, el sistema de identificación debe decidir la fiabilidad de su juicio aceptando o rechazando la identidad que encontró. Si el sistema la acepta debe además establecer el identificador del locutor al que pertenece la señal de voz de entrada.

Independientemente de que sea en conjunto abierto o cerrado, la IAL también tiene las modalidades de identificación *dependiente del texto* e identificación *independiente del texto*. En la primera, la identificación se hace mediante la elocución de una frase o frases predefinidas, mientras que en la segunda, la identificación se realiza no importando lo que el locutor exprese. En la primera es común que se utilicen en conjunto métodos de reconocimiento del habla mientras que en la segunda no es necesario.

2.6.1 Arquitectura de un sistema de IAL

Desde el punto de vista de *arquitectura* (Figura 2.4), una secuencia de voz representa la entrada a un sistema IAL. Para cada locutor ya conocido por el sistema, ésta secuencia es "comparada" a un conjunto de referencias, que son características de cada locutor. La identidad del locutor para la cual la referencia resulta la más "próxima" a la secuencia de entrada, es el resultado de salida del sistema IAL. Esta salida es en realidad una hipótesis del sistema, ya que lo que se pretende medir es la eficiencia del sistema en términos de sus aciertos.

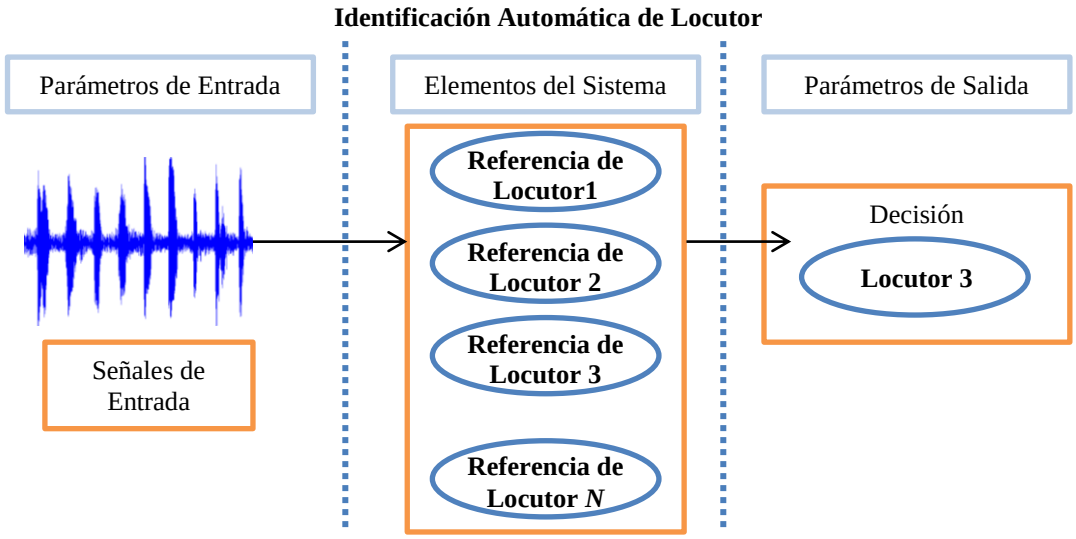

Figura 2.4 – Diagrama a bloques de la tarea IAL.

Un principio básico de los sistemas de IAL es determinar una identidad de entre todas las identidades potenciales y generalmente las eficiencias de estos sistemas se degradan cuando la población de locutores aumenta.

En IAL, las aplicaciones no son tan numerosas como las encontradas en reconocimiento del habla, sin embargo sí son complementarias con esta última. En este sentido, una aplicación común es la adaptación de modelos de lenguaje específicos para un locutor con el objetivo de mejorar la conversión de habla-a-texto (STT). Esto es, primero determinar quien está hablando

para poder elegir los parámetros específicos prestablecidos para ese locutor, lo cual asegura que el reconocimiento del habla se hará usando el modelo que mayormente se ajuste a la persona que está hablando. Ejemplos de esta fusión de IAL con RAH se presentan en [44] y [45].

2.7 Resumen y Conclusiones del Capítulo

En este capítulo se presentó una revisión del estado del arte en el reconocimiento de locutor. Se inicia con la ubicación del RAL dentro de la taxonomía de las aplicaciones de las tecnologías de la voz y las tareas en las que subdivide: VAL e IAL. Posteriormente se han presentado los retos que presentan las tareas de reconocimiento; por un parte, se presentan aquellos relacionados con los materiales con los que se adquiere la señal; y por la otra, los relacionados con el locutor propiamente, los cuales a su vez se pueden analizar por niveles de información: espectral, prosódico, fonético, sintáctico, dialógico y semántico. Para cada nivel se hace una descripción y se mencionan los trabajos previos que sustentan lo ahí expuesto. Un elemento esencial de los RAL es el método mediante el cual se modelan los locutores, por eso se presenta una clasificación de los principales métodos encontrados en la literatura, tales como GMM, VQ, SVM, ANN.

Después se presenta la descripción de lo que atañe directamente al alcance de esta tesis, la identificación de locutor, en esta parte se definen las modalidades: conjunto abierto, conjunto cerrado, dependiente del texto e independiente del texto.

De la información presentada se concluye que el RAL es una tarea muy compleja, los retos que presenta son numerosos y que siguen siendo problemas abiertos, existen muchos esfuerzos para abordarlo desde diferentes enfoques. Siendo los niveles espectral y prosódico los más estudiados debido principalmente a la posibilidad de aplicar las técnicas del procesamiento digital de señales. Por otra parte, los niveles superiores, fonético, sintáctico, dialógico y semántico

representan grandes retos porque requieren integración por un lado de sistemas de nivel espectral y prosódico, pero por otro de sistemas de reconocimiento del habla y técnicas estadísticas para modelar aspectos que las personas han aprendido y no sólo dependen de la conformación de su tracto vocal. Es en estos niveles superiores en los que el idioma de los locutores adquiere relevancia, ya que primero es necesario reconocer lo que están diciendo (fonemas, sílabas, palabras), mediante técnicas de reconocimiento del habla, para después caracterizar las particularidades de cada locutor. Se espera que en el futuro inmediato estos sean los temas de mayor énfasis en el RAL.

Capítulo 3

Integración de una Plataforma de Cómputo para la Identificación de Locutor

3 Integración de una Plataforma de Cómputo para la Identificación de Locutor

Sin importar la tarea a considerar, un sistema RAL se resume al encadenamiento de tres procesos principales: *parametrización*, *reconocimiento* y *decisión*. En la parametrización, los datos con los que se trabaja deben analizarse y caracterizarse en un espacio distinto al que se encuentran originalmente, es decir, quedan *parametrizados*; la finalidad de este análisis es extraer y realzar las características propias (o únicas) de los datos. En la etapa de reconocimiento, un sistema debe valerse de un algoritmo para medir, o calificar, los atributos de los datos de entrada; los algoritmos empleados en esta etapa varían de acuerdo a la técnica o a la aplicación final de la tarea de reconocimiento. Finalmente, en la etapa de decisión, a las medidas de la etapa de reconocimiento deben asignarles una etiqueta o clase, la cual representa el resultado final de la tarea RAL.

Contrariamente al proceso de parametrización, los principios para poner en marcha el reconocimiento y la decisión están muy estrechamente relacionados a la aplicación considerada [7, 14]. El proceso de reconocimiento (y por lo tanto el sistema asociado), queda restringido por una de las dos siguientes consideraciones:

- El sistema **es dependiente** de un modelado de las características de locutores que son conocidos por el sistema de reconocimiento. En dado caso, se crean modelos “cliente”, es decir, modelos de aquellos locutores a quienes se pretende diferenciar de un grupo genérico. Este es el caso de los sistemas IAL, VAL y de seguimiento de locutor.
- El sistema es **independiente** de un modelado de las características de locutores. Éste es el caso de los sistemas de indexación por locutor de un flujo de audio.

En IAL, una señal de prueba es comparada con todas las referencias de locutor conocidas por el sistema, lo que resulta en un conjunto de **valores de similitud** (o *scores*). Dichos scores son utilizados por el proceso de decisión: en éste, se determina el ***score* que presente la similitud máxima** con una referencia de locutor, lo que hace posible asignarle esta referencia a la señal de prueba. En otras palabras, se asigna una etiqueta/pertenencia/clase correspondiente a la señal de entrada, en base a sus scores.

3.1 Plataformas existentes para la tarea de IAL

Debido a la investigación poco abundante en identificación de locutor en el contexto del español mexicano, es preponderante seguir un protocolo antes de la integración de una plataforma. En este tenor, es importante revisar la aportaciones de otros países de lenguas similares (Ibéricas y de Latinoamericanas), así como las contribuciones por parte de investigadores mexicanos. Es preciso remarcar que la mayoría de las contribuciones en este campo en México han sido aportadas por instituciones como la UNAM[3], IPN[4], UDLAP[5] e INAOE[6].

Es posible realizar una contribución partiendo *desde cero*, es decir, implementando todas las rutinas necesarias (lo cual sería una labor compleja); por el contrario, es más práctico apoyarse en una búsqueda de herramientas disponibles, implementadas por otros laboratorios exitosos en el mundo.

En este apartado se presentan las herramientas de software orientadas al tratamientos de la voz encontradas durante la investigación. En la Figura 3.1 se muestra el protocolo que se llevó a cabo para contar con un sistema que sirviera para la experimentación con IAL.

[3] Universidad Nacional Autónoma de México (México D.F., México) [http://www.unam.mx]
[4] Instituto Politécnico Nacional (México D.F., México) [http://www.ipn.mx]
[5] Universidad de las Américas Puebla (Puebla, México) [http://www.udlap.mx]
[6] Instituto Nacional de Astrofísica, Óptica y Electrónica (Tonantzintla, México) [http://www.inaoep.mx]

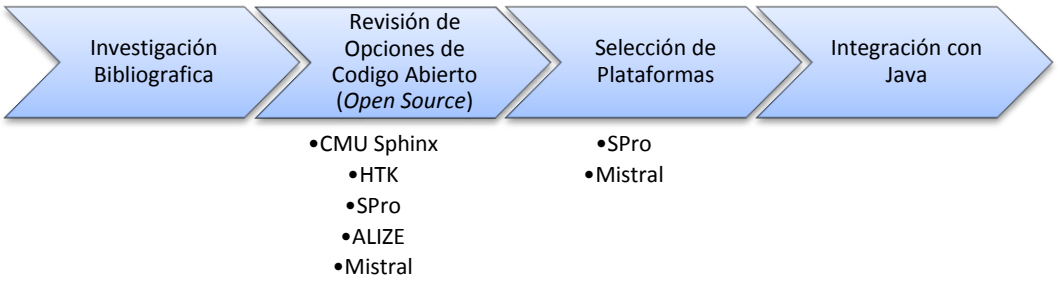

Figura 3.1 – Protocolo para la búsqueda de herramientas orientadas a integración de una plataforma.

3.1.1 CMU Sphinx

CMU Sphinx, son una serie de herramientas orientadas al reconocimiento del habla, desarrollados por CMU (Carnegie Mellon University, Pittsburg, Pennsylvania, USA). Dichas herramientas permiten construir aplicaciones en el dominio del reconocimiento y análisis de voz, e incluyen reconocedores y entrenadores. La lista de los módulos de los que se compone se resume a continuación:

- CMUclmtk: Herramientas para modelado de lenguaje.
- Pocketsphinx: Es una pequeña biblioteca de rutinas para el reconocimiento, escrita en lenguaje C.
- Sphinxbase: Una biblioteca de apoyo requerida por Pocketsphinx.
- Sphinxtrain: Rutinas de entrenamiento de modelos acústicos.
- Sphinx3: Decodificador para investigación en reconocimiento del habla, escrito en C.
- Sphinx4: Reconocedor modificable y ajustable, escrito en lenguaje Java.

3.1.2 HTK, Universidad de Cambridge

HTK (*HMM Toolkit*) es una serie de herramientas orientadas a la creación y manipulación de Modelos Ocultos de Markov (HMM, del inglés *Hidden Markov Models*). Este software fue desarrollado por el Departamento de Ingeniería de la Universidad de Cambridge (CUED). HTK ha sido utilizado principalmente en reconocimiento de locutor, aunque ha sido usado también

para síntesis de voz o secuenciación de ADN. HTK consiste de una librería de módulos escritos en C. Estas herramientas permiten crear, probar y analizar resultados con HMM.

3.1.3 SPro

SPro[7] es un conjunto de herramientas gratuitas, en el contexto de procesamiento de señales de voz. Provee algoritmos para extracción de características relacionadas con aplicaciones de voz y una biblioteca de módulos en C para implementar nuevos algoritmos, lo cual permite utilizar SPro con fines de desarrollo propios.

SPro fue diseñado originalmente para análisis espectral de resolución variable, pero también para técnicas de extracción clásicas en aplicaciones de voz. Existen apartados para las siguientes representaciones:

- Energías de banco de filtros.
- Coeficientes cepstrales.
- Representación derivada de predicción lineal.

A través de las herramientas de SPro, y en base a necesidades propias, es posible implementar aplicaciones dedicadas en el ámbito de RAL o VAL.

3.1.4 ALIZE

ALIZE es una plataforma libre (distribuida bajo la licencia LGPL) para la autenticación biométrica [46]. El objetivo de ALIZE es permitir desarrollos o implementaciones en la mayor parte de las aplicaciones biométricas, proveyendo un conjunto de aplicaciones de bajo y alto nivel. A fin de permitir a cada uno de los usuarios utilizar ALIZE en función de sus necesidades, ésta se compone de varias partes:

[7] http://www.irisa.fr/metiss/guig/spro/

- Una biblioteca de bajo nivel ALIZE: este módulo contiene todas las funciones necesarias para el uso de mezclados gaussianos.
- Un conjunto de rutinas de alto nivel, las cuales están separados en sub-partes:
  - LIA_SpkTools: Rutinas de integración de la biblioteca de alto nivel.
  - LIA_Utils: Las herramientas necesarias para la manipulación de diferentes formatos de datos utilizados en modelos GMM ALIZE, parámetros, etc.
  - LIA_SpkDet: Un conjunto de herramientas para efectuar las distintas tareas de un sistema de reconocimiento biométrico: aprendizaje de modelos (modelo de mundo y locutores), normalización de parámetros, normalización de *scores*, etc.
  - LIA_SPkSeg: Una paquetería nueva para hacer la segmentación en locutor.

En la Figura 3.2 se muestran los módulos de los que se compone ALIZE.

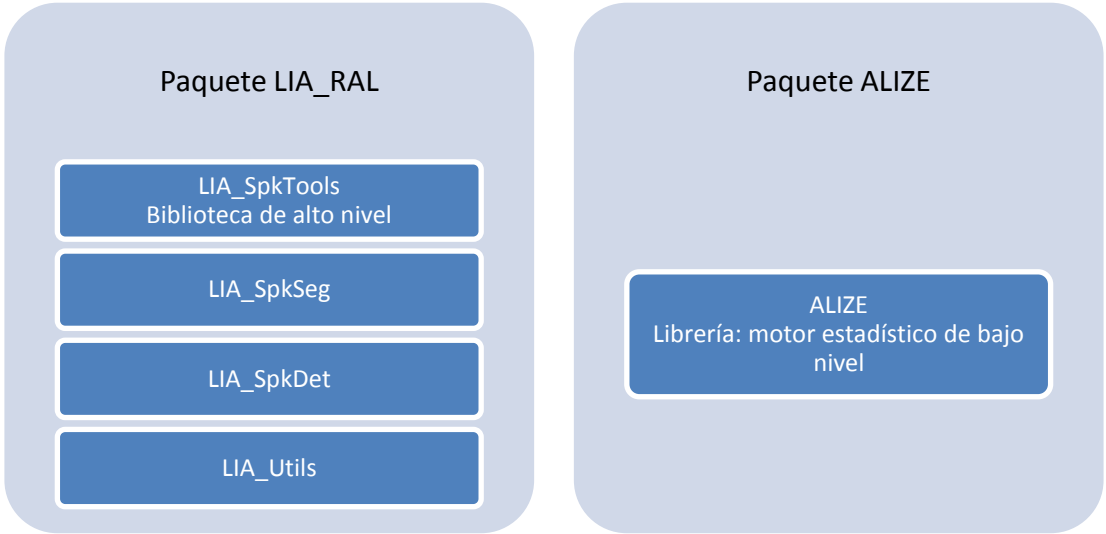

Figura 3.2 – Estructura general de los módulos de ALIZE.

3.1.5 Mistral

Mistral es un conjunto de herramientas de código abierto para aplicaciones biométricas [46]. Este software, basado HMM, en GMM y UBM (*Universal Background Model*) incluye también otros

desarrollos en reconocimiento de locutor, tales como análisis de factor y adaptación no supervisada para super-vectores SVM. Está conformado por módulos de modelado y reconocimiento, dentro de los cuales destaca ALIZE, que sirve como motor estadístico de Mistral.

Mistral crea y maneja modelos estadísticos a través de ALIZE (y sus algoritmos EM, ML y MAP), lo que permite hacer reconocimientos biométricos, entre los cuales destacan segmentación y clasificación de locutor, verificación de locutor, detección de lenguaje y reconocimiento facial. En la Figura 3.3 se muestra el diagrama de flujo para lograr RAL con Mistral.

