Segmentación del Habla con Independencia de Texto para Reconocimiento Fonético

por

Luis David Huerta Hernández

Tesis sometida como requisito parcial para obtener el grado de **Maestro en Ciencias** en la especialidad de **Ciencias Computacionales** en el Instituto Nacional de Astrofísica, Óptica y Electrónica

Supervisada por:

Dr. Carlos Alberto Reyes García

Sta. Ma. Tonanzintla, Puebla

Febrero 2007

Resumen

Actualmente, en las Tecnologías del Habla se están considerando con mayor importancia las unidades de sub-palabras como los fonemas, puesto que para el proceso de reconocimiento estas unidades reducen la complejidad de modelado, de clasificación, y de almacenamiento de información de los lenguajes. El problema a resolver en esta Tesis de Maestría es la segmentación fonética del habla con independencia de texto. Este problema consiste en obtener las posiciones de las fronteras entre fonemas, a partir de la onda de habla sin el apoyo de ningún tipo de información conocida a *priori*, como lo es comúnmente el texto. Aunque se han reportado trabajos encausados a la segmentación en sub-palabras, éstos han sido probados bajo una serie de restricciones como dependencia de hablante [1] [2], texto [3][4], vocabulario [5] [6], sin hacer uso de habla contínua expresada naturalmente y sin considerar la sobre-segmentación [7]. Recientemente se reportó un método [8] que suprime todas estas restricciones alcanzando una tasa de detecciones correctas de límites del 73.58 % y una tasa de sobre-segmentación cercana al 0 %. Considerando que existen fronteras fonéticas vagamente definidas, el desempeño se incrementó haciendo uso de medidas difusas y distintas representaciones del habla, obteniendo mayor detalle de esas fronteras. Para el idioma inglés se incrementó la tasa de detecciones correctas en un 4 % respecto a [8], y para el idioma español, se detectaron aproximadamente un 80 % de límites fonéticos presentes en la señal del habla.

Abstract

Nowadays, Speech Technologies are considering with high importance the sub-words units like phonemes, because for the recognition process, these units reduce the model complexity, classification and storage of the languages information. The problem to solve in this Master Thesis is the phoneme speech segmentation with text independence. The problem consists on obtaining phoneme boundaries, from the speech wave without any kind of information known *a priori*, as it happens commonly with text. Although some related works have been reported oriented to carry out segmentation in sub-words, they have been tested under a set of restrictions as speaker independence [1] [2], text [3][4], vocabulary [5] [6], without continuos speech expressed naturally and without considering the over-segmentation [7]. Recently it was reported a method [8] avoiding all the previous restrictions mentioned, reach 73.58 % of correct segmentation and a over-segmentation near to 0 %. Considering te existence of phoneme boundaries vaguely defined, the performances was increased by using fuzzy measures and different speech representations, obtaining major detail of some boundaries. The performance on the English language was increasing in 4 % with respect to [8], and for Spanish language were detected approximately 80 % of phoneme boundaries present on the speech signal.

Agradecimientos

Mis sinceros agradecimientos a todo el grupo de investigadores del Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE), por que gracias a sus críticas, consejos y presiones me han transmitido sabiduría. Al personal administrativo y técnico, por todas las facilidades y atenciones otorgadas. A todos mis compañeros que en algún momento me dieron su apoyo incondicional.

Se agradece especialmente al Dr. Jesus Ariel Carrasco Ochoa, por su paciencia, consejos y atenciones. Al Dr. Aurelio López López por el apoyo otorgado para la realización de mi Maestría.

Gracias a mi asesor, el Dr. Carlos Alberto Reyes García por su motivación en mis momentos de flaqueza, por sus conocimientos y experiencia compartidos, consejos y atenciones.

Sobre todo, doy gracias a dios por darme la sabiduría y bendición a lo largo de mi vida, por ponerme obstáculos en mi camino para aprender de ellos, por iluminarme para salir adelante. A mi Familia y Esposa por todo su apoyo y paciencia.

A todas aquellas personas que no están incluidas de forma explícita, pero que contribuyeron en mi formación académica y humana.
A todos ustedes mis más sinceros y humildes agradecimientos.

Dedicatoria

Dedicado a mi hijo Bryan Emmanuel, a mi Esposa Rocío, a mis padres y hermanos

Índice general

Resumen i

Abstract iii

Agradecimientos v

Dedicatoria vii

Lista de Figuras xiii

Lista de Tablas xv

1. Introducción **1**

1.1. Problemática 3

1.2. Objetivos 4

1.3. Justificación 4

2. El Habla como Medio de Comunicación **5**

2.1. El Sistema Respiratorio 5

2.1.1. Respiración Normal sin Producción de Habla 5

2.1.2. Expiración en la Producción del Habla. 6

2.2. El Aparato Fonador Humano 7

2.3. El Proceso de Producción y Percepción del Habla 8

2.4. Información Transportada en la Onda del Habla. 9

2.4.1. Información Linguística 9

2.4.2. Información Paralinguística 11

2.4.3. Información no Linguística 12

3. El Reconocimiento del Habla **15**
3.1. Importancia del Reconocimiento del Habla 15
3.1.1. Por que es Importante el Estudio del Reconocimiento de Habla 15
3.1.2. Por que es Complicada la Segmentación del Habla 16
3.2. Dimensiones del RAH 18
3.2.1. Parámetros 18
3.3. Enfoques Segmentales para el Reconocimiento 19
3.3.1. Importancia de Unidades de Sub-palabras 20
3.3.2. Fonos y Fonemas 21
3.3.3. Otras Unidades de Sub-palabras 23
3.3.4. Frases 24
3.4. Resumen 24

4. El Proceso de Segmentación del Habla **27**
4.1. Preproceso y Segmentación 27
4.1.1. Preproceso 28
4.1.2. Segmentación 29
4.2. Segmentación y Detección de Límites 29
4.2.1. Detección de Bordes 32
4.3. Enfoques de la Segmentación Automática del Habla 34
4.4. Resumen 35

5. Trabajos Relacionados **37**
5.1. Segmentación Automática de Fonemas por Aplicación de Reconocimiento Vocal 37
5.1.1. Síntesis del lenguaje basado en sílabas 37
5.1.2. Decremento de Sonoridad 38
5.1.3. Algoritmo 38
5.1.4. Observaciones 39
5.2. Segmentación Fonética de Habla Contínua Usando un Perceptrón Multicapa (MLP) 39
5.2.1. Arquitectura MLP para la Segmentación en Fonemas 40
5.2.2. Algoritmo de Aprendizaje 40
5.2.3. Experimentos 41

5.3. Método para la Segmentación de Fonemas Independiente de Texto . . 41
5.3.1. Preprocesamiento 41
5.3.2. Detección de Límites Fonéticos 42
5.3.3. Experimentos 43
5.4. Integración de Segmentación Independiente del Lenguaje y Modelado Basado en Fonemas Dependientes del Lenguaje 43
5.4.1. Modificación al Algoritmo Original 43
5.4.2. Experimentos 44
5.5. Resumen 44

6. Características Acústicas del Habla 47
6.1. Características en el Dominio del Tiempo 47
6.1.1. Frecuencia fundamental 48
6.1.2. Amplitud 48
6.1.3. Energía 49
6.1.4. Intensidad 50
6.2. Características en el Dominio de Frecuencia 51
6.2.1. Espectros Mel 52
6.2.2. Coeficientes Cepstrales de Frecuencia Mel 53
6.3. Resumen 56

7. Algoritmos de Segmentación 57
7.1. Introducción 57
7.2. Bases de Datos 58
7.2.1. Base de Datos DIMEx100 58
7.2.2. Base de Datos TIMIT 58
7.2.3. Datos Experimentales 59
7.3. Evaluación de Desempeño 59
7.4. El Filtro Pre-énfasis 60
7.5. Medidas Difusas Utilizadas en Algoritmos de Segmentación 61
7.5.1. Distancia Euclidiana 62
7.5.2. Distancia Manhattan 62
7.5.3. Distancia de Correlación Pearson 62
7.5.4. Distancia de Chebyshev 63

7.6. Algoritmos de Segmentación con Características en el Dominio del Tiempo 64
7.6.1. Algoritmo básico de segmentación con detección de bordes . . 64
7.6.2. Algoritmo con Medidas de Distancias Difusas de la Intensidad 67
7.7. Algoritmos de Segmentación con Características Vectoriales 72
7.7.1. Experimentos 74
7.8. Conclusión y Discusión 75
7.9. Discusión de Resultados 76
7.9.1. Análisis de Resultados 76

8. Conclusiones y Perspectivas **81**
8.1. Trabajo Futuro 82

A. Apéndice **83**
A.1. Publicaciones 83

Bibliografía **85**

Lista de Figuras

1.1. Esquema general para el reconocimiento de habla contínua. 3

2.1. Tracto vocal 7
2.2. Cadena de comunicación de producción a percepción de habla. 9

4.1. El objeto y el fondo que lo rodea 29
4.2. Las unidades de habla con sus respectivos grados de inconsistencia y dificultad de segmentación 31

5.1. (a)Intensidad acústica, b)Variación acústica, c)Decremento de sonoridad 38
5.2. Secuencia de puntos de segmentación, después del Nivel 1 de segmentación denotadas por O y puntos incluidos por el Nivel 2 de segmentación denotada por * 44

6.1. Amplitud del movimiento 48
6.2. Límites fonéticos sobre valores de la amplitud 49
6.3. Límites fonéticos sobre valores de la energía 50
6.4. Límites fonéticos sobre valores de la intensidad 51
6.5. Diagrama de un extractor de características con bancos de filtros . . . 52
6.6. Banco de fltros Mel 53
6.7. Descomposición de una señal de habla por Filtros de Mel 54
6.8. Forma de onda, MFCC, y Espectrograma 55

7.1. Intensidad con y sin pre-énfasis 61
7.2. Diagrama del algoritmo de segmentación con detección de bordes . . 65
7.3. Diferencias absolutas de las medias de la amplitud, y sus respectivos parámetros 66

7.4. Diagrama a bloques del algoritmo con membresias difusas de la intensidad 69
7.5. Diagrama a bloques del algoritmo con membresias difusas en sub-bandas 73
7.6. Puntos de segmentación del mejor caso TIMIT 77
7.7. Puntos de segmentación del mejor caso DIMEX 77
7.8. Puntos de segmentación del peor caso Timit 78
7.9. Puntos de segmentación del peor caso Dimex 78