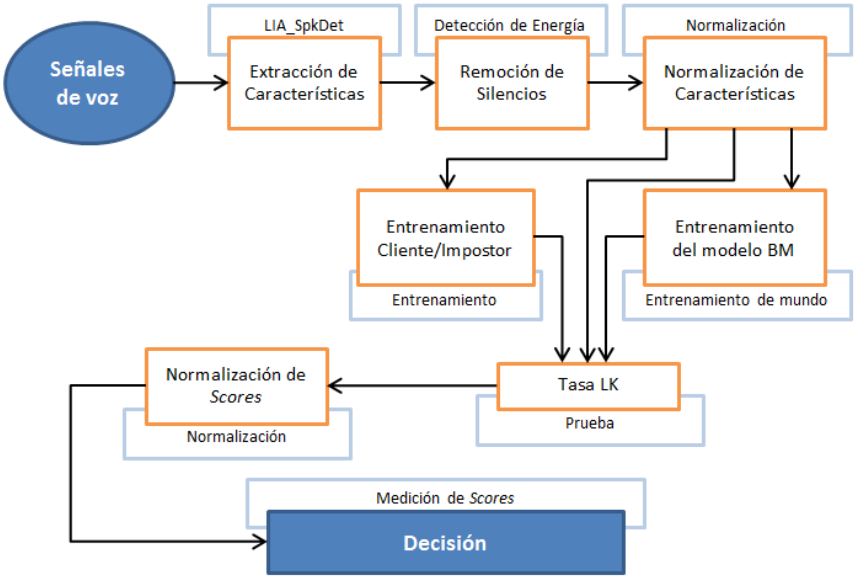

Figura 3.3 – Diagrama de flujo que ilustra las tareas de reconocimiento utilizando MISTRAL.

3.2 Integración de una Plataforma

Una vez analizadas las plataformas para procesamiento de voz y reconocimiento de locutor anteriormente expuestas, se concluyó que las herramientas que ofrecen lo necesario para integrar una plataforma de experimentación de IAL son SPro y Mistral [47]. Para integrar estas dos herramientas, se construyó una interfaz gráfica de usuario que facilita el acceso a los módulos de

cada software necesarios para las tareas relacionadas con identificación. La arquitectura de dicha integración se muestra en la Figura 3.4.

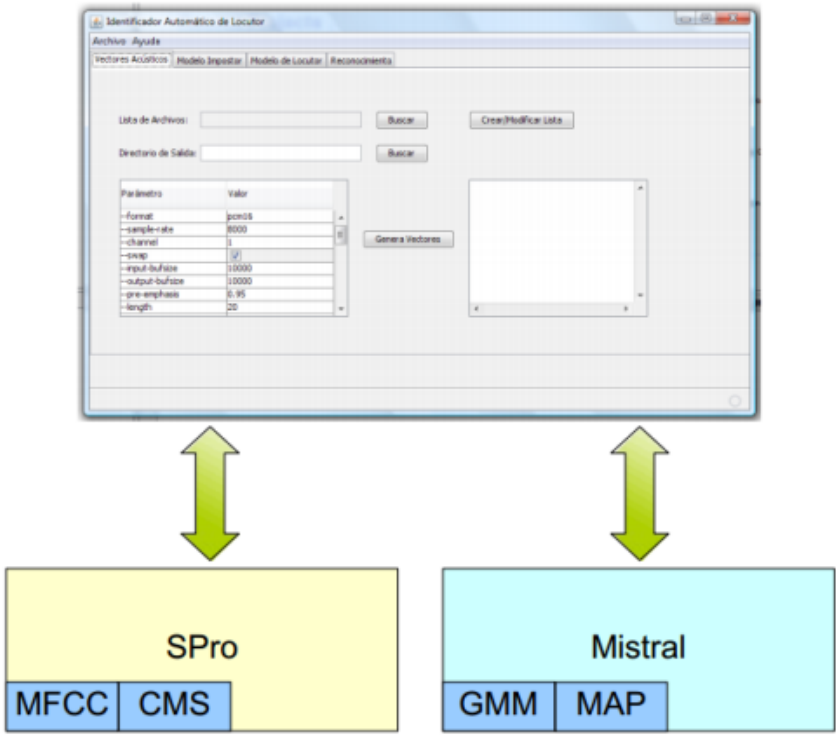

Figura 3.4 – Diagrama que ilustra la implementación de la plataforma.

Como se observa, consta de una interfaz gráfica en la cual se tiene la alternativa de buscar los archivos con los cuales se realizará el experimento, así como el directorio donde se guardarán los modelos y los resultados; otra de las ventanas nos da la posibilidad de seleccionar el formato del archivo a leer (*wav*, *pcm*, etc.), la frecuencia de muestreo, y otros parámetros y pre-procesamiento para generar el vector acústico.

Los vectores acústicos son generados utilizando SPro, dentro de las alternativas para la construcción de dichos vectores se puede elegir MFCC, con o sin Substracción de Media Cepstral (o CMS), con o sin valores para la primera y segunda derivada, con o sin cálculo del coeficiente de energía. Para el modelado de los locutores se utiliza Mistral, con generación de

modelos mezclados gaussianos (GMM) y adaptación MAP (Estimación de Máxima *A-Posteriori*, o Maximum *A-Posteriori* Estimation).

Los parámetros de los archivos de audio utilizados fueron se muestran en la Tabla 3.2.

3.3 Evaluación de la Plataforma y Resultados

Con la finalidad de hacer una evaluación del desempeño del sistema de IAL integrado con SPro y Mistral, se realizó un experimento con señales del corpus AHUMADA, el cual contiene grabaciones en español Ibérico [48]. La configuración de las particiones para los conjuntos de entrenamiento y prueba se muestran en la Tabla 3.1.

Tabla 3.1 - Partición de los conjuntos de entrenamiento y prueba.

Entrenamiento	Prueba
•5 locutores •3 sesiones de señal telefónica	•10 locutores •1 sesión de señal de micrófono •1 sesión de señal telefónica

La partición fue elegida de esa manera con la finalidad de evaluar el sistema en una de las condiciones más difíciles para un sistema IAL, la cual involucra entrenarlo con señal adquirida por un tipo de material, en este caso teléfono, y probarla con un conjunto de señales adquiridas por distintos materiales (incluyendo el mismo utilizado en el entrenamiento), además que el contenido de las grabaciones para entrenamiento consistieron de dígitos y las grabaciones para prueba consistieron en frases.

Para el la fase de entrenamiento primeramente se generaron los vectores acústicos a partir de los archivos de audio de todos los locutores. Posteriormente se generó un modelo de mundo (modelo impostor) a partir del cual se generaron los modelos clientes mediante la técnica *Maximum-A-*

Posteriori Estimation (MAP) [7, 14]. Los parámetros utilizados para la generación de los vectores acústicos y los modelos GMM, se muestran en la Tabla 3.2.

Tabla 3.2 - Características de los vectores acústicos y modelos.

Vectores Acústicos:	Modelado
• Tamaño de ventana: 30 ms • Desplazamiento: 10 ms • Canales en el banco de • Filtros: 26 • Coeficientes MFCC: 13 • Valores de la primera • derivada: 13 • Valores de la segunda derivada: 13 • Coeficientes de energía: 1 • Compensación CMS: Si • Coeficientes en cada vector: 40	• Componentes GMM: 32 • Matriz de covarianza: Diagonal • Entrenamiento EM: 7 iteraciones • Modelo de Mundo para modelo general: Todos los locutores (*world model*) • Un modelo por cada locutor (*cliente*) • *El modelo cliente se obtiene usando adaptación MAP

El resultado obtenido se muestra en la Tabla 3.3, en donde se observa que a pesar de las condiciones disímiles en las señales de entrenamiento y prueba, el sistema logró obtener un reconocimiento por encima del 50%.

Tabla 3.3 - Porcentajes de reconocimiento.

	%
Reconocimiento (aciertos/total)	60%
Identificaciones Erróneas	20%
Falsas Alarmas	20%
Total:	100.00%

El mismo resultado mostrado anteriormente se obtuvo utilizando las herramientas SPro y Mistral sin el ambiente de integración presentado en el apartado 3.2.

3.4 Resumen y Conclusiones del Capítulo

En este capítulo se ha presentado el resultado de la búsqueda de herramientas de software para el tratamiento de la voz. En este sentido, se encontró que existen una diversidad de herramientas que están a la libre disposición de quienes estén interesados en contribuir, tanto en el ámbito del

uso de las herramientas con fines de investigación, como contribuir en el propio desarrollo de las herramientas o en el desarrollo de aplicaciones prácticas de tecnologías de la voz.

De las herramientas encontradas se eligieron SPro y Mistral para integrarlas en una plataforma que permita la experimentación en IAL. Con la finalidad de probar si la plataforma integrada funciona adecuadamente, se construyó un sistema de identificación basado en un subconjunto de las señales del corpus Ahumada; se encontró que el desempeño en el reconocimiento fue igual con la plataforma integrada que haciéndolo con las herramientas por separado. Esta plataforma permitirá posteriormente realizar experimentos más completos.

Capítulo 4

Distribución Fonética del Español Mexicano

4 Distribución Fonética del Español Mexicano

Para experimentar con sistemas de RAL se requieren bases de datos de señales de voz las cuales deben cumplir con varias características, entre las que destacan: a) Incluir todos los fonemas del lenguaje; y b) Preservar la distribución fonética del lenguaje [49].

El Diccionario de la Real Academia de la Lengua Española define el término fonema como: *"Cada una de las unidades fonológicas mínimas que en el sistema de una lengua pueden oponerse a otras en contraste significativo".* En términos simples, se puede decir que los fonemas son los sonidos básicos utilizados en un idioma.

En [8, 50-53] se establece la ventaja de asignarles peso a los fonemas para mejorar el reconocimiento, por lo tanto es importante que un corpus de voz contenga al menos los fonemas más utilizados en el lenguaje nativo de los locutores. Por esta razón, el primer paso para la construcción de un corpus es la definición de una o más frases que los locutores deben grabar. Sin embargo, para definir estas frases es necesario conocer la distribución fonética del lenguaje y en base a ésta construir las frases fonéticamente equilibradas, tal y como se reporta en la construcción del corpus AHUMADA [48].

Pérez [54] presenta la distribución fonética del español latinoamericano obtenida mediante el análisis de grabaciones de audio correspondientes a noticieros chilenos, Villaseñor-Pineda [55] la obtuvo a partir de fuentes de texto de páginas web, ambos trabajos concuerdan en que la diferencia entre el español ibérico y el latinoamericano (chileno y mexicano) es la frecuencia de uso de los fonemas /a/ y /e/; en el español ibérico se utiliza más el fonema /a/ que el fonema /e/ y en el español latinoamericano es al contrario, tiene una mayor frecuencia de uso el fonema /e/ que el fonema /a/.

En los siguientes apartados se muestran las actividades realizadas con la finalidad de validar la distribución fonética del español mexicano encontrada en la bibliografía.

4.1 Definición de Fuentes y Recolección de Datos

Para este trabajo se utilizó un enfoque similar al presentado por [55], esto es, se utilizaron páginas web como única fuente de información. La principal diferencia consistió en el tipo de fuentes de datos utilizadas y el volumen total de información recolectada (se requirió alrededor de 90% menos datos). Las páginas web fueron seleccionadas de acuerdo a los siguientes criterios:

Páginas web en español de periódicos mexicanos de diferentes áreas geográficas. Para esto se utilizó la distribución de zonas propuesta por una editorial de periódicos mexicana cuyo portal de Internet publica los diarios que tiene a lo largo de todo el país (7 zonas: Norte, Golfo, Noroeste, Centro, Bajío, Sur, Este).

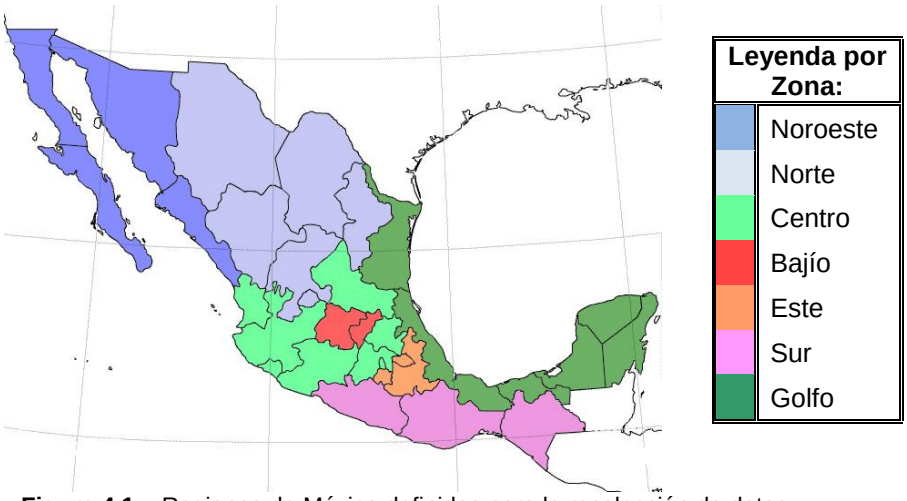

Figura 4.1 – Regiones de México definidas para la recolección de datos.

Con la finalidad de capturar las palabras utilizadas específicamente en esa zonas geográficas (*regionalismos*), sólo se utilizaron las páginas correspondientes a las noticias locales, las cuales se asume fueron escritas por reporteros locales.

4.1.1 Extracción de información a través de Web-Harvest

Web-Harvest[8] es una herramienta desarrollada en lenguaje de programación Java, diseñada para extraer información de la Web. A través ésta, es posible el analizar código de una página Web (HTML) y extraer datos que se consideren pertinentes a través de manipulación de texto común y XML. De esta manera, fue posible analizar de manera automática diversos sitios de noticias locales para extraer el texto de sus artículos, con el fin de modelar la distribución fonética del español mexicano.

En total, mediante esta herramienta se recabaron datos de 39 periódicos en línea, durante un lapso de 20 días, acumulando un total de 157MB de texto. Para cada región se generó una base de datos en la que se almacenó la lista de palabras encontradas y el número de veces que se encontró dicha palabra. El objetivo de almacenar los datos separados por región fue para poder hacer un análisis por áreas geográficas e identificar, en caso de que hubiera, las diferencias en el uso de los fonemas entre zonas.

4.2 Análisis de Datos y Resultados

Los fonemas analizados corresponden a la definición básica en [54], esta consiste de 22 fonemas básicos para el español y sus correspondientes alófonos (variaciones en la pronunciación de los fonemas en razón de su ubicación en palabras y/o frases), sin embargo, utilizar texto como fuente de análisis dificulta identificar estos últimos, razón por la cual los alófonos no están considerados dentro del alcance del presente trabajo.

Para realizar la obtención de la frecuencia de fonemas, se aplicó un algoritmo para analizar el texto recabado e ir acumulando las ocurrencias de cada uno, este algoritmo se muestra en la Tabla 4.1. La identificación de fonemas en el texto se realizó mediante el uso de expresiones

[8] Sitio oficial: http://web-harvest.sourceforge.net

regulares. Una vez obtenido el número de apariciones de cada fonema, se calculó la frecuencia de cada uno y se convirtió en porcentaje.

Tabla 4.1 - Algoritmo aplicado para contar las ocurrencias de los fonemas en el conjunto de palabras.

```
Analiza_Fonemas(Conjunto_de_Palabras)
1. Para cada Palabra del Conjunto_de_Palabras
 1.1. Para cada Fonema de la Lista_de_Fonemas
   1.1.1. acumula el número de ocurrencias del Fonema en Palabra
   1.1.2. almacena el resultado para la dupla Palabra-Fonema multiplicado
          por las ocurrencias de Palabra
 1.2 Fin-Para
2. Fin-Para
```

El porcentaje de aparición por zona se muestra en el Apéndice C, donde se observa que no existen diferencias significativas de la frecuencia de los fonemas entre las diversas zonas, para corroborarlo se calculó el coeficiente de correlación entre cada zona y el resultado en todos los casos es cercano a la unidad, lo que comprueba la similitud, tal como se muestra en la Tabla 4.2.

Tabla 4.2 - Análisis de correlación de la frecuencia de fonemas encontrada en las zonas.

	Norte	Golfo	Pacífico	Centro	Sur	Bajío	Oriente
Norte	1.000000	0.999615	0.999756	0.999593	0.999526	0.999601	0.999635
Golfo	0.999615	1.000000	0.999783	0.999530	0.999621	0.999679	0.999775
Pacífico	0.999756	0.999783	1.000000	0.999823	0.999493	0.999755	0.999895
Centro	0.999593	0.999530	0.999823	1.000000	0.999184	0.999808	0.999814
Sur	0.999526	0.999621	0.999493	0.999184	1.000000	0.999474	0.999607
Bajío	0.999601	0.999679	0.999755	0.999808	0.999474	1.000000	0.999883
Oriente	0.999635	0.999775	0.999895	0.999814	0.999607	0.999883	1.000000

El último paso consistió en agregar todos los datos recabados para obtener las frecuencias totales, en la Tabla 4.3 se muestran los resultados finales.

Tabla 4.3 - Frecuencia de pronunciación encontrada para cada uno de los Fonemas.

Fonema	Frecuencia	Fonema	Frecuencia	Fonema	Frecuencia
/i/	7.63%	/e/	13.85%	/a/	12.69%
/o/	9.37%	/u/	2.96%	/p/	2.81%
/t/	4.60%	/k/	4.07%	/b/	2.05%
/d/	5.41%	/g/	0.44%	/f/	0.78%
/s/	9.66%	/j/	0.87%	/ch/	0.18%
/ll/	0.45%	/m/	2.66%	/n/	7.21%
/ñ/	0.16%	/r/	6.59%	/rr/	0.18%
/l/	5.37%				

4.3 Validación con Trabajos Previos

Los resultados obtenidos en la distribución fonética se contrastaron con lo reportado en [54, 55] para el español latinoamericano, el análisis de correlación [56, 57] entre lo encontrado y esos trabajos se muestra en la Tabla 2. Se observa que la correlación es alta, por lo que los resultados obtenidos en este trabajo como los anteriormente publicados son coincidentes.

Tabla 4.4 - Coeficiente de correlación con trabajos previos.

Investigación previa	Coeficiente de correlación
Villaseñor-Pineda (2003)	0.9960
Pérez (2003)	0.9974

4.4 Resumen y Conclusiones del Capítulo

En este capítulo se establece el primer requisito la construcción de un corpus de voz orientado al reconocimiento de locutor, el cual consiste en grabar frases que se apeguen a la distribución fonética del idioma objetivo, en este caso Español Mexicano. Para tal fin, se realizó un análisis de la distribución de fonemas contenidos en textos de periódicos de todo México publicados en línea en la Web. El resultado se contrastó con los trabajos previos encontrados y se concluye que lo obtenido es congruente con estos trabajos, por lo que se establece la distribución fonética de las frases que se tienen que definir para las grabaciones del corpus. Como resultado adicional se

observa que es posible obtener la distribución fonética del español mexicano analizando texto de páginas web de periódicos locales de varias regiones de México y se comprobó que la distribución fonética, obtenida por este medio, no varía entre las diferentes regiones en las que se dividió el conjunto de fuentes de datos.

Capítulo 5

Corpus de Voz en Español Mexicano

5 Corpus de Voz en Español Mexicano (VoCMex)

Para propósitos del presente trabajo, los corpora de voz para experimentación se pueden clasificar en dos tipos: los orientados a RAH y los orientados a RAL. Los primeros deben permitir analizar las diferencias fonéticas del idioma independientemente del locutor del cual provenga la señal, para lograr esto es necesario grabar las mismas frases para una gran cantidad de locutores, no es tan importante el número de sesiones por locutor, lo que importa es caracterizar los fonemas para un amplio rango de locutores. Por otro lado, los corpora para RAL deben ser multi-sesión para que el sistema pueda discriminar los elementos de variabilidad intra-locutor a lo largo del tiempo y resaltar la variabilidad interlocutor para poder caracterizar a las personas. Esto impone una mayor complejidad en la construcción de este tipo de corpora porque es necesario llevar un seguimiento del locutor a lo largo de la etapa de grabación, debido al tiempo que debe transcurrir entre una sesión y otra, el cual puede ser de semanas a meses.

En el presente capítulo se describe la construcción del corpus VoCMex, el cual consta de grabaciones multi-sesión de frases apegadas a la distribución fonética del Español Mexicano descrita en el capítulo 4.

5.1 Corpora para RAL en Idiomas Distintos al Español

Aunque el interés del presente trabajo de investigación está centrado en el español mexicano, es necesario revisar las bases de datos en otros idiomas, con la finalidad de conocer sus características. En esta sección se describen los corpora en diversos idiomas, que se consideraron de mayor relevancia debido a la aparición constante en la literatura consultada.

5.1.1 Polycost

Frases en inglés de 134 locutores de 13 países europeos (de habla inglesa y habla no inglesa), consta de más de 5 sesiones por cada locutor grabadas por medio de aparatos telefónicos a través de líneas digitales ISDN [58].

5.1.2 YOHO

Diseñado para evaluar sistemas de verificación de locutor en situaciones dependientes de texto. Consiste de 138 locutores (106 hombres, 32 mujeres) que grabaron frases compuestas de dígitos en idioma inglés [51].

5.1.3 Switchboard I-II

Un corpus en inglés con más de 500 locutores, las grabaciones fueron realizadas mediante un sistema telefónico automático que conectaba a un participante con otro y grababa la conversación. Un subconjunto de este corpus es utilizado en las evaluaciones de reconocimiento de locutor que regularmente organiza el Instituto Nacional de Estándares y Tecnología de Estados Unidos (NIST) [59].

5.1.4 SIVA

Base de señales en italiano, contiene 691 locutores (335 hombres, 336 mujeres), grabado en múltiples sesiones utilizando líneas telefónicas [52, 58].

5.1.5 XM2VTSDB

Corpus multimodal que incluye cerca de 300 locutores ingleses en grabaciones de audio y video a lo largo de cuatro sesiones. Es un producto en constante evolución cuyos orígenes fueron M2VTS en francés y XM2VTS en inglés [60].

5.2 Corpora para RAL en el Idioma Español

Siendo el idioma español el interés primario de la tesis, se realizó una búsqueda de corpora para reconocimiento en las diversas variaciones de este idioma, sin embargo lo encontrado fue escaso, tal como se describe a continuación.

5.2.1 AHUMADA.

Corpus en Español Ibérico, consiste de 144 locutores hombres. Fue desarrollado en el contexto de un proyecto de investigación forense [48]. Fue el único corpus encontrado para RAL, que ha sido desarrollado en un país de habla hispana. Otro corpus en Español Ibérico encontrado en la literatura es el denominado Albayzin, pero está orientado al reconocimiento del habla [53, 61].

5.2.2 Corpora de Voz en Español Mexicano

De acuerdo al alcance de las fuentes de información consultadas en esta investigación, no se encontró un corpus para RAL que haya sido desarrollado en México, todas las bases de datos encontradas fueron construidas para experimentación en reconocimiento del habla.

5.2.2.1 DIMEX-100

Desarrollado en el Instituto de Investigaciones en Matemáticas Aplicadas de la Universidad Nacional Autónoma de México (IIMAS-UNAM). Es un corpus en español mexicano orientado al reconocimiento del habla, incluye grabaciones de 100 locutores del centro del México, incluye un estudio detallado de los fonemas y sus correspondientes alófonos para el español mexicano [49].

5.2.2.2 TLATOA

Desarrollado por la Universidad de las Américas en Puebla (UDLA). Consiste de grabaciones de 550 adultos principalmente del centro de México, está orientado a experimentación con reconocimiento del habla [62].

Una vez revisada las fuentes bibliográficas y al no haber encontrado un corpus con las características deseadas para el reconocimiento de locutor, se refuerza la idea de construirlo. Los siguientes apartados muestran la metodología realizadas para tal efecto.

5.3 Desarrollo de VoCMex

Para construir el corpus se llevaron a cabo las siguientes actividades:

1. Definición de frases a grabar
2. Definición del protocolo de grabación
3. Habilitación de la infraestructura, física y de aplicaciones de cómputo.
4. Reclutamiento de locutores
5. Realización de las sesiones de grabación
6. Revisión y tratamiento de las grabaciones
7. Evaluación del corpus

En las siguientes secciones se describen los aspectos más relevantes de las actividades anteriormente citadas.

5.3.1 Definición de las frases fonéticamente equilibradas

Para definir las frases a grabar en el corpus, se siguió el enfoque utilizado por el corpus Ahumada, el cual consta de frases de 10 a 12 palabras. Sin embargo, con el objetivo de poder experimentar con frases de diferentes tamaños, se optó por frases de mayor longitud. Para formar cada una de estas frases se eligió un conjunto de palabras que se aproximaran a la distribución

fonética, a estas se les llamó *frases candidatas*. Una vez que se definieron varias frases candidatas, se calculó y comparó el coeficiente de correlación de cada una de ellas con respecto a la distribución fonética del idioma. Las tres frases con el mayor coeficiente fueron las elegidas para el corpus. Dichas frases son las siguientes:

Frase 1: "El teatro San Isidro no pudo caer en manos de Isabel y Kazel"

Frase 2: "Primero en casa del Jaime, Tina y Rafael, cocinan Celeste y don Arturo, después, todos van al bosque"

Frase 3: "Dentro de la escuela de Fátima, Yanet, Piter, Cris, Tom y Cabino planeaban hacer un juego de carreras, sólo que necesitaban a muchos niños, por lo que deciden primero ir al salón a decirles a todos"

Aun y cuando las frases por incluyen palabras que no necesariamente son comunes en el idioma, el objetivo principal es que satisfagan la frecuencia de los fonemas. La Tabla 5.1 muestra la correlación de la frecuencia porcentual de cada una de las frases anteriores con la del español mexicano y se aprecia que las tres cumplen con el requisito de ser fonéticamente balanceadas.

Tabla 5.1 - Coeficientes de correlación de cada frase.

	Frase 1	Frase 2	Frase2
Villaseñor-Pineda et al. (2003)	0.9817	0.9857	0.9024
Pérez (2003)	0.9759	0.9945	0.9100
Propia	0.9905	0.9941	0.9903

5.3.2 Protocolo de Grabación

Una vez construidas las frases se definió el protocolo de grabación, el cual incluyó las siguientes características:

- Por cada locutor, grabar 3 frases fonéticamente equilibradas (4, 6 y 10) segundos respectivamente y un texto adicional no balanceado (aprox. 60 segundos).

- Realizar tres sesiones por locutor.
- Utilizar dos tipos de dispositivo de adquisición: micrófono y aparato telefónico.
- Tres modalidades: sistema telefónico analógico, grabación directa de micrófono y sistema telefónico de VoIP.

5.3.3 Software y Hardware utilizado

Para las grabaciones telefónicas se configuró el software Asterisk [63-65], el cual es un servidor de comunicaciones de voz para líneas telefónicas analógicas y líneas de VoIP, bajo el sistema operativo Linux. El hardware utilizado fue una computadora con procesador Pentium 4 de 3GHz de velocidad, 2GB de RAM, una tarjeta de puertos telefónicos Digium, modelo 1TDM422EF. Las grabaciones de VoIP se realizaron a través de una red local de datos inalámbrica 803.11g utilizando el *softphone Zoiper* y un micrófono tipo cardiode. Para las grabaciones analógicas y las de VoIP se configuró un sistema de contestación automática en el servidor Asterisk, el cual consistió de un menú interactivo para establecer el número de locutor, tipo de medio, número de frase y número de sesión. Todas las grabaciones telefónicas se almacenaron en el sistema de archivos del servidor Linux. Para las grabaciones directas se usó un micrófono de escritorio tipo cardiode conectado por puerto USB a una computadora Pentium 4 de 3GHz, 2GBde RAM con sistema operativo Windows XP SP3 y el software Audacity para procesar las señales de audio. Una descripción más detallada del hardware y software utilizados se puede consultar en el Anexo D.

5.3.4 Sesiones de Grabación

Todas las grabaciones se realizaron en un cuarto especialmente acondicionado para tal fin, con dimensiones de 3x3m y las paredes cubiertas con espuma fonoabsortora de poliuretano con cuñas anecoicas para disminuir la reverberación acústica. Las sesiones se llevaron a cabo a lo

largo de 18 meses, participaron 50 personas, 31 hombres y 19 mujeres. 23 hombres completaron dos sesiones y 18 completaron las tres sesiones; en el caso de las mujeres, 14 completaron dos sesiones y 13 completaron las 3 sesiones. En total el corpus contiene 31 locutores con las tres sesiones completas. El tiempo transcurrido entre sesiones varía de 1 a 5 semanas. El tiempo promedio que cada locutor le dedicó por sesión fue de 20 minutos.

5.3.5 Organización del Corpus

El corpus está organizado por directorios, uno para cada locutor, el nombre del directorio es el identificador del locutor, *e.g.* L01 indica al locutor 01. Cada archivo del corpus se nombró utilizando la siguiente nomenclatura que se muestra en la Tabla 5.2

Tabla 5.2 - Formato de identificación de cada archivo del corpus.

L<ID de locutor><Tipo de Grabación>F<ID de Frase>S<ID de Sesión>

Descripción de cada campo:

<ID de locutor>	Número consecutivo asignado a cada locutor, con fines de seguimiento e identificación en el corpus.
<Tipo de Grabación >	Identificador del tipo de adquisición de la señal, MI para micrófono, T1 para teléfono analógico, T2 para VoIP.
<ID de Frase >	Identificador de la frase, puede tomar los valores de 1, 2 , 3 ó 4.
<ID de Sesión >	Identificador de la sesión en que fue grabada esa señal, puede ser 1, 2 ó 3.

Por ejemplo, el archivo *L01MIF2S3* corresponde a la grabación directa de micrófono del locutor 01, frase 2, sesión 3.

5.4 Evaluación del Corpus

Para la evaluación del corpus se implementó un sistema de IAL basado en modelos GMM [20], utilizando la plataforma expuesta en el capítulo 3, cuya arquitectura se muestra en la Figura 5.1.

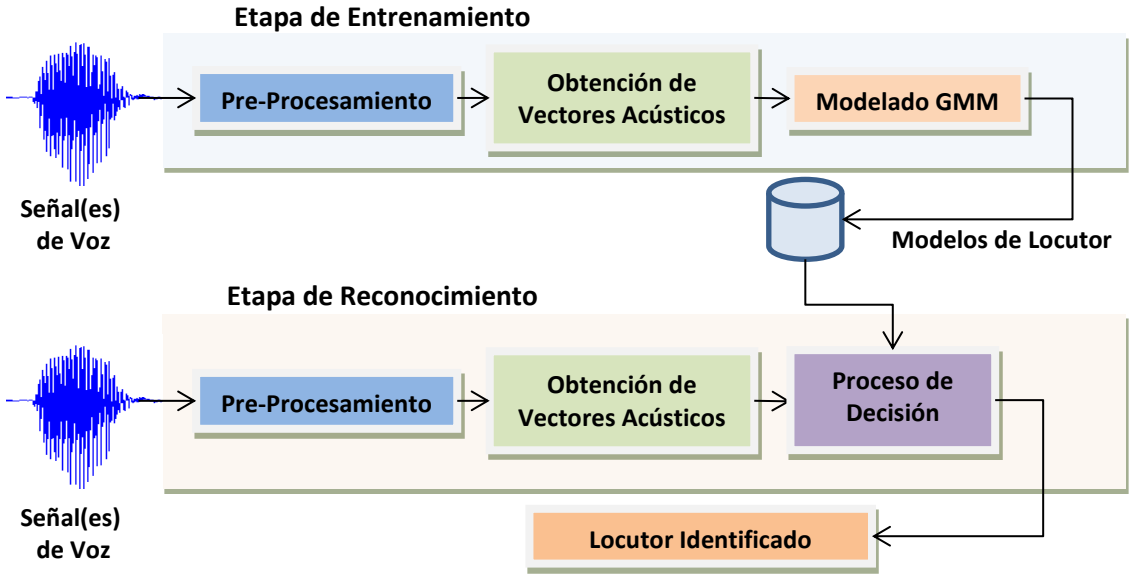

Figura 5.1 – Arquitectura del Sistema de IAL utilizado para la evaluación de VoCMex.

Para la construcción de vectores acústicos se utilizó una ventana de 30 ms con un traslape de 10 ms, se aplicó un banco de filtros de 26 canales para obtener un vector de 41 elementos (13 coeficientes MFCC, 13 valores de la primera derivada, 13 de la segunda derivada y 2 coeficientes de energía) además se aplicó sustracción de medias cepstrales (CMS) para atenuar el ruido del medio. Para el modelado de locutores primero se construyó un modelo impostor con 32 mezclados gaussianos con matriz de covarianza diagonal, el entrenamiento se llevó a cabo con el algoritmo Expectación-Maximización (EM); cada modelo cliente se obtuvo mediante adaptación máxima a posteriori (MAP).

5.4.1 Evaluación

El corpus se dividió en dos particiones (o *conjuntos*): uno de entrenamiento y otro para prueba, quedando de la siguiente forma:

- Frases 1, 2 y 3 de las sesiones 1 y 2 (aprox. 40 segundos de señal).
- Frase 4 de la sesión 3 (aprox. 60 segundos de señal).

La evaluación del corpus se realizó sólo con las señales de los locutores hombres, con la finalidad de poder comparar los resultados con otros corpora disponibles, los cuales sólo contienen señales de locutores masculinos.

El resultado de las pruebas se muestra en la Tabla 5.3. El mejor reconocimiento se obtuvo con las grabaciones de micrófono, tal y como se esperaba, ya que este tipo de datos son los que presentan mayor calidad en términos del medio de adquisición; en segundo lugar los de VoIP, los cuales también fueron adquiridos por medio del mismo micrófono. Por otro lado, el menor reconocimiento corresponde a los datos adquiridos con un aparato telefónico, como era de esperarse, ya que este tipo de medio es el que tiene menor calidad de captura.

Tabla 5.3 - Resultados de la evaluación del corpus.

	Micrófono	Teléfono	VoIP
% Reconocimiento	95.24	71.43	90.48
% Identificación errónea	0.00	0.00	0.00
% Falsos Positivos	0.00	28.57	9.42
% Falsos Negativos	4.76	0.00	0.00
TOTAL	100.00	100.00	100

5.4.2 Comparación con AHUMADA

Además de la evaluación anteriormente presentada, se realizó otra con el mismo sistema de identificación pero utilizando los datos del corpus Ahumada para micrófono y teléfono ya que este corpus no contiene señales de VoIP. Las particiones del conjunto de datos para AHUMADA se eligieron de manera semejante a las utilizadas con nuestro corpus:

1. Diez frases fonéticamente equilibradas grabadas en dos sesiones distintas (Aprox. 40s de señal).

2. Lectura de un texto de una tercera sesión, un texto distinto al usado en el entrenamiento (Aprox. 60s de señal). De la misma manera que con nuestro corpus, sólo se utilizaron señales de locutores masculinos.