Lista de Tablas

7.1. Parámetros empleados en cada una de las características sobre el corpus DIMEx100 66
7.2. Desempeño del algoritmo usando distintas características sobre el corpus DIMEx100 66
7.3. Desempeño del algoritmo usando distintas características sobre el corpus TIMIT 67
7.4. Desempeño del algoritmo sin utilizar membresias difusas 70
7.5. Desempeño del algoritmo, con distintos tamaños de frames y sus respectivos parámetros 70
7.6. Desempeño del algoritmo con membresias difusas normalizadas 71
7.7. Desempeño del algoritmo con frames no adyacentes 71
7.8. Desempeño del algoritmo utilizando intensidad mínima de 25 dB . . . 71
7.9. Desempeño del algoritmo utilizando distintas medidas de disimilaridad sobre el corpus TIMIT 72
7.10. Desempeño del algoritmo utilizando distintas medidas de disimilaridad sobre el corpus DIMEx100 72
7.11. Desempeño del algoritmo con características vectoriales sobre el corpus TIMIT 74
7.12. Desempeño del algoritmo con características vectoriales sobre el corpus DIMEx 75
7.13. Parámetros de desempeño en los mejores casos en TIMIT y DIMEX . 78
7.14. Parámetros de desempeño en los peores casos en TIMIT y DIMEX . 79

Capítulo 1

Introducción

En el transcurso de la historia, el ser humano ha ido perfeccionando sus medios y herramientas de comunicación para expresar pensamientos e ideas, y de esta manera se ha ido contribuyendo a la socialización, transmisión de conocimientos y lo más importante, a la supervivencia. Sin duda, a medida que la comunicación entre los seres humanos ha evolucionado, se ha impulsado el desarrollo en todas las facetas de la humanidad. Actualmente, la expresión oral es considerada el medio de comunicación natural por excelencia, de la cual, el ser humano se vale para realizar sus actividades cotidianas y sobrevivir en la sociedad.

Con la aparición, generalización e inclusión de computadoras en dispositivos electrónicos industriales y domésticos, el ser humano tiene la necesidad de comunicarse con ellas de una manera eficiente. Se han desarrollado tecnologías para la interacción hombre-máquina que han ido incluyendo progresivamente medios naturales de comunicación. Las tecnologías que buscan como objetivo genérico proveer a las máquinas de conocimientos necesarios (en forma de algoritmos, modelos) para hacerlas comunicarse con los humanos de la forma más natural para los mismos, mediante el lenguaje oral, son conocidas como Tecnologías del Habla [9].

Actualmente, se están realizando esfuerzos en la Tecnología del Habla para tener un mejor entendimiento y manipulación de la expresión oral por ser la forma más espontánea y natural de comunicación entre las personas. La Tecnología del Habla es tomada como uno de los factores determinantes en la mejora de la interacción entre personas y computadoras. La progresiva integración de la voz como interfaz de comunicación entre los hombres y las máquinas, permite aumentar la cooperación con los

sistemas informáticos, aprovechando en gran medida los servicios que estos sistemas proporcionan con una mayor rapidez y eficiencia. La telematización de la computadora personal, así como la llegada de teléfonos móviles de tercera generación, hacen que los sistemas de reconocimiento de habla fiables sean una necesidad determinante. Aunque actualmente se han alcanzado avances significativos en las tecnologías del habla, aun estamos muy lejos de un Sistema de Reconocimiento Automático de Habla que tenga un buen desempeño en la aplicación para la que es destinado. Se han desarrollado aplicaciones de Reconocimiento Automático de Habla con algunas restricciones de vocabulario, de locutor, disfluencia por mencionar algunos.
Dentro de las Tecnologías del Habla se encuentran varias tecnologías que tienen como objetivo, el tratamiento del lenguaje oral. Una de estas tencologías es conocida como Reconocimiento Automático de Habla (Automatic Speech Recognition), que a su vez se puede dividir en tres importantes campos de aplicación:

- *Sistemas de Transcripción de Voz*. Son frecuentemente utilizados en oficinas, en los que por medio de un micrófono el sistema transforma las palabras en su correspondiente transcripción textual que aparecen en la pantalla de la computadora. Estas aplicaciones, también conocidas como *Sistemas de Dictado*, se apoyan de un diccionario de palabras y habrá que enseñarles aquellas que no conocen. El funcionamiento de estos sistemas se basa en el modelado del lenguaje, se puede mencionar a Dragon System Naturally Speaking de Nuance Communications e IBM ViaVoice dentro de los comerciales.

- *Sistemas de Comando y Control (Interactive Voice Response)*. Suelen controlar microprocesadores en vehículos, software de consultas bancarias entre otros. Importantes aplicaciones se encuentran en el campo de la telefonía, donde se proporciona al usuario una interfaz para conectarlo con bases de datos y recuperar información de su interés. La tecnología IVR ha alcanzado importantes avances como para ser adecuados para muchas aplicaciones. Algunas aplicaciones pioneras podemos mencionar como: reserva y consulta de horarios de viajes en líneas aéreas, consultas bancarias. El inconveniente de estas aplicaciones es su baja adaptabilidad a nuevas tareas , y en general de aquellas que reciben como entrada la voz es la alta sensibilidad al ruido ambiental.

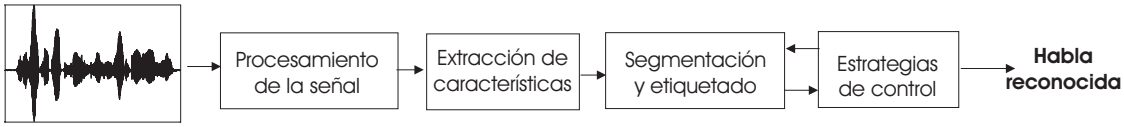

Figura 1.1: Esquema general para el reconocimiento de habla contínua.