La comparación de resultados se presenta en la Tabla 5.4. Se observa que para el caso de la señal de micrófono, el reconocimiento de los datos de AHUMADA fue ligeramente mayor, mientras que nuestra señal telefónica obtuvo un mejor reconocimiento.

Una posible explicación para la gran diferencia en el reconocimiento de la señal telefónica es que las condiciones de grabación en nuestro caso fueron más cuidadas que lo reportado por [48] y otra posibilidad radica en la diferencia de calidad del aparato telefónico.

Tabla 5.4 - Comparación de resultados de reconocimiento usando el corpus AHUMADA.

	Micrófono	Micrófono AHUMADA	Teléfono	Teléfono AHUMADA
% Reconocimiento	95.24	100.00	71.43	47.62
% Identificación errónea	0.00	0.00	0.00	4.76
% Falsos Positivos	0.00	0.00	28.57	33.33
% Falsos Negativos	4.76	0.00	0.00	14.29
TOTAL	100.00	100.00	100.00	100

5.4.3 Experiencias Obtenidas

Durante la etapa de grabación se identificaron dos aspectos a considerar en los proyectos de este tipo:

Cuando se construyeron las frases fonéticamente equilibradas, no se consideró importante que fueran frases coherentes, lo único que se buscaba es que cumplieran con la distribución de frecuencia de fonemas, se trató de seguir la idea de las frases de XM2VTSDB [60], las cuales son una secuencia de palabras sin significado. Sin embargo los locutores tendían a confundirse fácilmente en las grabaciones, por lo que se optó por cambiar las frases por otras más coherentes, esto disminuyó el número de errores que el locutor cometió por sesión y por lo tanto disminuyó el tiempo promedio por sesión.

Dado que los participantes no tienen claro el objetivo de las grabaciones, es necesario estar constantemente explicando los alcances de la investigación, ya que sienten cierta incomodidad al

no tener certeza del uso que se les dará a sus datos. La certeza es indispensable para conseguir de ellos las sesiones subsecuentes.

5.5 Resumen y Conclusiones del Capítulo

En este capítulo, se describió la construcción del corpus de voz en español mexicano orientado al reconocimiento de locutor. Es importante contar con una base de señales de este tipo ya que algunas tareas de reconocimiento (especialmente la clasificación), requieren de señales de voz que contengan características fonéticas específicas de los habitantes de una región geográfica dada. Una vez construido el corpus, se evaluó mediante la construcción de un sistema de IAL, cuyos resultados se contrastaron con el mismo IAL pero basado en el corpus Ahumada. Los resultados obtenidos mostraron coherencia con el tipo de material utilizado para la adquisición y lo reportado en trabajos previos.

Por otro lado, se establece la conveniencia de que las frases que se vayan a grabar tengan cierto grado de coherencia, aún y cuando expresen ideas incompletas, ya que los locutores tienden a equivocarse menos cuando las frases a pronunciar son menos complejas.

Capítulo 6

Identificación de Locutor aplicando Vectores Acústicos Cuantílicos

6 Identificación de Locutor aplicando Vectores Acústicos Cuantílicos

En la contextualización de la problemática relacionada con los sistemas de reconocimiento de locutor y en particular de identificación establecida en el capítulo 1, se hace la pregunta: ¿*En cuáles aspectos se puede mejorar la IAL en general*? Contestarla no es una tarea sencilla, ya que como se discutió en el capítulo 2, la comunidad científica continúa trabajando en los retos asociados a los diferentes niveles de información aprovechable para el RAL en general y la IAL en particular.

Una técnica escasamente encontrada en la literatura de reconocimiento automático de locutor es una que ha sido utilizada para aprovechar información concerniente a la capacidad pulmonar, a partir de medidas acústicas. Dicho enfoque ha sido aplicado con el fin de apoyar diagnósticos médicos, midiendo las tasas de flujo y volumen de la respiración con cuantiles [66-68].

En el presente capítulo se propone el uso de vectores acústicos construidos con cuantiles, los cuales son una aportación novedosa en el tratamiento de la señal de voz a nivel espectral.

6.1 Vectores Acústicos Aplicados en Voz

La entrada principal de los sistemas de reconocimiento de voz son los vectores acústicos, los cuales constituyen la representación acústica de las señales originales pero en un espacio de menor dimensión, siendo estas representaciones conformadas por parámetros que capturan la estructura fina del aparato fonador. Los principios de la construcción de vectores acústicos se sustentan en modelar ya sea el tracto vocal desde la perspectiva de la generación de voz; o el paradigma del funcionamiento del oído humano y la interpretación de la voz en el cerebro humano [12, 16]. El análisis de la señal de voz comúnmente se realiza sobre una sucesión de segmentos elementales cuasi-estacionarios llamados ventanas de análisis (o tramas); en este

proceso se toma una señal de voz, se parte en ventanas de cierto tamaño (típicamente en el orden de los milisegundos) las cuales están traslapadas entre sí, a cada una de estas ventanas se le aplica un análisis espectral con la intención de generar un vector acústico correspondiente.

Existen diferentes técnicas para la obtención de los vectores acústicos, de las cuales aquí se presentan algunas de las más recurrentes en los trabajos de investigación en el área del tratamiento de la voz. Desde el punto de vista de la aplicación, algunas fueron desarrolladas específicamente para la compresión de señales de voz, otras son más utilizadas en sistemas de reconocimiento del habla y del locutor [69].

Para fines prácticos, la señal de voz contiene dos elementos de información: la fuente vocal y la transformación efectuada por el conducto vocal (la cual se asume como lineal). Para separar estos dos elementos de información es necesario efectuar una deconvolución a posteriori de la señal, esta operación debe permitir conocer la contribución de las cuerdas vocales y la del conducto vocal durante la generación de voz, esta contribución es la que se utiliza como entrada en los sistemas automáticos de reconocimiento. Una aproximación exitosa de este enfoque es la denominada deconvolución Cepstral [16], la cual permite aislar las frecuencias fundamentales de la señal de voz, de aquellas que son generadas por el conducto vocal. Esto se obtiene aplicando la transformada discreta cosenoidal (cosine discrete transform, CDT) a los vectores espectrales previamente calculados mediante la transformada rápida de Fourier (FFT):

$$c_n = \sum_{k=1}^{K} S_k . \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], n = 1,2,\ldots,L \qquad \textbf{Ecuación } 6.1$$

Siendo K el número de coeficientes espectrales, S_k los coeficientes espectrales y L el número de coeficientes cepstrales que se quieren calcular, $(L \leq K)$.

Una extensión de los principios cepstrales y su paso a un espacio frecuencial no lineal relacionado con la audición humana y muy exitoso son los *Coeficientes Cepstrales en*

Frecuencia de Mel (MFCC) [8, 16, 70, 71]. El interés del análisis MFCC se origina en su similitud con el funcionamiento del sentido humano de la audición. La naturaleza logarítmica de la técnica es coherente ya que el sistema auditivo humano percibe los sonidos en forma logarítmica sobre un intervalo de frecuencias, dichas distribución de frecuencias se muestra en la Figura 6.1.

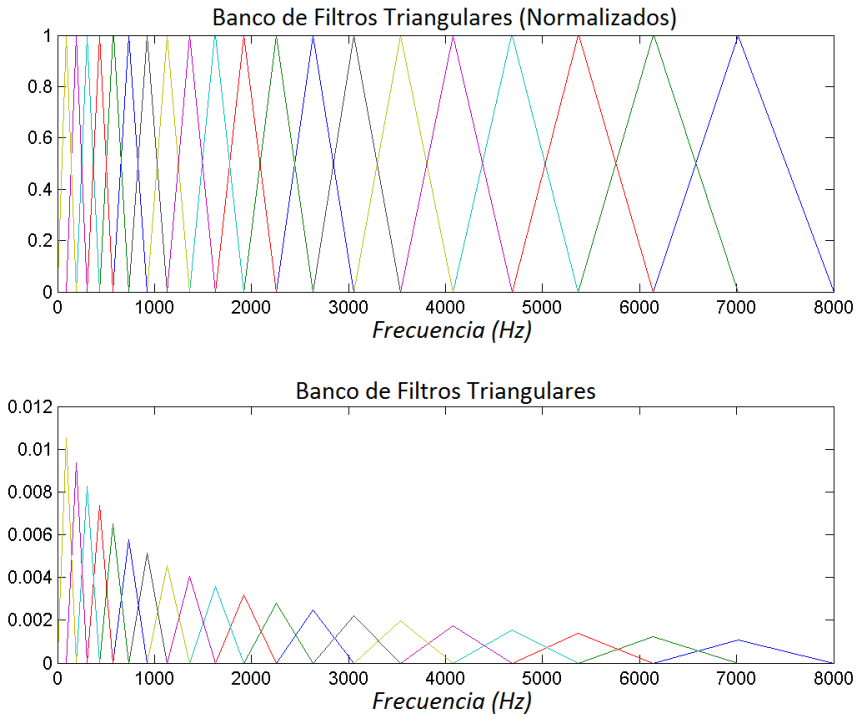

Figura 6.1 – Distribución de las frecuencias en la escala Mel.

Una variante de los vectores MFCC son los *Coeficientes Cepstrales en Frecuencia Lineal (LFCC)*, donde a diferencia de los anteriores, las frecuencias de los filtros son repartidas uniformemente sobre una escala lineal de frecuencias y no sobre la escala Mel [16, 72].

Una aproximación muy difundida en el ámbito de transmisión de voz y sistemas de telecomunicación sobre IP, es la *Codificación Lineal Predictiva (LPC)*. Este método se basa principalmente sobre la hipótesis de que la voz puede ser modelada por un proceso lineal predictivo [12, 73-76]:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + e(n) \qquad \textbf{Ecuación 6.2}$$

Donde *s(n)* representa la predicción de la señal en un instante *n* a partir de las *p-muestras* precedentes. La voz no es, sin embargo, un proceso lineal perfecto, el valor que constituye la suma ponderada sobre *p* pasos de tiempo introduce un error que es necesario corregir con la introducción de un término *e(n)*. La codificación por predicción lineal consiste entonces en determinar los coeficientes a_k que minimizan el error *e(n)*, esto en función de un conjunto de señales. Estos coeficientes de predicción son utilizados como parámetros acústicos. El método de codificación por predicción lineal es utilizado tanto en sistemas de reconocimiento como en sistemas de transmisión de voz como la radio, teléfono o las redes de datos IP.

Existen otros métodos como *PLP*, que modifican el espectro de potencia de la voz antes de obtener una aproximación por un modelo auto-regresivo [77]. Además, hay otras variantes de esta metodología que son más adaptadas a un canal de comunicación tales como *Relative Spectral PLP* (RASTA PLP) [76, 78].

6.2 Antecedentes en Vectores basados en Cuantiles

Con la finalidad de investigar si aplicando modelos acústicos en sonidos pulmonares se podían identificar pacientes y monitorearlos, se desarrollaron algunos trabajos [66-68]. Viendo estos experimentos desde el punto de vista del reconocimiento de locutor, es evidente que los cambios en la estructura fisiológica influyen en el tracto respiratorio, tracto vocal y por consecuencia en la voz. Considerando estas evidencias se propone analizar señales de voz pero con un tratamiento basado en vectores cuantílicos, con la finalidad de distinguir elementos que permitan caracterizar a los locutores y por lo tanto reconocerlos. El uso de los vectores es motivado por el interés de

identificar los valores frecuenciales relevantes en un vector acústico dada la densidad de energía presente y su relación con las formantes de la señal.

6.2.1 Cuantiles

En teoría estadística, las medidas de tendencia no centralistas permiten conocer otros aspectos particulares de una distribución (como pudiera ser el caso de vectores acústicos). Dentro de estas medidas, unas de las más importantes son los *cuantiles*, en estas variables los datos son ordenados de forma creciente, dividiendo la función de distribución en partes, de tal forma que cada una de estas partes contiene la misma cantidad de área. Los cuantiles son puntos tomados a intervalos regulares de la *Función de Distribución Acumulativa (CDF, del inglés Cumulative Distribution Function)* de una variable aleatoria. Dividiendo los datos ordenados en intervalos de igual tamaño; los cuantiles son los valores de los datos que marcan el límite entre subconjuntos de datos consecutivos.

Los cuantiles se basan en la CDF, la cual juega un papel muy importante en teoría estadística. El cuantil q_p de una variable aleatoria está definido como el número q más pequeño tal que la función de distribución acumulativa es mayor o igual a algún valor p, donde p se encuentra entre $0 < p < 1$. Esto puede ser calculado para el caso de una función de distribución continua con su función de densidad $f(x)$ resolviendo:

$$p = \int_{-\infty}^{q_p} f(x)dx \qquad \textbf{Ecuación } 6.3$$

Para nuestros propósitos lo que se desea es encontrar q_p, para lo cual nos valemos de la transformada inversa de la CDF.

$$q_p = F^{-1}(p) \qquad \textbf{Ecuación } 6.4$$

Estableciéndolo de otra manera, el *p-ésimo* cuantil q_p de una variable aleatoria **X**, es el valor tal que:

$$F(q_p) = P(X \leq q_p) = p; 0 < p < 1 \qquad \textbf{Ecuación } 6.5$$

Un caso particular de los cuantiles, son los *cuartiles*. El objetivo de los cuartiles es conocer los valores que toma una variable aleatoria cuando su probabilidad corresponde a un 25% , 50% y 75% de su CFD, dada su *Función de Densidad de Probabilidad (PDF, del inglés Probability Density Function)*. Se les denomina cuartiles debido a que dividen el área de una función de distribución de probabilidad en cuatro segmentos cuya área bajo la curva es igual. En la Figura 6.2 se muestra este concepto.

Los cuartiles pueden representarse por la siguiente notación: $q_{0.25}$, $q_{0.5}$, *y* $q_{0.75}$, respectivamente. En esencia, estos dividen la distribución en cuatro segmentos de igual probabilidad bajo la curva. Un caso especial es el segundo cuartil, conocido como la *mediana*, es uno de los cuantiles más utilizados en la estadística, el cual satisface la siguiente ecuación:

$$0.5 = \int_{-\infty}^{q_{0.5}} f(x)dx \qquad \textbf{Ecuación } 6.6$$

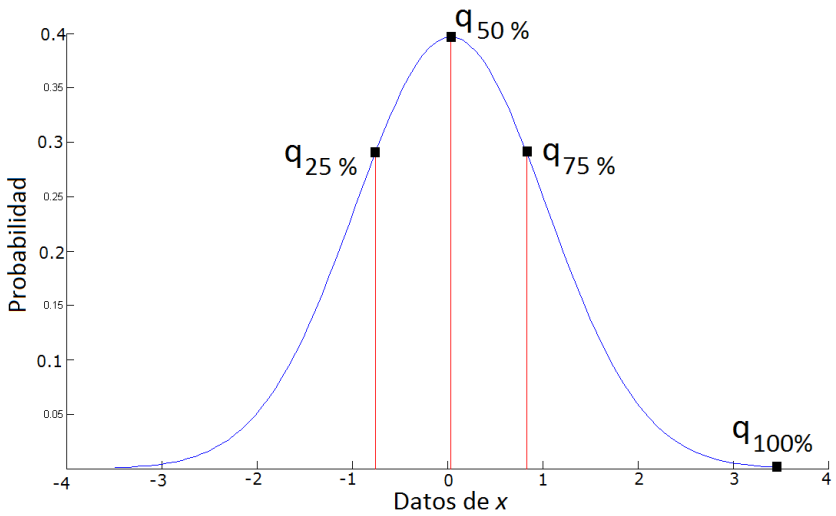

Figura 6.2 – Representación gráfica del concepto de cuartiles.

Otra alternativa para obtener un estimado de los cuantiles está basada en la función de distribución empírica [56]. Si $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denota el orden estadístico para una muestra aleatoria de tamaño n, entonces $X_{(j)}$ es un estimado del cuantil $(j - 0.5)/n$:

$$X_{(j)} \approx F^{-1}\left(\frac{j - 0.5}{n}\right) \quad \textbf{Ecuación 6.7}$$

La teoría estadística no limita al uso exclusivo de cuartiles, es decir, no se está limitado a valores de 0.5. En general, se puede estimar el *p-ésimo cuantil* aplicando la siguiente ecuación:

$$q_p = X_{(j)} \frac{j-1}{n} < p \leq \frac{j}{n} j = 1, \ldots, n \quad \textbf{Ecuación 6.8}$$

Un cuartil, decil o en lo general un *percentil* (como también son conocidos los cuantiles) son conceptos ligados a la PDF. Estas son utilizadas en estadística con el objetivo de conocer cómo se distribuyen las probabilidades en un evento, en relación a un suceso. Un requisito indispensable que debe cumplir una PDF es que su área total encerrada bajo la curva debe ser igual a 1. La probabilidad de que una variable *x* tome un valor a lo largo de la curva, es igual al área bajo la curva de dicha PDF. Matemáticamente esto se define como:

$$F(x) = \int_{-\infty}^{x} f(u)du \quad \textbf{Ecuación 6.9}$$

Una de las PDF's más conocidas es la *Distribución Gaussiana* o *Distribución Normal*. De aquí se desprende el concepto de función de distribución empírica, que para una variable aleatoria continua está dada por:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \quad \textbf{Ecuación 6.10}$$

Donde el valor cuantílico de interés x; que para una variable aleatoria discreta está dada por:

$$F(a) = \sum_{x_i \leq a} f(x_i)$$ **Ecuación** 6.11

Cuando no es conveniente suponer una distribución específica para la variable aleatoria, entonces podemos aplicar la CDF, siendo el principio a partir del cual se obtiene histogramas de frecuencias de los vectores acústicos, sobre los cuales se pueden ajustar modelos gaussianos, como se presenta más adelante en la Sección 6.3 de modelado con GMM.

6.2.2 Vectores Acústicos basados en Cuantiles

Lo que se propone en el presente trabajo es, a partir de los datos de la señal de voz **X** obtener su *Transformada Rápida de Fourier* (*FFT*), normalizarla y tomarla como la función de distribución de frecuencias o una función de densidad de frecuencias, de esta forma se puede obtener un vector $Q = (q_1 , q_2 , \ldots , q_n)$ sobre los valores frecuenciales, donde cada $q_i \in Q$ es el valor de la frecuencia asociado al valor porcentual acumulado bajo la CDF específico del cuantil *i*; en otras palabras, *Q* representará un vector cuantílico para la señal *X*. De esta forma podemos hablar de vectores cuartílicos, cuando sus elementos representan los valores frecuenciales para el 25%, 50% y 75% del área bajo la curva normalizada de la FFT de *X*; o vectores octílicos en el caso de obtener estos valores para 12.5%, 25%, 37.5%, 50%, 62.5%, 75% y 87.5% del área.

Las etapas del proceso para la construcción de vectores cuantílicos son las siguientes:

a) Se lee la señal (archivo de audio), la cual está en el dominio de las amplitudes (Figura 6.3)

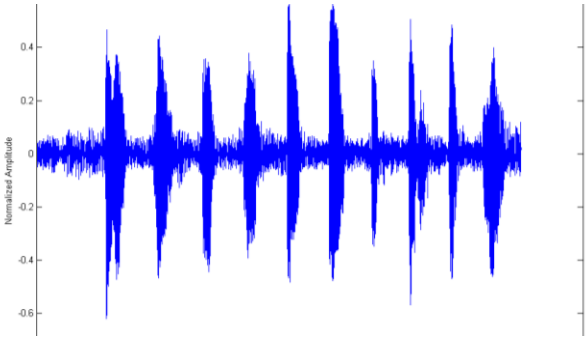

Figura 6.3 – Señal de voz en representación Tiempo vs Amplitud.

b) Aplicar la FFT, con lo que se obtiene su representación en el dominio de las frecuencias (ver la Figura 6.4)

c) Se normaliza el área de la FFT para cumplir con el principio básico de una PDF, para lo cual se utiliza la siguiente ecuación:

$$F_N(f) = \frac{\int_{-\infty}^{\infty} f(t) e^{-j2\pi f t}\, dt}{area(\, F(f)\,)} \qquad \textbf{Ecuación } 6.12$$

La ecuación anterior garantiza que la suma de la distribución de valores frecuenciales obtenidos a partir de la FFT será igual a 1

d) Por último se buscan los valores para los cuales se cumplen los cuantiles *($q_1, \ldots q_n$)*, de la siguiente forma:

$$A = q1 = \int_{-\infty}^{f_{q1}} F_N(f)\, df \quad , \ldots , \quad A = qn = \int_{-\infty}^{f_{qn}} F_N(f)\, df \qquad \textbf{Ecuación } 6.13$$

Algorítmicamente, esto se calcula mediante una sumatoria para obtener el área y detectando los valores frecuenciales para los cuales el área es $A = q_1, \ldots A = q_n$. Un aspecto a remarcar, es que calcular el último cuantil no importante en un proceso normalizado, pues siempre será igual a 1.

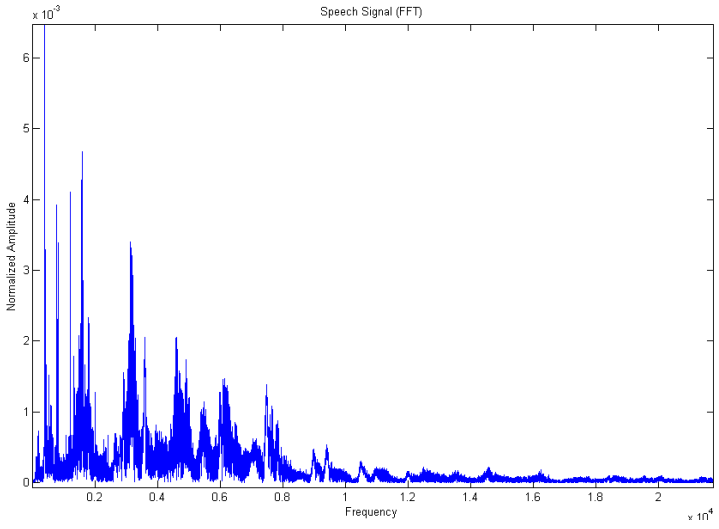

Figura 6.4 – Señal de voz en representación Frecuencia vs Amplitud.

El vector cuantílico se tomará como el vector acústico que representa un segmento, dígase estacionario, de la señal original **X** para efectos del modelado del locutor. Bajo esta idea los vectores se pueden obtener analizando la señal de voz de un locutor en dos modalidades diferentes: en tiempo largo y en tiempo corto.

El análisis en tiempo largo consiste en obtener un sólo vector cuantílico resultado de procesar la señal de un registro completo de entrada, ver Figura 6.5; por otra parte para el análisis en tiempo tiempo corto, es necesario definir ventanas de procesamiento de cierta longitud de tiempo w (regularmente en el orden de los milisegundos y dentro del rango estacionario de la señal) de tal forma que por cada ventana se genera un vector. Además, se establece un valor de traslape o < w, de tal forma que el análisis se va realizando por tramas traslapadas de señal, con la finalidad de hacer la extracción de características en segmentos estacionarios. Así, en el análisis de tiempo corto se tendrán T/(w-o) vectores para cada señal procesada, donde T es el tiempo total de duración de la señal, w es el tamaño de la ventana y o representa el tiempo de traslape. Esta modalidad se esquematiza en la Figura 6.6.

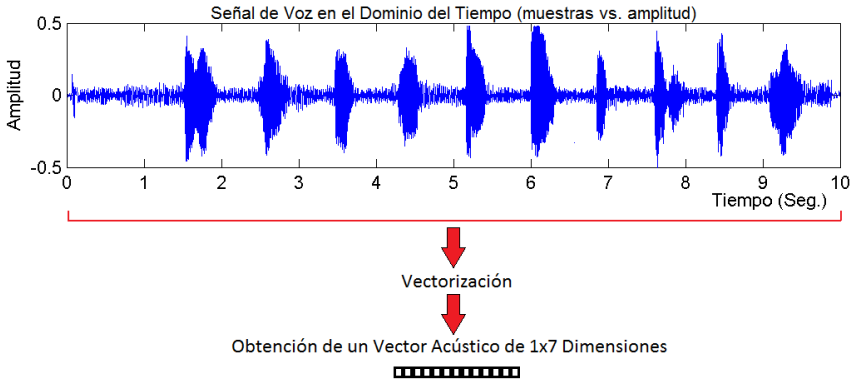

Figura 6.5 – Vectores en Tiempo Largo.

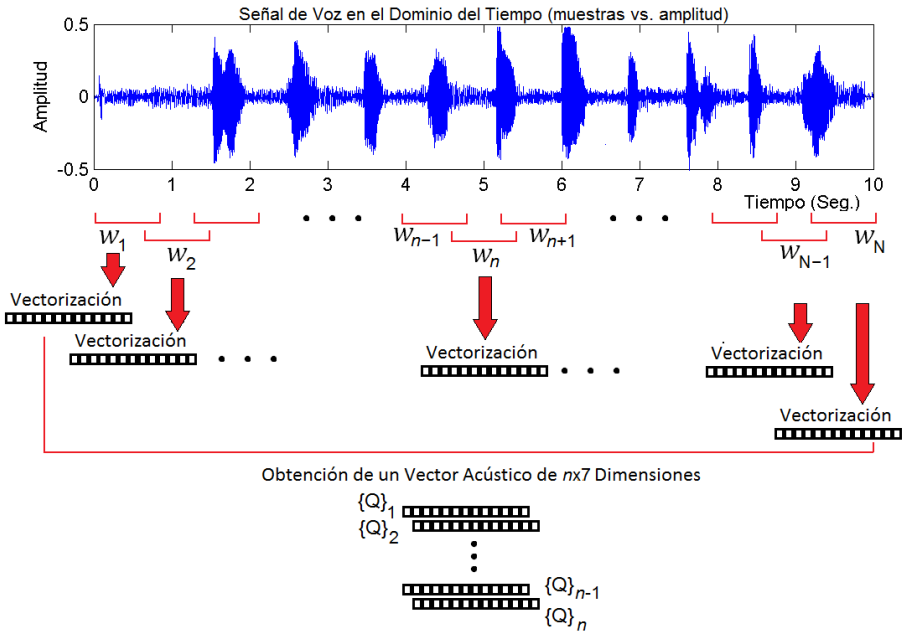

Figura 6.6 – Vectores en Tiempo Corto.

6.3 Modelado GMM para Locutor

Existen dos tipos de problemas en clasificación, el aprendizaje supervisado y el no supervisado. En *aprendizaje supervisado*, se tienen datos de los cuales se conoce la clasificación de cada observación, es decir, cada caso o dato tiene una etiqueta asociado a él. Se construye un clasificador basado en este conjunto de datos, el cual es usado para clasificar futuras observaciones. En caso contrario, es decir que no se tienen las etiquetas, membresía de los datos, observaciones o que no se conozca la cantidad de clases que representan los datos, se denomina aprendizaje *no supervisado*. Además, en reconocimiento de patrones y/o reconocimiento de formas existen diversas aproximaciones y una de las más extendidas es la aproximación estadística. Por lo tanto, para entender estas ideas es importante partir de algunos conceptos fundamentales en procesos aleatorios, como la concepción misma de una función de densidad de probabilidad (PDF).

Para crear un modelo correcto es importante cumplir con algunos axiomas de PDF, como lo marca la siguiente ecuación:

$$\int p(x)dx = 1 \qquad \textbf{Ecuación } 6.14$$

Asimismo, es necesario conocer el cálculo de los parámetros de un modelo, particularmente los valores de expectación, es decir media y covarianza (o varianza, cuando es el caso de una dimensión). Valor esperado, también conocido como media. Otros conceptos trascendentales dentro del procesamiento aleatorio son las ideas de *probabilidad a priori*, *probabilidad a posteriori* y *probabilidad condicional de clase*. Un teorema trascendental que establece una relación fundamental entre ellas es el teorema de Bayes.

La idea es crear una regla de decisión o clasificador, que tome un vector ***x*** de características cuya membresía de clase es desconocida y que regrese la clase más probable a la que pertenezca. Una forma lógica es asignándole la etiqueta al vector de características que corresponda a la probabilidad a posteriori más alta dada por:

$$P(\omega_j|\, x) = \frac{P(\omega_j)P(x|\omega_j)}{P(x)} \qquad \textbf{Ecuación } 6.15$$

Donde:

$$P(x) = \sum_{j=1}^{J} P(\omega_j)\, P(x|\omega_j) \qquad \textbf{Ecuación } 6.16$$

Esto representa la probabilidad de que el caso pertenezca a la *j-ésima* clase dado un vector observado ***x*** de características. Para usar esta regla, evaluamos todas las ***J*** probabilidades a posteriori, y seleccionamos la clase con más alta probabilidad. Aquí, es necesario conocer las *probabilidades a priori* de cada clase *j* dada por:

$$P(\omega_j)\; para, j = 1, \dots, J, \qquad \textbf{Ecuación } 6.17$$

Y la *probabilidad condicional-clase* (algunas veces llamada *probabilidad condicional-estado*):

$$P(x|\omega_j), para, j = 1, \dots, J, \qquad \textbf{Ecuación } 6.18$$

La probabilidad *condicional-clase* representa la distribución de probabilidades de características para cada clase. La probabilidad *a priori* significa nuestro grado inicial de creencia de que un conjunto observado de características corresponda al caso de la *j-ésima* clase. El proceso de calcular estos dos tipos de probabilidad es la forma de construir un clasificador. En nuestro caso el clasificador estará determinado por una suma ponderada de densidades gaussianas o GMM, como se muestra en la siguiente ecuación:

$$P(\vec{x}|\lambda) = \sum_{i=1}^{M} m_i b_i(\vec{x}) \qquad \textbf{Ecuación } 6.19$$

Donde x es un vector aleatorio D-dimensional, $b_i(x)$ para $i=1,\dots,M$ son las densidades gaussianas utilizadas para los modelos, como se muestra en la siguiente ecuación:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(\vec{x} - \overrightarrow{\mu_i}) \sum_i^{-1} (\vec{x} - \overrightarrow{\mu_i})\right\} \qquad \textbf{Ecuación } 6.20$$

Con μ_i representando el vector media y Σ_i la matriz de covarianza. El modelo que representa a un locutor estará dado por:

$$\lambda = \{m_i, \overrightarrow{\mu_i}, \Sigma_i\} \;\; i = 1, \dots, M \qquad \textbf{Ecuación } 6.21$$

Mientras que la regla de clasificación para seleccionar la mejor hipótesis o el modelo al cual se ajusta con más alta probabilidad la señal de entrada es:

$$C = arg_{1 \le i \le I} \max \; p(X|\lambda_i)p(\lambda_i) \qquad \textbf{Ecuación } 6.22$$

Si se aplica suponiendo cierta independencia estadística entre los vectores acústicos (lo cual no es estrictamente cierto, pero da buenos resultados y reduce enormemente el cálculo), conduce a:

$$\prod_{t=1}^{T} p(\overrightarrow{x_t}|\lambda_i) > \prod_{t=1}^{T} p(\overrightarrow{x_t}|\lambda_r); \quad r = 1, \dots, I \qquad \textbf{Ecuación } 6.23$$

Luego, si se aplica la regla de comparación en el dominio de los logaritmos (una operación válida, debido a que el logaritmo es una función monótona y las relaciones *mayor que* y *menor que* se conservan), esto es:

$$L(\lambda_I) = \sum_{t=1}^{T} \log p(\vec{x}|\lambda_i) \qquad \textbf{Ecuación } 6.24$$

Se obtiene el modelo para el cual la señal *X* presenta la mayor similitud. De esta forma se obtiene la referencia al modelo para el cual la señal de entrada se considera que pertenece, completando el proceso de identificación.

6.3.1 El concepto de Modelo de Mundo o UBM

El trabajo de Carey [79] introduce la noción del concepto de *Modelo de Mundo* (varias denominaciones están dadas en la literatura para el modelo de mundo: modelo genérico, modelo de mundo, o incluso modelo universal UBM), para la aproximación del modelo "no-locutor". Este modelo de mundo tiene por objetivo representar una población genérica de locutores.

La principal ventaja de esta aproximación es considerar un modelo que sea independiente de los locutores cliente. Aquí no se requiere ningún procedimiento de selección durante la fase de entrenamiento o de prueba. Sin embargo, es importante notar que la pertinencia de un modelo de mundo depende fuertemente de la composición de la población genérica. Esta última debe contener un número suficiente de locutores para cubrir un espacio acústico lo más grande posible sin sobre-representar el espacio acústico de un cliente particular. En la práctica, la población

genérica es seleccionada independientemente de la población de los locutores cliente e impostores.

Algunos de los métodos fundamentales en reconocimiento de locutor incluyen MAP de un UBM y modelado de super-vectores de máquinas de soporte vectorial (SVM) [32, 33, 80]. Ambos métodos pueden mejorarse al utilizar técnicas de normalización, como análisis de factores [81], *eigen-voice* [82], o proyección de atributos de perturbación (NAP) [83]. Muchos grupos emplean dichas técnicas con éxito como sub-sistemas en tareas del reconocimiento del habla o locutor.

El UBM es esencialmente un modelo GMM de gran tamaño que se ha entrenado para representar la distribución de las características del habla que son independientes de locutor, para todo locutor en general, y que es usado como la alternativa esperada de modelo de locutor durante una tarea de verificación. También es empleado en sistemas de conjunto abierto.

La tendencia general es usar una gran cantidad de locutores y habla, lo que genera un amplio rango de condiciones de canal/voz.

Existen distintos parámetros involucrados en el proceso de entrenamiento del UBM. Es posible clasificar esos parámetros en dos grandes categorías: 1) Parámetros del algoritmo; y 2) parámetros de datos.

Los parámetros del algoritmo son las variaciones en el proceso de entrenamiento, comúnmente incluyen el método de entrenamiento, número de iteraciones y valores de inicialización. Por otro lado, los parámetros de los datos incluyen los distintos métodos de definir los subconjuntos de datos disponibles para entrenamiento. Dichos parámetros consideran el corpus, cantidad de datos, número de datos por locutor, método de selección de locutores, métodos de selección de locutor, maneras de usar los vectores de características, balanceo de datos de acuerdo al canal, micrófono, lenguaje u otra variabilidad, entre los más comunes.

Ya que el único criterio objetivo disponible que puede usarse para medir la calidad de un UBM es el desempeño final del sistema, encontrar un mejor UBM se convierte en un reto debido a que se cae en un proceso de prueba y error continuo, haciendo que variar los parámetros mencionados para encontrar un punto óptimo de desempeño se convierta en una tarea impráctica.