- *Sistemas de Segmentación y Etiquetado Automático del Habla*. Estos sistemas han sido demandados por muchas aplicaciones que necesitan grandes cantidades de habla segmentada y etiquetadas fonéticamente, donde el etiquetado se define como la identificación de unidades de habla con sus respectivos tiempos de inicio y fin contenidas en las oraciones habladas. De hecho, el aprendizaje de los Reconocedores Automáticos de Habla hace uso de segmentos de oraciones etiquetados fonéticamente. Incluso, los segmentos son utilizados para la conversión de texto a voz sintética.

1.1. Problemática

La mayoría de los sistemas pioneros de Reconocimiento Automático del Habla que han reportado algún éxito han sido probados en vocabularios pequeños. Aunque el tamaño del vocabulario no representa directamente la dificultad del proceso de reconocimiento, existen una serie de problemas que se presentan cuando el tamaño del vocabulario se incrementa. Otra dificultad es el aumento en la complejidad de la búsqueda, mientras que en vocabularios pequeños es posible realizar búsquedas menos complejas. En vocabularios pequeños es posible modelar cada palabra, y almacenar sus parámetros. Por otro lado, cuando el tamaño del vocabulario aumenta, se hace imposible el entrenamiento de cada palabra explícitamente, debido a que tanto el entrenamiento como la cantidad de almacenamiento consumirían demasiados recursos computacionales. En esta situación, se tiende a utilizar otras unidades de habla generalmente en la forma de sub-palabras como sílabas, difonos y fonemas.
Esta tesis de maestría se desarrolla en el campo de la Inteligencia Artificial conocida como Reconocimiento Automático del Habla, específicamente en la obtención de sub-palabras por medio de los Sistemas de Segmentación Automática.

1.2. Objetivos

El objetivo principal de esta tésis es mejorar un aspecto concreto del Reconocimiento Automático del Habla, la segmentación. Se desarrollará un algoritmo para la segmentación fonética del habla que soporte la independencia de texto, de hablante y de vocabulario, el habla contínua y la naturalidad de la misma. El algoritmo deberá minimizar el fenómeno de sobre segmentación para lograr un desempeño de alta calidad.

1.3. Justificación

La segmentación de habla contínua en sus correspondientes fonemas ha llegado a ser un tema muy importante en muchas areas de procesamiento de habla como el reconocimiento, análisis, síntesis, y bases de datos de habla. La exactitud y confiabilidad de la segmentación automática de fonemas es un factor crucial para un adecuado reconocimiento, y por lo tanto, para el adecuado funcionamiento de sistemas completos. Actualmente, muchos sistemas de reconocimiento de habla estan usando fonemas como unidades de habla por que proporcionan las siguientes ventajas: 1) Los fonemas son lingüísticamente bien definidos y pueden ser buscados fácilmente en diccionarios; 2) la variabilidad de pronunciación debido al contexto lingüístico, acento o diálogos pueden ser fácilmente representados por la aplicación de reglas en base a formas; 3) el número de unidades es pequeño; y 4) los fonemas requieren significativamente menos datos para entrenar, de los que deberían ser necesarios para el modelado de una palabra completa [5][10].
Se han hecho pocos esfuerzos por desarrollar métodos para la segmentación de una onda de habla a nivel de fonemas, sin las restricciones que otros algoritmos han presentado como la dependencia de hablante, habla discreta, y la dependencia de texto. Por otro lado, la cuantificación del fenómeno de sobre-segmentación no ha sido usada como una de las métricas de calidad en muchos algoritmos de segmentación previos, por lo que han llegado a reportar altas tasas de correcta segmentación con tasas de sobre-segmentación de hasta 63 % [3].

Capítulo 2

El Habla como Medio de Comunicación

En este capítulo se abordarán los conceptos básicos relacionados con el reconocimiento de habla, así como los conceptos esenciales en los medios de producción y percepción del habla humanos.
Es importante considerar que el reconocimiento de habla contínua hace uso del proceso de segmentación en unidades manejables de habla como pueden ser palabras, sílabas o fonemas, de tal manera que las dificultades del proceso de reconocimiento mismo son heredadas por el proceso de segmentación.

2.1. El Sistema Respiratorio

En esta sección se revisará de forma breve, los mecanismos respiratorios, y como ellos influyen en la producción del habla.

2.1.1. Respiración Normal sin Producción de Habla

La respiración es una actividad que da oxígeno al cuerpo, y expele dióxido de carbono de él. Los sonidos de habla son producidos usando el flujo de aire expiratorio. Cuando el flujo de aire pasa por las cuerdas vocales, se genera una vibración quasi-periódica como la fuente de sonido.
Durante el respiro normal, la inspiración es realizada por la contracción de los músculos intercostales externos para elevar las costillas hacia arriba y afuera, con el fin de

agrandar el volumen de los pulmones con la ayuda del movimiento hacia abajo del diafragma. En expiración, por otro lado, los músculos intercostales internos mueven las costillas hacia abajo y hacia adentro con el fin de decrementar el volumen de los pulmones con la ayuda de la contracción del músculo abdominal y el elástico retroceso de los pulmones, la viscera abdominal, y costillas. La respiración es producida por el control recíproco de activación y relajación de estos dos grupos de músculos para inspiración y expiración. La respiración es importante en el proceso de producción del habla, puesto que el habla nace a partir de una fuente de aire (energía) generada por los pulmones.