La primera decisión que se tiene que tomar en relación al UBM es la cantidad de datos con los cuales se va a formar. Es común suponer que mientras más datos se utilicen, mejor sea el desempeño del sistema. UBM's con 512, 1024, 2048 o más mezclas se han construido [84], con la suposición de que representarán el espacio acústico de mundo definitivo. Sin embargo, no hay evidencia concreta de que utilizar una máxima cantidad de datos generará el mejor desempeño neto.

De acuerdo a [80], mientras la población de locutores de desarrollo se mantenga igual, una cantidad de datos pequeña es suficiente para un desempeño de sistema razonable. Esto sugiere que la variación interlocutor en los datos es más importante que la cantidad absoluta de datos por locutor.

6.4 Resultados en IAL aplicando Vectores Cuantílicos

Para demostrar la aplicación de los vectores cuantílicos en la IAL, se realizaron experimentos considerando tres corpora para efectos de comparación del desempeño. El primero fue el corpus desarrollado para esta tesis y explicado en el capítulo 5, el segundo corpus utilizado fue AHUMADA [48] y el tercero fue XM2VTS8k [60].

Para la construcción de vectores acústicos se siguieron las etapas descritas en la sección 6.2.2 para la obtención de vectores basados en cuartiles, octiles y deciles.

Una vez obtenidos los vectores cuantílicos para todos los locutores, se construyeron los modelos GMM de cada locutor con dos configuraciones. La primera utilizando un modelo

aleatorio como modelo inicial para el algoritmo Expectación-Maximización (EM). Para la segunda configuración se creó un UBM utilizando los datos de entrenamiento de todos los locutores combinados en un solo modelo GMM, posteriormente los modelos para cada locutor se crearon utilizando el UBM como modelo inicial para el algoritmo EM.

Los modelos GMM también fueron creados con configuraciones de 10 y 32 componentes gaussianas para las distintas combinaciones de tamaños de vectores (cuartiles, octiles y deciles).

Debido a que las bases de datos no tienen el mismo tamaño ni fueron realizadas siguiendo el mismo protocolo bajo la misma norma, los experimentos tuvieron variaciones en cuanto al número de archivos utilizados para entrenamiento y evaluación. En consecuencia, los resultados se explicarán por separado para cada base de datos. Sin embargo, lo importante dentro de esta contribución yace en mostrar la potencialidad de los vectores cuantílicos con distintas bases de datos, las cuales fueron creadas bajo distintas condiciones de grabación e idioma.

Para el caso particular de la base de datos VoCMex, se usaron los 20 locutores masculinos, tomando 40 segundos de entrenamiento, correspondiendo a las dos primeras sesiones de las frases 1,2 y 3, que son fonéticamente equilibradas. Para la evaluación se utilizaron 60 segundos, correspondientes a la frase 4 (texto fijo) de la sesión 3. Debido a que nuestro sistema está orientado a la identificación de locutor, la medición de eficiencia fue efectuada en función de las clasificaciones correctas de locutor. La Tabla 1 muestra los resultados de eficiencia de reconocimiento considerando entrenamiento y evaluación de señales adquiridas por el mismo medio micrófono (M1), teléfono (T1) y VoIP (T3). La parte superior de la tabla indica el tipo de cuantil, el número de componentes gaussianas en los modelos GMM y los medios de adquisición/evaluación. El número de componentes de los vectores fueron 3, 7 y 9 para cuartiles,

octiles y deciles. En este experimento se obtuvieron los mejores resultados con los vectores de mayor dimensión (deciles). Lo anterior puede visualizarse en la siguiente tabla:

Tabla 6.1 - Experimentos con VoCMex sin pre-procesamiento.

Cuantil	GMM	MI	T1	T3
Octil	10	35 %	32.00 %	36.66 %
Cuartil	10	25 %	8.33 %	31.66 %
Decil	10	47 %	20.00 %	40.00 %
Octil	32	20 %	17.00 %	23.00 %
Cuartil	32	17 %	8.33 %	26.66 %
Decil	32	33 %	12.00 %	35.00 %

Lo anterior pone de manifiesto que el vector cuantílico con más dimensiones captura con mayor precisión la estructura fina y/o densidad espectral importante del aparato fonatorio del locutor. Asimismo, se observa que los resultados no mejoran cuando se aumenta el número de componentes gaussianas en los modelos GMM, denotándose mejores resultados con mezclados de 10 componentes gaussianas. Lo anterior no significa que modelos más pequeños mejoren los resultados, puesto que aquí no mostramos otras evaluaciones menos satisfactorias con modelos de menor tamaño. Lo que sí se puede observar es que modelos con más de 10 gaussianas pueden conducir a errores de otra naturaleza como sobre-entrenamiento o redondeo para este sistema en particular. Otro aspecto interesante que permite emitir un juicio es que el mejor resultado fue arrojado al aplicar deciles y directamente con micrófono. Este hecho pone de manifiesto por un lado que el decil capturó mejor la estructura fina del locutor y que el usar directamente micrófono implicaba menos degradación en los modelos debida a ruido o a efectos de canal.

En base a estos resultados, se decidió construir modelos con vectores decílicos y tratando de mejorar los resultados se aplicó pre-procesamiento que consiste en preénfasis en segmentos de tiempo corto de 30ms y traslape de 20ms. Además, se construyó el modelo impostor aplicando UBM. En la Tabla 6.2 se puede observar una mejoría evidente en la eficiencia de identificación de locutor. De nuevo los modelos con 10 GMM arrojaron resultados ligeramente mejores que

los de 32 GMM. Lo anterior refuerza la idea de que modelos más complejos pueden conducir a errores de redondeo o problemas de sobre-entrenamiento y que los vectores decílicos son más consistentes con modelos de 10 gaussianas. Otro aspecto relevante es que tanto el preénfasis como el uso de UBM lograron paliar en gran medida los problemas derivados de los medios de transmisión o canal como se observa para el caso de la señal de teléfono (T1). Asimismo estos resultados refuerzan la hipótesis de que los cuantiles para señales orientadas a la identificación de locutor tienen potencial de aplicación.

Tabla 6.2 - Experimentos con VoCMex y modelos GMM usando UBM.

Cuantiles	GMM	MI	T1
Decil	10	75 %	80.00 %
Decil	32	75 %	75.00 %

Debido a que este trabajo presenta dos variantes con respecto a la mayoría de los trabajos en identificación de locutor, las cuales son, el uso de un corpus en español mexicano y la aplicación de vectores cuantílicos, a manera de comparación se realizaron experimentos con la bases de datos AHUMADA, la cual contiene grabaciones en español Ibérico. Se trató de establecer particiones entrenamiento/prueba con condiciones similares a las utilizadas en los experimentos con VoCMex. Para este caso se tomaron las grabaciones de 20 locutores, el entrenamiento se realizó con las 10 elocuciones fonéticamente equilibradas de las sesiones 1 y 2 (identificadas con el código C), para tener un total aproximado de 40 segundos de señal. Para la evaluación se usó la elocución fonéticamente equilibrada (identificada con el código D) de la sesión 3 de aproximadamente 60 segundos de duración.

En la Tabla 6.3 se muestran los resultados de estos experimentos, notándose la similitud con los mostrados en la Tabla 6.2. Los resultados aquí obtenidos con micrófono son iguales, lo cual demuestra consistencia en los métodos de grabación, sin embargo las grabaciones efectuadas por

teléfono evidencian un mejor control en el caso de VoCMex. Lo cual es consistente con lo reportado por los autores de AHUMADA [48]. Un aspecto interesante a resaltar es que los modelos acústicos con 32 gaussianas lograron mejorar la eficiencia de reconocimiento con respecto a los de 10, es decir de alguna manera se adaptaron más a la distorsión debida al canal de transmisión. Pero lo más importante es que nuevamente los resultados refuerzan el potencial de los vectores cuantílicos.

Tabla 6.3 - Experimentos con AHUMADA y modelos GMM usando UBM.

Cuantiles	GMM	MI	T1
10	10	75 %	25.00 %
10	32	75 %	30.00 %

Habiendo obtenido resultados alentadores con las dos bases de datos en español y habiendo establecido una línea base para los parámetros de preprocesamiento, número de cuantiles, número de componentes gaussianas, tamaño de la ventana y tamaño del traslape, se decidió utilizar el mismo sistema de identificación con la base de señales XM2VTS8K. Aunque esta base de datos sólo contiene una frase fonéticamente equilibrada, la cual se grabó dos veces por sesión, se tuvieron que incluir las frases que constan de dígitos para poder contar con señales de tamaño similar a las utilizadas con los corpus anteriores. Para el entrenamiento se utilizaron las grabaciones 1 a la 3 de las tres primeras sesiones, aproximadamente 50 segundos de señal. Para la evaluación las frases 4 a la 6 de la sesión 4, aproximadamente 20 segundos. Adicionalmente del total de 295 locutores sólo se eligieron las grabaciones de 134, aquellas que después de una revisión se consideraron las de mejor calidad, sin ruidos de fondo o señales de otros locutores que no corresponden al de la grabación. En la Tabla 6.4 se muestran los resultados en las dos modalidades usadas para el entrenamiento, con y sin UBM. Para estos experimentos sólo se

aplicó la configuración de vectores decílicos y 10 componentes gaussianas, ya que en los experimentos con los otras bases de datos es la que demostró mejores resultados.

Tabla 6.4 - Resultados de Identificación con XM2VTS.

Cuantiles	GMM	Sin UBM	Con UBM
10	10	45 %	54.00 %

El resultado de identificación obtenido se observó por debajo del logrado con las otras bases de datos. Con el objetivo de mejorar el nivel de identificación, se restringió la selección de locutores a los 40 que tuvieran las mejores grabaciones, es decir, además de que no tienen ruidos de fondo o locutores adicionales, las señales están bien grabadas con frases completas sin saturación y con mayor nitidez. Con esta modificación, los resultados obtenidos fueron los mejores de todos los experimentos, tal y como se muestran en la Tabla 6.5.

Tabla 6.5 - Resultados de Identificación con XM2VTS, utilizando los 40 locutores con mejores grabaciones.

Cuantiles	GMM	Con UBM
10	10	83.00 %
10	32	90.83 %

Lo anterior pone de manifiesto varios aspectos, uno de ellos que la técnica UBM mejora de manera perceptible los resultados, además que cuando las señales son de muy buena calidad el aumento de gaussianas en los modelos mejora en gran medida la eficiencia y finalmente refuerza el potencial de nuestra propuesta aplicando vectores cuantílicos al grado de ser completamente viable en muchas aplicaciones actuales.

6.5 Conclusiones en IAL con Vectores Cuantílicos

Los vectores acústicos cuantílicos se presentan como una propuesta novedosa en el campo de vectores acústicos, en especial en el dominio de la Identificación de Locutor. Esta representación

se sustenta en trabajos de medicina sobre la capacidad de flujo respiratorio; para el caso de voz, siguió un tratamiento de transformada de Fourier en tiempo corto para tomar en cuenta la cuasi-estacionaridad de la señal.

Dichos vectores proporcionan la oportunidad de relacionar la energía con los valores frecuenciales más importantes de la señal de voz. Hasta el momento se obtuvo una tasa de identificación correcta del 80% en análisis de tiempo corto para elocuciones obtenidas de VoCMex y arriba del 90% en XM2VTS. Los vectores cuantílicos fueron aplicados sobre señales con un mínimo de preprocesamiento (preénfasis y antitraslape) lo cual deja la posibilidad de explorar variantes más elaboradas de esta técnica e intentar obtener mejores resultados.

Capítulo 7

Conclusiones

7 Conclusiones

7.1 Resumen de Resultados

En esta sección se presenta un resumen de los resultados obtenidos en el trabajo de tesis y su relación con los objetivos planteados en la sección 1.2 de este documento.

En el análisis del estado del arte se encontró que los trabajos de investigación relacionados con el Español Mexicano están enfocados en el área del Reconocimiento del Habla, destacan los esfuerzos realizados en la UNAM y la UDLAP. En contraste, no se encontraron referencias de trabajos previos que aborden específicamente la relación del Español Mexicano con el RAL.

Por otra parte, sobre la pregunta planteada al inicio de este documento: ¿Es relevante para un sistema de IAL el idioma de los locutores?, se puede deducir la respuesta a partir de la clasificación de niveles de información presentados en [8] y de los sistemas que fusionan información de varios de estos niveles [9-11]. Con base en estos resultados se observa que existen sistemas de RAL (los cuales abarcan a los de IAL) para los cuales no es relevante el idioma de locutor, siendo estos los que tratan con la información a nivel espectral o prosódico. Esto se debe a que el análisis se hace exclusivamente sobre las características físicas de la señal (e.g. amplitud, frecuencia, energía). Sin embargo, los sistemas que aprovechan la información fonética, sintáctica, dialógica y semántica, sí están estrechamente relacionados con el idioma utilizado por los locutores, ya que es necesario primero hacer reconocimiento del habla para posteriormente hacer el reconocimiento de locutor. De lo anterior se obtiene que las características dependientes del idioma que son relevantes para caracterizar a un locutor son los mismos que se utilizan para el reconocimiento del habla, esto es: fonemas, palabras, sintaxis y

semántica. Por lo tanto, un sistema de IAL para el Español Mexicano deberá incluir al menos una de esas características.

En relación al segundo objetivo particular de la tesis, la construcción del corpus de voz con características del Español Mexicano constituyó la parte fundamental del trabajo realizado. Tomando en consideración lo encontrado en el análisis del estado del arte en relación a las características dependientes del idioma aprovechables para el RAL, se decidió que la característica del Español Mexicano a incluir en el corpus sería la distribución fonética del idioma. Esto implica que el corpus se forma de elocuciones cuyos fonemas tienen una frecuencia de uso similar a la frecuencia de uso de los fonemas en el Español Mexicano. En este sentido, el primer paso para construir el corpus fue encontrar la distribución fonética del Español Mexicano, la cual se obtuvo a partir de textos recabados de la Web, particularmente de portales de periódicos en Español Mexicano. La distribución obtenida por este método fue contrastada mediante un análisis de correlación con la publicada en [55], arrojando un coeficiente de 0.995, lo cual indica un alto grado de similitud entre las dos. Con esta distribución se definieron las frases a incluir en el corpus, para hacerlo se tomaron en consideración las características presentadas en [48,60], resultando en tres frases fonéticamente equilibradas y un texto no equilibrado. Una vez definidas las frases a grabar, el siguiente paso fue realizar las sesiones de grabación. En total participaron 50 locutores, de los cuales 33 (20 hombres y 13 mujeres) completaron tres sesiones y los restantes 17 (11 hombres y 6 mujeres) realizaron solo una o dos sesiones. En cada sesión se grabaron las frases mediante tres modalidades: teléfono, micrófono y VoIP. Al corpus resultante se le denominó *VoCMex*. Para validarlo se construyó un sistema de IAL, mismo que obtuvo un reconocimiento del 95% para señal de micrófono y 70% para señal de teléfono. Estos resultados se contrastaron con los obtenidos al utilizar el mismo sistema IAL

con el corpus ibérico Ahumada, resultando VoCMex con un mejor reconocimiento para la señal de micrófono. Con estos resultados se demuestra que el corpus es útil para la experimentación en IAL, siendo la primera base de señales de voz que contiene frases con la distribución fonética del Español Mexicano y que además contiene señales adquiridas mediante un sistema telefónico de VoIP.

Adicionalmente, VoCMex se utilizó para demostrar el uso de vectores acústicos cuantílicos en un sistema de IAL. Los vectores acústicos son la transformación de la señal de voz de la representación tiempo/amplitud a otra representación que resalte la información útil para el RAL. En este sentido los vectores acústicos cuantílicos se presentan como una alternativa a los vectores acústicos obtenidos por medio de técnicas conocidas en RAL (e.g. MFCC o LPCC). Para la obtención de los vectores acústicos cuantílicos se parte de la representación de la señal de voz en el dominio de las frecuencias (obtenida mediante la FFT), se toma esta representación como si fuera la función de densidad de probabilidad y entonces se ubican los valores frecuenciales para los cuales la probabilidad acumulativa va cumpliendo con los criterios de los cuantiles elegidos (i.e. cuartiles, octiles, deciles, etc.), estos valores cuantílicos obtenidos conforman al vector acústico. Para probar esta técnica se utilizó el sistema de IAL con VoCMex, con el cual se obtuvo una tasa de reconocimiento del 80% en señal de micrófono usando deciles y modelado GMM con 10 componentes. Con esta misma configuración del sistema IAL se probó para el corpus Ahumada obteniendo un reconocimiento del 75% y con el corpus XM2VTS se logró una configuración que alcanzó el 90% de reconocimiento.

7.2 Contribuciones

Como primer contribución se tiene la obtención de la distribución fonética del Español Mexicano a partir de páginas web de periódicos mexicanos, en particular de las secciones de noticias

locales. El análisis realizado en dos modalidades, por regiones y global, mostró que la frecuencia de uso de los 22 fonemas del idioma no varía entre regiones y con respecto al uso global.

La segunda contribución es VoCMex, el primer corpus de voz orientado para apoyar la investigación en RAL que incluye grabaciones con la distribución fonética del Español Mexicano y también el primero en español en incluir señales adquiridas mediante un sistema de VoIP. La relevancia de esta contribución radica en el hecho de que al no haberse encontrado trabajos previos relacionados con el RAL en Español Mexicano, no había un punto de partida para realizar este tipo de investigaciones, VoCMex representa ese primer esfuerzo y punto de partida para nuevas investigaciones.

Las contribuciones anteriores aparecen en el artículo: "Corpus de Voz en Español Mexicano para Experimentación en Reconocimiento Automático de Locutor", presentado en el *VII Congreso Internacional en Tecnologías Inteligentes y de la Información (CITII 2010)* y publicado en la revista *Research in Computing Science Vol. 50, ISSN 18770-4069.*

La segunda contribución aparece en el artículo: "VoCMex: a voice corpus in Mexican Spanish for research in speaker recognition", publicado en la revista *International Journal of Speech Technology* con DOI 10.1007/s10772-012-9183-z.

La tercera contribución es la aplicación de cuantiles en la IAL, una técnica que no había sido utilizada en el área del reconocimiento de locutor. Los resultados obtenidos muestran su viabilidad como una alternativa a las técnicas ya utilizadas en la construcción de vectores acústicos, con la ventaja de no requerir la aplicación de filtros a la señal en el domino de las frecuencias, lo cual la hace más simple de calcular.

Esta tercera contribución aparece en el artículo: "Identificación de Locutor usando Vectores Acústicos basados en Cuantiles", presentado en el *VII Congreso Internacional en Tecnologías*

Inteligentes y de la Información (CITII 2012) y publicado en la revista *Research in Computing Science Vol. 60, ISSN: 18770-4069.*

7.3 Trabajo Futuro

Siendo esta tesis un trabajo pionero en el área de RAL en Español Mexicano, la perspectiva de trabajo futuro se considera amplia. A continuación se presentan algunas de las que se identifican como una continuación inmediata a los resultados aquí presentados.

En relación a la distribución fonética del Español Mexicano, el siguiente paso es realizar un análisis que incluya también los alófonos de cada fonema, es decir, las variaciones que cada fonema tiene dependiendo de la posición en la palabra. Esto permitirá tener un nivel más detallado de información a considerar cuando se analicen las señales de voz de los locutores.

En relación al corpus de voz, aparte de agregar más locutores, se pueden incluir grabaciones que contengan conversaciones espontáneas entre dos locutores para contar con información de niveles dialógico y semántico con la finalidad de utilizarlo en sistemas de RAL que aprovechen estos niveles. También hacer un análisis de las particularidades del Español Mexicano por regiones en los niveles prosódico y sintáctico, esto permitirá reconocer no solo a la persona por su forma de hablar, sino también la región de donde proviene.

En relación al uso de cuantiles para IAL, un resultado interesante sería realizar la comparación del desempeño (rapidez, eficiencia) con respecto a otras de las técnicas de obtención de vectores acústicos, con la finalidad de determinar en qué tipo de aplicaciones conviene utilizarlos.

Referencias

[1] J. A. Markowitz, "Voice biometrics," *Commun. ACM,* vol. 43, pp. 66-73, 2000.

[2] G. Friedland and O. Vinyals, "Live speaker identification in conversations," presented at the ACM Multimedia, 2008.

[3] S. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech & Language Processing,* vol. 14, pp. 1557-1565, 2006.

[4] T. Kinnunen, E. Karpov, and P. Fränti, "Real-time speaker identification and verification," *IEEE Transactions on Audio, Speech & Language Processing,* vol. 14, pp. 277-288, 2006.

[5] H. Do, I. Tashev, and A. Acero, "A new speaker identification algorithm for gaming scenarios," presented at the ICASSP, 2011.

[6] F. Burkhardt, R. Huber, and A. Batliner, "Application of Speaker Classification in Human Machine Dialog Systems," presented at the Speaker Classification (1), 2007.

[7] C. Fredouille, "Approche Statistique pour la Reconnaissance Automatique du Locuteur: Informations Dynamiques et Normalisation Bayesienne des Vraisemblances," Avignon, France, These pour obtenir le grade du Docteur2000.

[8] M. Faúndez-Zanuy and E. Monte-Moreno. (2005, May) State-of-the-Art in Speaker Recognition. *IEEE Aerospace and Electronic Systems Magazine*. 7-12.

[9] J. P. Campbell, D. A. Reynolds, and R. D. Dunn, "Fusing High- and Low-Level Features for Speaker Recognition," in *Eurospeech Proc.*, Geneva, Switzerland, 2003, pp. 2665-2668.

[10] D. A. Reynolds, W. Andrews, J. P. Campbell, J. Navratil, B. Peskin, A. Adami*, et al.*, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, 2003, pp. IV-784-7 vol.4.

[11] Y. A. Solewicz and M. Koppel, "Using Post-Classifiers to Enhance Fusion of Low- and High-Level Speaker Recognition," *Trans. Audio, Speech and Lang. Proc.,* vol. 15, pp. 2063-2071, 2007.

[12] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*: Prentice Hall, 1993.

[13] B.-H. Juang and T. Chen. (1998) The past, present, and future of speech processing. *IEEE Signal Processing Magazine, IEEE, Vol. 15, No. 3. (1998), pp. 24-48 Key: citeulike:109117*. 24-48.

[14] C. Fredouille, J. Mariéthoz, C. Jaboulet, J. Hennebert, C. Mokbel, and F. Bimbot, "Behavior of a Bayesian Adaptation Method for Incremental Enrollment in Speaker Verification," in *IEEE Int. Conf. on Acoustics, Speech ans Signal Processsing (ICASSP2000)*, Istambul, Turkey, 2000, pp. 1197-1200.

[15] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," in *Eurospeech Proc.*, Aalborg , Denmark, 2001, pp. 2521-2524.

[16] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier*, et al.*, "A Tutorial on Text-Independent Speaker Verification," *EURASIP J. on Advances in Signal Processing,* pp. 430-451, 4 21 2004.

[17] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Communication,* vol. 17, pp. 91-108, 8 1995.

[18] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2002)*, Orlando, FL, USA, 2002, May 13-17, pp. IV-4072 - IV-4075.

[19] D. A. Reynolds, "A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification," Georgia, Ph. D. ThesisAug. 1992.

[20] D. A. Reynolds and R. C. Rose, "Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models," in *IEEE Transactions on Speech and Audio Processing*, Jan. 1995, pp. 72-83.

[21] I. Magrin-Chagnolleau, G. Durou, and F. Bimbot, "Application of time-frequency principal component analysis to text-independent speaker identification," *Speech and Audio Processing, IEEE Transactions on,* vol. 10, pp. 371-378, 2002.

[22] T. Kinnunen, J. Saastamoinen, V. Hautamaki, M. Vinni, and P. Franti, "Comparing maximum a posteriori vector quantization and Gaussian mixture models in speaker verification," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 4229-4232.

[23] D. Hosseinzadeh and S. Krishnan, "Combining Vocal Source and MFCC Features for Enhanced Speaker Recognition Performance Using GMMs," in *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, 2007, pp. 365-368.

[24] D. Hosseinzadeh and S. Krishnan, "On the use of complementary spectral features for speaker recognition," *EURASIP J. Adv. Signal Process,* vol. 2008, pp. 1-10, 2008.

[25] D. Hosseinzadeh and S. Krishnan, "Gaussian Mixture Modeling of Keystroke Patterns for Biometric Applications," *Trans. Sys. Man Cyber Part C,* vol. 38, pp. 816-826, 2008.

[26] M. Grimaldi and F. Cummins, "Speaker Identification Using Instantaneous Frequencies," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, pp. 1097-1111, 2008.

[27] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust Speaker Recognition in Noisy Conditions," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 1711-1723, 2007.

[28] L. Ferrer, E. Shriberg, S. Kajarekar, A. Stolcke, K. Sonmez, A. Venkataraman*, et al.*, "The Contribution of Cepstral and Stylistic Features to SRI's 2005 NIST Speaker Recognition Evaluation System," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 2006, pp. I-I.

[29] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 2095-2103, 2007.

[30] S. Kajarekar, L. Ferrer, K. Sönmez, J. Zheng, E. Shriberg, and A. Stolcke, "Modeling NERFs for Speaker Recognition," in *Proc. Odyssey 2004: The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004, pp. 51-56.

[31] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling Prosodic Features Sequence s for Speaker Recognition," *Speech Communication,* vol. 46, pp. 455-472, 2005.

[32] W. M. Campbell, J. P. Campbell, D. A. Reynolds, J. Reynolds, and T. R. Leek, "Phonetic Speaker Recognition with Support Vector Machines," presented at the Advances in Neural Information Processing Systems 16, Neural Information Processing Systems (NIPS 2003), Vancouver and Whistler, British Columbia, Canada, 2003.

[33] W. M. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE,* vol. 13, pp. 308-311, 2006.

[34] D. Klusacek, J. Navratil, D. A. Reynolds, and J. P. Campbell, "Conditional pronunciation modeling in speaker detection," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, 2003, pp. IV-804-7 vol.4.

[35] K.-Y. Leung, M.-W. Mak, M.-H. Siu, and S.-Y. Kung, "Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification," *Speech Communication,* vol. 48, pp. 71-74, 2006.

[36] A. O. Hatch, B. Peskin, and A. Stolcke, "Improved Phonetic Speaker Recognition Using Lattice Decoding," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 2005, pp. 169-172.

[37] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective," *Speech Communication,* vol. 31, pp. 225-254, 6// 2000.

[38] C. S. Greenberg and A. Martin, "NIST speaker recognition evaluations 1996-2008," in *Proc. SPIE 7324, Atmospheric Propagation VI*, 2009.

[39] M. Przybocki, A. Martin, and A. N. Le, "NIST Speaker Recognition Evaluation Chronicles - Part 2," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, 2006, pp. 1-6.

[40] D. Baum, "Recognising speakers from the topics they talk about," *Speech Communication,* vol. 54, pp. 1132-1142, 12// 2012.

[41] M.-I. Faraj and J. Bigun, "Synergy of Lip-Motion and Acoustic Features in Biometric Speech and Speaker Recognition," *IEEE Trans. Comput.,* vol. 56, pp. 1169-1175, 2007.

[42] K. S. Rao, A. K. Vuppala, S. Chakrabarti, and L. Dutta, "Robust speaker recognition on mobile devices," in *Signal Processing and Communications (SPCOM), 2010 International Conference on*, 2010, pp. 1-5.

[43] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication,* vol. 52, pp. 12-40, 1// 2010.

[44] P.-Y. Shih, P.-C. Lin, J.-F. Wang, and Y.-N. Lin, "Robust several-speaker speech recognition with highly dependable online speaker adaptation and identification," *Journal of Network and Computer Applications,* vol. 34, pp. 1459-1467, 9// 2011.

[45] T. Herbig, F. Gerl, and W. Minker, "Self-learning speaker identification for enhanced speech recognition," *Computer Speech & Language,* vol. 26, pp. 210-227, 6// 2012.

[46] E. Charton, A. Learcher, C. Levy, and J.-F. Bonastre, "Mistral: open source biometric platform," in *SAC '10 Proceedings of the 2010 ACM Symposium on Applied Computing*, 2010, pp. 1503-1504.

[47] B. G. B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J. S. D. Mason, "State-of-the-Art Performance in Text-Independent Speaker Verification Through Open-Source Software," *Trans. Audio, Speech and Lang. Proc.,* vol. 15, pp. 1960-1968, 2007.

[48] J. Ortega-García, J. González-Rodríguez, and V. Marrero-Aguiar, "AHUMADA: A Large Speech Corpus in Spanish for Speaker Characterization and Identification," *Speech Communication,* vol. 31, pp. 255-264, 6 2000.

[49] L. A. Pineda, H. Castellanos, J. Cuétara, L. Galescu, J. Juárez, J. Llisterri*, et al.*, "The Corpus DIMEx100: Transcription and Evaluation. Language Resources and Evaluation," *Language Resources and Evaluation,* vol. 44, pp. 347-370, 2010.

[50] R. Auckenthaler, E. S. Parris, and M. J. Carey, "Improving a GMM Speaker Verification System by Phonetic Weighting," in *Proceedings of the Acoustics, Speech and Signal Processing*, Phoenix, Arizona, USA, 1999.

[51] J. P. Campbell, "Testing with the YOHO CD-ROM Voice Verification Corpus," in *Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP-95*, Detroit, USA, 1995, pp. 341-344.

[52] J. P. Campbell and D. A. Reynolds, "Corpora for the Evaluation of Speaker Recognition Systems," in *Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-99*, Phoenix, Arizona, USA, 1999, pp. 829-832.

[53] F. Casacuberta Nolla, R. García Gómez, J. Llisterri Boix, C. Nadeu Camprubi, J. M. Pardo Muñoz, and A. J. Rubio Ayuso, "Desarrollo de Corpus para Investigación en Tecnologías del Habla (ALBAYZIN)," *Procesamiento del Lenguage Natural,* vol. 12, pp. 35-42, 7 1992.

[54] H. E. Pérez. (2003, 3) Frecuencia de Fonemas. *Revista Electrónica de la Red Temática en Tecnologías del Habla*.

[55] L. Villaseñor-Pineda, M. Montes-y-Gómez, D. Vaufreydaz, and J.-F. Serignat, "Elaboración de un Corpus Balanceado para el Cálculo de Modelos Acústicos usando la Web," in *Internationnal Conference on Computing (CIC-2003)*, Mexico City, 2003, pp. 198-200.
[56] W. L. Martinez and A. R. Martinez, *Computational Statistics Handbook with MATLAB*: Chapman & Hall/CRC, 2008.
[57] W. L. Martinez, A. R. Martinez, and J. L. Solka, *Exploratory Data Analysis with MATLAB*: Chapman & Hall/CRC, 2010.
[58] J. Hennebert, H. Melin, D. Petrovska, and D. Genoud, "POLYCOST: A telephone-speech database for speaker recognition," *Speech Communication,* vol. 31, pp. 265 - 270, 2000.
[59] M. Przybocki and A. Martin, "NIST Speaker Recognition Evaluation Chronicles," in *Odyssey 2004: Procedings of The Speaker and Language Recognition Workshop* Toledo, Spain, 2004.
[60] K. Messer, J. Matas, J. V. Kittler, J. Lüttin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," in *Proc. of the 2nd Int. Conf. on Audio and Video Based Biometric Person Authentication (AVBPA'99)*, Washington DC, USA, 1999.
[61] M. Zamalloa, G. Bordel, L. J. Rodríguez, M. Peñagarikano, and J. P. Uribe, "Selección y pesado de parámetros acústicos mediante algoritmos genéticos para el reconocimiento del locutor," in *IV Jornadas en Tecnologia del Habla*, 2006, pp. 349-354.
[62] I. Kirschning, "TLATOA: Developing Speech Technology & Applications for Mexican Spanish. ," in *2nd. Intl.Workshop on Spanish Language Processing and Language Technologies*, Jaén, Spain, 2001.
[63] D. Gomillion, B. Dempster, A. Epshteyn, R. Clews, J. Kolasinski, and Books24x7 Inc. (2005). *Building telephony systems with Asterisk an easy introduction to using and configuring Asterisk to build feature-rich telephony systems for small and medium businesses* [Text].
[64] J. V. Meggelen, J. Smith, L. Madsen, and Safari Books Online (Firm). (2005). *Asterisk the future of telephony (1st ed.)*. Available: http://proquest.safaribooksonline.com/0596009623
[65] J. V. Meggelen, J. Smith, L. Madsen, and Safari Books Online (Firm). (2007). *Asterisk the future of telephony (2nd ed.)*. Available: http://proquest.safaribooksonline.com/9780596510480
[66] P. Mayorga, C. Druzgalski, and O. H. González, "Quantile Vectors based Verification of Normal Lung Sounds," in *PAHCE-2012 (Pan American Health Care Exchanges)*, Miami, FL, USA, 2012, Mar. 26-31, pp. 7-11.
[67] P. Mayorga, C. Druzgalski, O. H. González, A. Zazueta, and M. A. Criollo, "Expanded Quantitative Models for Assessment of Respiratory Diseases and Monitoring," in *PAHCE-2011 (Pan American Health Care Exchanges Conf. 2011)*, Rio de Janeiro, Brazil, March 2011.
[68] P. Mayorga, C. Druzgalski, and J. Vidales, "Quantitative Models for Assessment of Respiratory Diseases," in *PAHCE-2010 (Pan American Health Care Exchanges Conf. 2010)*, Lima, Peru, 2010.
[69] H. Perez-Meana, *Advances in Audio and Speech Signal Processing: Technologies and Applications*: IGI Global, 2007, Feb. 28.
[70] J. P. Campbell. (1997, Sept.) Speaker Recognition: A Tutorial. *Proceedings of the IEEE*. 1437-1462.
[71] D. Pearce, "An Overview of ETSI Standards Activities for Distributed Speech Recognition Front-Ends," in *AVIOS 2000: The Speech Applications Conference*, San Jose, CA, USA, May 22-24 2000.
[72] D. M. Istrate, "Detection et Reconnaissance des Sons pour la Surveillance Médicale," These pour obtenir le grade de docteur de l'INPG: spécialité Signal, Image, Parole, Télécoms;2003.
[73] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, and H. Leich, *Traitement de la Parole*: PPUR Presses Polytechniques et Universitaires Romandes, 2000.
[74] B. Milner and A. James, "Robust Speech Recognition Over Mobile and IP Networks in Burst-Like Packet Loss," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 14, pp. 223-231, Jan. 2006.

[75] B. Milner and S. Semnani, "Robust Speech Recognition over IP Networks," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP2000*, Istambul, Turkey, Jun 2000.