2.1.2. Expiración en la Producción del Habla.

La producción del habla es realizada durante la fase de expiración, sin embargo, esto no significa que solo los músculos expiratorios estan involucrados en la producción del habla; ambos músculos para expiración e inspiración son cooperativamente activados durante la expiración, para mantener estable la presión subglotal. El volumen de los pulmones cae gradualmente, mientras se mantiene casi el mismo flujo de aire, y la presión subglotal es mantenida casi al mismo nivel durante el habla. Sin embargo, para producir este flujo constante de aire, los músculos para la inspiración son activados hasta le momento en que la presión de relajación alcanza el nivel de presión subglotal. Entonces los músculos expiratorios son gradualmente activados para después reducir el volumen de los pulmones, extendiendo la exhalación. Estudios recientes de cinemática y dinámica para monitorear los movimientos anterior-posterior y vertical de las paredes del pecho, demostraron que la actividad muscular abdominal continua en toda la fase de expiración en la respiración del habla. Estas actividades cooperativas de músculos juegan un papel muy importante en controlar la diferencia de presión transglotal y también como en la tasa de flujo de aire, produciendo la apropiada intensidad del sonido de habla.
La intensidad es una característica acústica que puede obtenerse en el dominio del tiempo, y que será considerada para el desarrollo de uno de los algoritmo de segmentación.

2.2. El Aparato Fonador Humano

La señal de voz es transmitida principalmente a través de un canal que está constituido por ondas de presión que se propagan a través del aire. El aparato fonador está constituido básicamente de tres elementos: un generador de energía (pulmones), un sistema vibrante (laringe y cuerdas vocales), y cavidades resonantes constituidas por el tracto vocal y el tracto nasal. El tracto vocal es el espacio comprendido entre las cuerdas vocales o glottis y los labios. El tracto vocal consiste de la faringe y la boca, o cavidad oral. En promedio, la longitud del tracto vocal es de 17 cm, 14 cm y 10 cm para hombres, mujeres y niños respectivamente. El área representativa del tracto vocal, determinada por la posición de la lengua, labios, mandíbulas y velo, varía desde cero cuando está completamente cerrada hasta 20 cm^2 aproximadamente. El tracto nasal empieza en el velo y termina en los orificios nasales. Cuando el velo es bajado, el tracto nasal es acústicamente unido al tracto vocal para producir los sonidos nasales del habla.

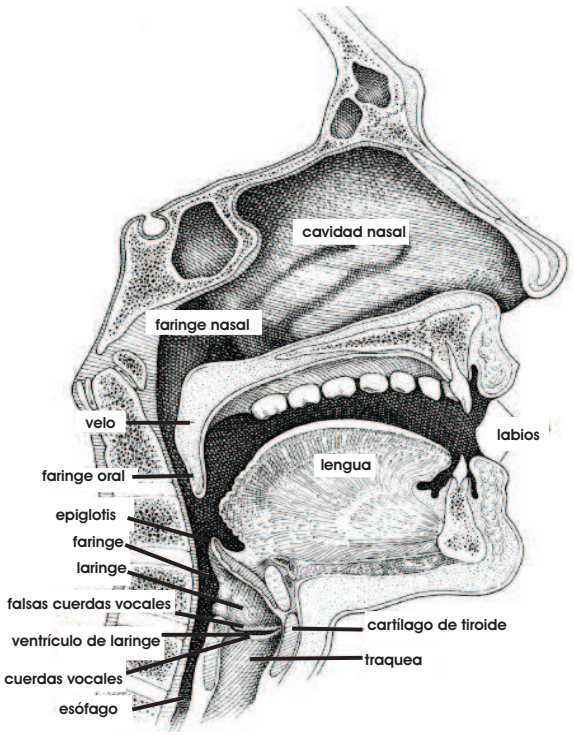

Figura 2.1: Tracto vocal

2.3. El Proceso de Producción y Percepción del Habla

Una señal de habla es una forma de onda acústica que transporta varios tipos de información útiles para la comunicación humana, de hecho, el habla ha sido un canal central de telecomunicaciones desde la llegada de la telefonía.
El proceso de producción del habla inicia cuando el hablante estructura un mensaje en su mente para transmitirlo al oyente por medio del habla. El próximo paso es la codificación del mensaje en unidades lingüísticas como palabras o fonemas. Posteriormente el hablante debe de ejecutar una serie de comandos neuromusculares para hacer vibrar las cuerdas vocales cuando sea necesario, dar forma al tracto vocal para que la secuencia de sonidos de habla sean creados y dichos por el hablante. Los comandos neuromusculares deben simultáneamente controlar todos los aspectos de movimientos articulatorios de los labios, mandíbulas, lengua y velo. Dependiendo de la posición de los articuladores, diferentes sonidos son producidos.

La generación de sonidos de habla inicia cuando introducimos aire a los pulmones vía el mecanismo normal de respiración. Los pulmones proporcionan una diferencia de presión necesaria para crear el flujo de aire que activará la laringe y las demás cavidades del tracto vocal. El flujo de aire es convertido a pulsos *quasi*-periódicos, los cuales son modulados en frecuencia a través de la laringe, la cavidad vocal, y la cavidad nasal. La frecuencia de vibración de las cuerdas vocales es la frecuencia fundamental o *pitch* del sonido producido. Las resonancias originadas en el conducto vocal, producen señales acústicas en las que la energía está concentrada en mayor o en menor grado alrededor de las frecuencias de resonancia correspondientes. A estas concentraciones se les conoce con el nombre de *formantes*. Los formantes contienen la mayor parte de la información acústica transportada por la señal vocal.
Una vez que la señal de habla es generada y propagada al oyente, el proceso de percepción del habla inicia. La señal acústica es procesada a lo largo de la membrana basilar en el oído interno, el cual provee un análisis espectral a partir de la señal entrante. El proceso de traducción neural convierte la señal espectral de la salida de la membrana basilar en actividades de señales en el nervio auditivo, correspondiente al proceso de extracción de características. Esta actividad neural, a lo largo del

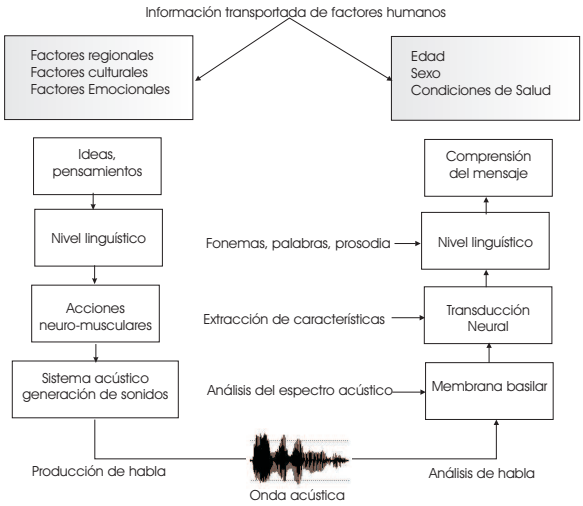

Figura 2.2: Cadena de comunicación de producción a percepción de habla.

nervio auditivo es codificada en unidades lingüísticas produciendo un entendimiento del mensaje hablado.

2.4. Información Transportada en la Onda del Habla.

Existen muchos factores durante la producción del habla que influyen en la variabilidad de sus características, ampliando la gama de clases en proceso de reconocimiento, y por lo tanto, también facilitan o dificultan el proceso de segmentación. Por otro lado, estos factores pueden ser explotados para obtener información útil relacionada con el hablante. La información incluida en los sonidos del habla puede ser dividida en términos de su contenido en tres categorías: lingüística, paralinguística y no linguística.

2.4.1. Información Linguística

El principal objetivo de la comunicación humana hablada es transferir información linguística. La información linguística puede ser definida como "información simbólica que es representada por un conjunto de símbolos discretos y reglas para su combi-

nación". Sin embargo, la información linguística es discreta y categórica. Una característica importante de la información linguística, en contraste con la no linguística, es que la primera puede ser controlada por el hablante. Como transportadora de información linguística, una *característica segmental* juega un papel importante, por que es usada para distinguir el significado discreto de palabras.

Características Segmentales

Suponga que un hombre desea transmitir su amor a una mujer, y ambos son hablantes nativos de español. Para componer la sentencia "Yo te amo", él debe seleccionar tres palabras: "Yo", "te", y "amo". En esta etapa, cada palabra es seleccionada como una unidad que tiene una única correspondencia de significado y función. Sin embargo, cada palabra puede ser dividida jerárquicamente en pequeños segmentos: Una palabra es dividida en sílabas, y cada sílaba puede ser dividida en fonemas. Los fonemas son los segmentos más pequeños del sonido que funcionan en un lenguaje para señalar las diferencias en el significado. En este sentido, los fonemas juegan un papel importante en transferir información linguística en la comunicación del habla humana.
El *fono* es también usado para determinar los segmentos más pequeños de sonido. Un fono es un sonido particular del habla categorizado discretamente de acuerdo a muchas características acústicas del sonido. Estas características son llamadas segmentarias, y son determinadas por una única combinación del proceso de producción del habla independiente, tal como el lugar y manera de articulación, contrastes de voz contra ausencia de voz.
La diferencia entre *fono* y *fonema* es que el fonema es una familia de *fonos* los cuales tienen una idéntica función en un cierto lenguaje. Por ejemplo, los sonidos plosivos aspirados sordos [th] y el no aspirado [t], son diferentes fonos, pero ellos son alófonos del mismo fonema /t/ en inglés.

Características Suprasegmentales

Características suprasegmentales, o prosódicas, son características acústicas distribuidas sobre una secuencia de muchos segmentos, tal como fonemas, sílabas, palabras , frases y oraciones. Las características acústicas que transfieren características suprasegmentales incluyen la frecuencia fundamental de vibración glotal para

vocales y consonantes sonoras, la intensidad del habla, y las características temporales de segmentos. Las características suprasegmentales juegan un papel principal en la transferencia paralinguística y no linguística de la información. Sin embargo, las características citadas arriba, juegan un papel importante en la transferencia de información linguística.
El acento de palabras es un ejemplo de características suprasegmentales usadas para alterar el significado de palabras por controlar la frecuencia fundamental (F0), la cual es la frecuencia de vibración de las cuerdas vocales, correlacionada con la intensidad. En español, estos cambios de entonación son expresados con símbolos llamados precisamente *acentos*. Es necesario hacer notar que las restricciones fisiológicas de producción de habla son importantes para mantener un canal robusto en la transferencia de información acentuada. Puesto que la frecuencia e intensidad de la vibración de cuerdas vocales tienen alta correlación, la intensidad del sonido puede ser cambiada al ser ajustada la F0, y la F0 puede ser influenciada por la alteración del estrés o por la acentuación [11]. Estas influencias mutuas entre características acústicas correlacionadas pueden ser usadas para preservar la confiabilidad de la transferencia de información acentuada.