[76] J. Solé-Casals and V. Zaiats, "Advances in Nonlinear Speech Processing," in *NOLISP 2009 (Int. Conf. on Nonlinear Speech Processing) revised selected papers in Lecture Notes in Computer Science*, Vic, Spain, 2010.

[77] H. Hermansky, "Perceptual Linear Prediction (PLP) Analysis for Speech," *J. of the Acoustical Society of America,* vol. 87, pp. 1738-1752, 1990.

[78] H. Hermansky and P. Fousek, "Multi-Resolution RASTA Filtering for TANDEM-based ASR," in *Proc. of INTERSPEECH 2005 (European Conference on Speech Communication and Technology)*, Lisbon, Portugal, 2005, pp. 361-364.

[79] M. J. Carey and E. S. Parris, "Speaker verification using connected words," in *ProceedingsofInstituteofAcoustics*, 1992, pp. 95-100.

[80] D. A. Reynolds, T. Quatieri, and R. D. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, pp. 19-41, 2000.

[81] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, pp. 980-988, 2008.

[82] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *Speech and Audio Processing, IEEE Transactions on,* vol. 8, pp. 695-707, 2000.

[83] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances In Channel Compensation For SVM Speaker Recognition," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 2005, pp. 629-632.

[84] T. Hasan and J. H. L. Hansen, "A Study on Universal Background Model Training in Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 1890-1899, 2011.

Anexo A

Archivos de Configuración de Web-Harvest

Los siguientes archivos XML fueron utilizados para configurar la extracción de texto de los sitios Web de noticias. A través de ellos, se definieron las URL (direcciones Web) a analizar y el método de análisis del texto para el modelado de la distribución fonética.

Archivo de definición de fuentes de datos

```
<?xml version="1.0" encoding="UTF-8"?>

<!--
    Document   : diarios.xml
    Created on : 25 de enero de 2008, 03:05 PM
    Author     : molguin
    Description:
        Purpose of the document follows.
-->

<diarios>
  <diario nombre="http://www.oem.com.mx/lavozdelafrontera"/>
  <diario nombre="http://www.oem.com.mx/diariodequeretaro"/>
  <diario nombre="http://www.oem.com.mx/elsoldeirapuato"/>
  <diario nombre="http://www.oem.com.mx/elsoldesalamanca"/>
  <diario nombre="http://www.oem.com.mx/elsoldesanjuandelrio"/>
  <diario nombre="http://www.oem.com.mx/elsoldelbajio"/>
  <diario nombre="http://www.oem.com.mx/diariodexalapa"/>
  <diario nombre="http://www.oem.com.mx/elsoldecordoba"/>
  <diario nombre="http://www.oem.com.mx/elsoldeorizaba"/>
  <diario nombre="http://www.oem.com.mx/elsoldetampico"/>
  <diario nombre="http://www.oem.com.mx/diariodelsur"/>
  <diario nombre="http://www.oem.com.mx/elheraldodechiapas"/>
  <diario nombre="http://www.oem.com.mx/elheraldodetabasco"/>
  <diario nombre="http://www.oem.com.mx/elsoldedurango"/>
  <diario nombre="http://www.oem.com.mx/elheraldodechihuahua"/>
  <diario nombre="http://www.oem.com.mx/elsoldezacatecas"/>
  <diario nombre="http://www.oem.com.mx/elsoldepuebla"/>
  <diario nombre="http://www.oem.com.mx/elsudcaliforniano"/>
  <diario nombre="http://www.oem.com.mx/elsoldemexico"/>
  <diario nombre="http://www.oem.com.mx/elsoldetoluca"/>
  <diario nombre="http://www.oem.com.mx/elsoldemazatlan"/>
</diarios>
```

Archivo de configuración para la extracción de texto específico de las secciones locales de los diarios

```
<?xml version="1.0" encoding="UTF-8"?>

<config charset="ISO-8859-1">

      <var-def name="diariosOEM">
        <xpath expression="//diario/@nombre">
            <file action="read" path="diariosOEM.xml"/>
        </xpath>
    </var-def>
      <loop item="diario" index="s" filter="unique">
            <list>
                  <var name="diariosOEM"/>
            </list>
            <body>
            <var-def name="notas">
                  <xpath expression="//div[@class='cabezasecu_esto']/a/@href">
                        <var-def name="principal">
                              <html-to-xml>
                                    <try>
                                          <body>
                                    <http url="${diario}/locales.aspx"/>
                              </body>
                              <catch>
                                    <![CDATA[<html></html>]]>
                              </catch>
                        </try>
                        </html-to-xml>
                  </var-def>
            </xpath>
            <xpath expression="//div[@class='resumenesto']/a/@href">
                        <var name="principal"/>
            </xpath>
      </var-def>


     <loop item="link" index="i" filter="unique">
        <list>
            <var name="notas"/>
        </list>
        <body>

                  <var-def name="textoCasiLimpio">
                        <regexp replace="true">
                              <regexp-pattern><![CDATA[<br/>]]></regexp-pattern>
                              <regexp-source>
                                    <xpath expression="//div[@class='texto']">
                                          <html-to-xml>
                                                <try>
                                                      <body>
                                                            <http url="${diario}/${link}"/>
                                                <catch>

      <![CDATA[<html></html>]]>
                                                </catch>
                                          </try>
```

```
56                                          </html-to-xml>
57                                    </xpath>
58                                          </regexp-source>
59                                    <regexp-result>
60                                          <template></template>
61                                    </regexp-result>
62                                    </regexp>
63                              </var-def>
64
65                              <var-def name="textoCasiLimpio">
66                              <regexp replace="true">
67                                          <regexp-
    pattern><![CDATA[</div>]]></regexp-pattern>
68                                    <regexp-source>
69                                          <var name="textoCasiLimpio"/>
70                                          </regexp-source>
71                                    <regexp-result>
72                                          <template></template>
73                                    </regexp-result>
74                                    </regexp>
75                              </var-def>
76                        <file action="append" type="text"
    path="texto20080805.txt">
77                              <regexp replace="true">
78                                          <regexp-pattern><![CDATA[<div
    class="texto">]]></regexp-pattern>
79                                    <regexp-source>
80                                          <var name="textoCasiLimpio"/>
81                                          </regexp-source>
82                                    <regexp-result>
83                                          <template></template>
84                                    </regexp-result>
85                                    </regexp>
86                  </file>
87              </body>
88        </loop>
89        </body>
90        </loop>
91    </config>
```

Anexo B

Programa Aplicado en la Detección Automática de Fonemas en el Texto

El código mostrado a continuación corresponde a la clase en lenguaje Java utilizada para analizar el texto recabado de la Web y separarlo en palabras que son almacenadas en una base de datos. (ver apartado 4.2)

```
package procesadordefonemas;

import java.io.*;
import java.util.logging.Level;
import java.util.logging.Logger;
import java.util.regex.*;
import java.sql.*;

/**
 *
 * @author molguin
 */
public class ProcesadorDePalabras {

    /**
     * @param args the command line arguments
     */
    public     void     inicia(String     arch,     String     base)     throws
IllegalAccessException{
        BufferedReader in = null;
        String nomInFile = arch;

        String linea=null;
        String palabra=null;

        Pattern patronNormal=Pattern.compile("[a-záéíóúñ]+\\b");
        Matcher empatadorNormal;

        Statement stm=null;

        try {

            Connection                          conn                          =
DriverManager.getConnection("jdbc:mysql://localhost/"+base+"?user=&password="
);
            stm=conn.createStatement();

            in = new BufferedReader(new FileReader(nomInFile));
            System.out.println("Procesando...");
            int progress=0;
            while((linea= in.readLine())!=null){
                if((++progress%100)==0)System.out.print("#");
                linea=linea.toLowerCase();
                empatadorNormal = patronNormal.matcher(linea);
                while(empatadorNormal.find()){
                  palabra=linea.substring(empatadorNormal.start(),
empatadorNormal.end());
                  guardaPalabra(stm,palabra);
                }
            }
```

```
            in.close();
            stm.close();
        } catch (FileNotFoundException ex) {

Logger.getLogger(ProcesadorDePalabras.class.getName()).log(Level.SEVERE,
null, ex);
        } catch(SQLException e){
            e.printStackTrace();
        } catch (IOException ex) {

Logger.getLogger(ProcesadorDePalabras.class.getName()).log(Level.SEVERE,
null, ex);
        } finally {
            try {
                in.close();
            } catch (IOException ex) {

Logger.getLogger(ProcesadorDePalabras.class.getName()).log(Level.SEVERE,
null, ex);
            }
            System.out.println("\nFin");
        }
    }

    static void guardaPalabra(Statement s, String p){

        try {
            String palabra = p;
            ResultSet res = s.executeQuery("Select ocurrencia from palabra
where palabra='" + palabra + "'");
            if (res.first()) { //Ya existe en la BD
                int valor = res.getInt(1);
                valor++;
                s.executeUpdate("update palabra set ocurrencia=" + valor + "
where palabra='" + palabra + "'");
            }else{ //Insertarla
                s.executeUpdate("Insert into palabra (palabra,ocurrencia)
values('"+palabra+"',1)");

            }
            res.close();
        } catch (SQLException ex) {

Logger.getLogger(ProcesadorDePalabras.class.getName()).log(Level.SEVERE,null,
ex);
        }
    }
}
```

La siguiente clase en lenguaje Java fue diseñada para analizar las palabras almacenadas en la base de datos.

```
package procesadordefonemas;

import com.mysql.jdbc.exceptions.*;
import java.sql.*;
import java.util.*;
import java.util.Hashtable;
import java.util.logging.*;
```

```
import java.util.regex.*;

/**
 *
 * @author molguin
 */
public class Fonemizador {

    Connection conn=null;
    Hashtable<String,String> fonemaPatron = new Hashtable<String,String>();
     String kk = "algo";

    Fonemizador(String bd, String usr, String pass){
        try {
            Class.forName("com.mysql.jdbc.Driver").newInstance();
        } catch (Exception ex) {
            Logger.getLogger(Fonemizador.class.getName()).log(Level.SEVERE,
null, ex);
        }
        try {
            conn                                                          =
DriverManager.getConnection("jdbc:mysql://localhost/"+bd+"?"              +
"user="+usr+"&password="+pass);
            inicializaTablaDeFonemas();
            guardaFonemas();

        } catch (SQLException ex) {
            Logger.getLogger(Fonemizador.class.getName()).log(Level.SEVERE,
null, ex);
        }
    }

    private void inicializaTablaDeFonemas(){
        fonemaPatron.put("a", "[aá]");
        fonemaPatron.put("e", "[eé]");
        fonemaPatron.put("i", "[ií]");
        fonemaPatron.put("o", "[oó]");
        fonemaPatron.put("u", "[uúw]");
        fonemaPatron.put("b", "b");
        fonemaPatron.put("cs", "ce|ci");
        fonemaPatron.put("cc","cc");
        fonemaPatron.put("k", "ka|ke|ki|ko|ku|ca|co|cu");
        fonemaPatron.put("d", "d");
        fonemaPatron.put("f", "f");
        fonemaPatron.put("j", "j|ge|gi");
        fonemaPatron.put("gu", "gu");
        fonemaPatron.put("gago", "ga|go");
        fonemaPatron.put("g","gue|gui");
        fonemaPatron.put("q", "que|qui");
        fonemaPatron.put("l","l");
        fonemaPatron.put("ll","ll");
        fonemaPatron.put("m", "m");
        fonemaPatron.put("n", "n");
        fonemaPatron.put("eñe", "ñ");
        fonemaPatron.put("p", "p");
        fonemaPatron.put("r","r");
        fonemaPatron.put("rr","rr");
        fonemaPatron.put("s", "[sz]");
        fonemaPatron.put("t", "t");
        fonemaPatron.put("v", "v");
        fonemaPatron.put("iy", "y");
        fonemaPatron.put("y", "ya|ye|yi|yo|yu");
    }
```

```
    private void guardaFonemas(){
        Enumeration<String> llaves=fonemaPatron.keys();
        String llave=null;
        try{
          Statement stm= conn.createStatement();

          while(llaves.hasMoreElements()){
              llave=llaves.nextElement();
              try{
                stm.executeUpdate("Insert  into  fonemas  (fonema,ocurrencia)
values('"+llave+"',0)");
              }catch(MySQLIntegrityConstraintViolationException e){
                  System.out.println("Fonema ya existente en la BD: "+llave);
              }
          }
          stm.close();
        }catch(Exception e){
            e.printStackTrace();
        }
    }
    public void fonemiza(){
        String llave=null;
        ResultSet rs=null;
        String palabra=null;
        try{
            PreparedStatement pStmPalabras = conn.prepareStatement("Select *
from palabra");
            Statement st = conn.createStatement();
            rs=pStmPalabras.executeQuery();
            if(rs.first()){
                do{
                    palabra=rs.getString(2);

                    Iterator<String>                llaves=(Iterator<String>)
fonemaPatron.keys();
                    while(llaves.hasNext()){
                        llave=llaves.next();
                        Pattern p = Pattern.compile(fonemaPatron.get(llave));
                        Matcher m = p.matcher(palabra);
                        if(m.matches()){
                          int ocurrencias=1;
                          while(m.find()){
                              ocurrencias++;
                          }
                          st.executeUpdate("update  fonemas  set  ocurrencia =
ocurrencia + "+ocurrencias+" where fonema='"+llave+"'");
                        }

                    }
                }while(rs.next());
            }
            pStmPalabras.close();
        }catch(SQLException e){
            e.printStackTrace();
        }
    }

}
```

Anexo C

Frecuencia de fonemas encontrada en cada zona geográfica

Fonema	Norte	Golfo	Pacífico	Centro	Sur	Bajío	Oriente
/i/	0.076139	0.075148	0.076233	0.078239	0.074139	0.077107	0.076252
/e/	0.139139	0.138630	0.138360	0.139033	0.137420	0.138802	0.138024
/a/	0.127752	0.128082	0.127172	0.124644	0.129010	0.126306	0.126516
/o/	0.092585	0.093272	0.094490	0.094330	0.093518	0.093804	0.094020
/u/	0.030675	0.029093	0.028958	0.029129	0.029194	0.029921	0.029872
/p/	0.027196	0.028359	0.027690	0.028033	0.028960	0.028358	0.028565
/t/	0.044642	0.045197	0.045767	0.045857	0.046780	0.046750	0.047191
/k/	0.039738	0.041438	0.040943	0.040677	0.040884	0.040278	0.041371
/b/	0.020537	0.021586	0.020589	0.020228	0.021052	0.019908	0.020287
/d/	0.055484	0.053066	0.054779	0.054481	0.052975	0.053308	0.053325
/g/	0.004523	0.004104	0.004234	0.004251	0.004363	0.004563	0.004363
/f/	0.007803	0.007986	0.007839	0.007887	0.007750	0.007685	0.007573
/s/	0.097720	0.096194	0.095704	0.096066	0.099028	0.095755	0.095852
/j/	0.008524	0.008583	0.009180	0.008797	0.008106	0.009067	0.008817
/ch/	0.002060	0.001480	0.001563	0.001914	0.002463	0.001512	0.001688
/ll/	0.004902	0.004653	0.004409	0.004140	0.004768	0.004652	0.004312
/m/	0.025613	0.026749	0.026115	0.026980	0.027224	0.027793	0.026746
/n/	0.072383	0.071196	0.072170	0.072257	0.071446	0.072591	0.072127
/ñ/	0.001463	0.001744	0.001537	0.001361	0.001760	0.001538	0.001681
/r/	0.065319	0.067862	0.065865	0.065874	0.064799	0.066393	0.065902
/rr/	0.002106	0.001887	0.001714	0.001629	0.001809	0.001816	0.001549
/ll/	0.053696	0.053689	0.054690	0.054196	0.052550	0.052094	0.053967

Anexo D

Software y Hardware para el Desarrollo del Corpus

Asterisk

Asterisk[9] es una plataforma de software de código abierto, orientada a la creación de aplicaciones de comunicación. De este modo, Asterisk puede convertir una computadora en un servidor de comunicaciones. Era plataforma es compatible con sistemas IP PBX, VoIP, servicios de conferencia y demás soluciones en telefonía por internet.

ZoIPer

ZoIPer[10] es una aplicación multiplataforma para emular telefonía a través de internet, es decir un *softphone* (del inglés software y telephone, un teléfono en software). El fin de ZoIPer es crear enlaces telefónicos entre los distintos dispositivos y sistemas operativos sobre los cual opera, a través de un entorno de voz sobre IP (VoIP). Esta herramienta es comercial, y está disponible para diversos sistemas operativos y dispositivos.

Audacity

Audacity[11] es una aplicación para grabación y edición de audio, disponible para los principales sistemas operativos. Esta aplicación es de código abierto, bajo la Licencia Pública General (GLP, del inglés General Public Licence). Este editor permite realizar modificaciones en señales de audio digital en diversos formatos a través de una interfaz gráfica (incluyendo midi, wav, mp3, etc.).

[9] http://www.asterisk.org
[10] http://www.zoiper.com
[11] http://audacity.sourceforge.net/

Tarjeta Telefónica

Para lograr conectar una línea telefónica estándar a una computadora, se utilizó una tarjeta de telefonía análoga marca Digium, modelo 1TDM422EF (puerto PCI). Ésta permite conectar teléfonos y líneas telefónicas POTS convencionales (del inglés Plain-Old Telephone Service, Servicio Viejo de Teléfono) hacia o a través de una PC. Gracias a esto es posible utilizar algún software para crear ambientes de telefonía para diversos propósitos. En el caso del presente trabajo, se utilizó en conjunto con Asterisk.