2.4.2. Información Paralinguística

La información paralinguística es definida como "información que no es inferible de una contraparte escrita, pero es deliberadamente agregada por el hablante para modificar o suplementar información linguística". La información paralinguística puede tener características discretas y continuas. Por ejemplo, un hablante puede controlar una oración así como categorizarla como una oración declarativa o interrogativa. Tales cualidades como el grado de suspicacia del hablante puede ser representado continuamente en la elocución de una pregunta. Para producir estas funciones de información paralinguística, el hablante principalmente controla características suprasegmentales, pero las características segmentarias también entran en juego.
Por ejemplo, cuando se está conversando con un amigo, se puede hablar rápidamente y no articular las palabras claramente. Por el otro lado, cuando hablamos con una persona mayor quien podría tener dificultades auditivas o a una persona que no habla nuestro lenguaje nativo muy bien, hablaremos de manera lenta y claramente articulado. Durante la conversación, podemos elevar la F0 y/o hacer énfasis sobre un

segmento de una palabra para hacer el tema de conversación claro o para responder a una pregunta. La manipulación de la F0 es también usada para transferir la intención de oraciones habladas. Por ejemplo, la F0 de una oración interrogativa se incrementa en la parte final de la oración.

2.4.3. Información no Linguística

La información no lingüista es responsable de factores idiosincrásicos y estados emocionales del hablante. Generalmente, el hablante no puede controlar estos factores, aunque es posible imitar algunas características de estos factores.

Factores Idiosincrásicos

Las características acústicas del habla son influenciadas por las características idiosincrásicas tal como edad, sexo, características morfológicas individuales, condiciones de salud, y en algunos casos las discapacidades. Por ejemplo, el sexo de los hablantes adultos es identificable con solo escuchar la voz. Una octava de diferencia en la F0 (que es de aproximadamente 100 Hz para hombres vs. 200 Hz para mujeres) es una importante indicación para la identificación del sexo. Puesto que el rango de F0 para niños es frecuentemente similar a las mujeres adultas, así no siempre aplica para identificar la edad (identificable con las F1 y F2).
Las características acústicas del habla también varían debido al envejecimiento, donde se han hecho experimentos indicando que la F0 converge en un rango similar tanto para la voz masculina y femenina. Estos cambios en las características acústicas pueden originarse por cambios morfológicos en las cuerdas vocales.
Las características acústicas de las cavidades nasales y paranasales juegan un papel muy importante para la evaluación de la salud del hablante. Es bien sabido que la calidad de voz cambia cuando una persona esta resfriada, esto es debido a la inflamación parcial de la membrana mucosa de la cavidad nasal.

Factores Emocionales

Es bien conocido que condiciones emocionales como enojo, tristeza y cansancio pueden tener efectos sobre el sonido del habla. Estos efectos pueden ser observados principalmente en las características suprasegmentales tales como la F0, intensidad, y

características de habla temporales. Puesto que la tensión de los músculos puede ser ocasionada por algunas condiciones emocionales, existe la posibilidad que algunas características segmentales sean también influenciadas por las condiciones emocionales del hablante.
Hoy en día, las emociones presentes en la señal del habla están siendo objeto de estudio. Algunos experimentos han extraído características acústicas del habla conteniendo características emocionales y sus efectos en la percepción. Sin embargo, el material del habla usado en estos estudios ha sido habla imitada para replicar condiciones emocionales. En el sentido de investigar características emocionales que usualmente no pueden ser controladas por el hablante, los datos deben ser tomados de conversaciones naturales, fuera de ambientes experimentales.

Resumen

La producción del habla es un proceso en el que intervienen factores tanto físicos, ideológicos y emocionales del hablante. El habla, no solo transporta el mensaje explícito estructurado por el hablante, si no información adicional como estado de salud, estado de ánimo por mencionar algunos importantes. El RAH, y por lo tanto la segmentación deben lidiar con estos factores que afectan el desempeño del sistema en su totalidad. Puesto que es complicado controlar dichos factores, la inclusión de una amplia gama de hablantes incrementa la posibilidad de experimentar con ellos y obtener resultados más realistas. En el Capítulo 8 se da una breve descripción de las bases de datos utilizadas para llevar a cabo la segmentación fonética.

Capítulo 3

El Reconocimiento del Habla

3.1. Importancia del Reconocimiento del Habla

En esta sección se hará un repaso de la importancia del reconocimiento automático del habla, así como los retos que esto representa.

3.1.1. Por que es Importante el Estudio del Reconocimiento de Habla

Por que estudiamos Reconocimiento Automático del Habla (RAH)?, una de las razones, es que hay muchos intereses económicos de por medio: el reconocimiento del habla es una industria de multi billones de dólares actualmente. Como en 1998, las ganancias de aplicaciones simples de telefonía son reputadas en cientos de millones de dólares por año.

Hay muchos aspectos del reconocimiento del habla que son bien entendidos. Sin embargo, es claro que hay muchos que aún no se conocen. Nuestros métodos y tecnología, no tienen la calidad de los humanos para el reconocimiento del habla; el funcionamiento se degrada considerablemente cuando diversos factores afectan la señal del habla, algunos de estos efectos pueden ser provocados, por ejemplo, por el cambio de micrófonos.

El reconocimiento del habla puede potencialmente ser muy utilizado. Algunas aplicaciones son las siguientes:

Aplicaciones Telefónicas. Para la mayoría de las aplicaciones actuales de correo de

voz, uno tiene que seguir una serie de botones a través de un menú jerárquico. El reconocimiento del habla tiene la potencial forma de hacerlo directamente a través del menú jerárquico.

Operación del Manos Libres. Hay muchas situaciones en la cual las manos no están disponibles para enviar comandos a los dispositivos. Usando un teléfono de automóvil y controlando la posición del microscopio en un cuarto de operación son dos ejemplos para los cuales sistemas de vocabulario limitado ya existen.

Aplicaciones para Discapacitados. El reconocimiento de habla es una alternativa natural de interfaz en computadoras para gente con limitada movilidad en sus brazos y manos, o para aquellas con limitaciones de vista. Para algunos aspectos de aplicaciones de computadoras, el habla podría ser una interfaz más natural que el teclado o el ratón.

Dictado. Dictado general es una aplicación avanzada, requiere de un vocabulario muy grande, que reemplaza a sistemas de menú. Desde 1998, hay muchos sistemas de dictado en el mercado que aceptan habla continua en entrada, como el actualmente conocido sistema Dragón de IBM.

Traducción. Otra aplicación avanzada es la traducción de un lenguaje a otro. El proyecto Verbmobil de Alemánia, es un esfuerzo tanto colaborativo como competitivo que da una traducción lenguaje a lenguaje. La tarea es facilitar una conversación entre hablantes nativos de Alemán y Japonés, usando Inglés como un lenguaje intermedio; el sistema es para actuar como un asistente de un participante Alemán, traduciendo palabras y frases como se necesiten del Alemán en Inglés.

3.1.2. Por que es Complicada la Segmentación del Habla

Hay muchas razones por lo que el reconocimiento del habla es muy difícil.

Primero, el habla natural es contínua; frecuentemente no tiene pausas entre las palabras. Esto hace difícil de determinar, donde se encuentran los límites de las palabras, entre otras cosas. También, el habla natural contiene disfluencias. Los hablantes cambian los pensamientos acerca de lo que quieren decir a la mitad de las oraciones, y accidentalmente intercambian fonos, o generan oraciones llenas de pausas con expresiones sin significado (p.e: "mmm", .^{eh}hh", etc.) mientras estan pensando el siguiente mensaje.

Segunda, el habla natural puede también cambiar por diferencias en parámetros globales (distintos hablantes) o locales(un hablante) de habla. La manera de pronunciar las palabras de un hablante a otro pueden variar debido a diversos factores como características físicas del tracto vocal; o aquellas psicológicas como el estrés o estado de ánimo del hablante, o simplemente por el contexto en que se pronuncian, por mencionar algunos. Como resultado, el espectro del habla cambiará, y con frecuencia muy dramáticamente, si alguna de esas condiciones son cambiadas.

Tercera, grandes vocabularios son frecuentemente confusos. Un vocabulario de 20,000 palabras, tendrá mas probabilidad de contar con palabras parecidas entre sí que aquellos vocabularios de solo 10 palabras. Existe también la situación de hacer uso de palabras que esten fuera del vocabulario; para algunas tareas, no importa que palabras estén en el vocabulario, el reconocimiento siempre encontrará las palabras que no han sido vistas antes. Como modelar esas palabras desconocidas es un problema importante no resuelto.

Cuarta, como se hace notar previamente, las grabaciones del habla son variables sobre cuartos acústicos, características del canal de transmisión, características del micrófono, y el ruido de fondo. En habla telefónica, el canal usado por la compañía en una llamada particular, tendrá efectos espectrales y temporales sobre la señal del habla transmitida. Ruido de fondo y acústica del ambiente en el que un hablante de teléfono se encuentra, también tendrá efectos tangibles en la señal. Diferentes microteléfonos, o en general, diferentes micrófonos, tienen diferente frecuencia de respuesta; la inclinación de un teléfono también cambiará la frecuencia de respuesta. Efectos no lineales son particularmente significantes en micrófonos de carbón granulado, pero en general, ellos pueden complicar los efectos usando algún microteléfono en particular. Algunos efectos serán dependientes del teléfono; por ejemplo, los sonidos nasales pueden ser más ruidosos si los micrófonos se encuentran más cercanos a la nariz.

Todos lo factores antes mencionados pueden cambiar las características de la señal del habla. El sistema auditivo humano puede frecuentemente compensar estos cambios de características en la señal del habla percibida, mientras que los sistemas de

reconocimiento del habla artificiales no.
Los algoritmos para el sistema de entrenamiento de reconocimiento deben ser cuidadosamente escogidos, por que grandes tiempos de entrenamiento no son prácticos para propósito de investigación. Existen algoritmos que toman demasiado tiempo de ejecución sobre hardware disponible y pueden ser de gran interés teórico, pero puesto que los programas tienen fallas, son una opción que realmente no permite el desarrollo de un enfoque experimental.

3.2. Dimensiones del RAH

Una vez expuestas algunas dificultades del reconocimiento del habla, estas se pueden dimensionar de alguna manera.

3.2.1. Parámetros

Un calificador de la tarea de un reconocedor automático del habla es la dependencia o independencia del hablante. Un sistema dependiente del hablante es entrenado con un hablante particular y probado con él mismo. Por otro lado, un sistema independiente del hablante es entrenado con muchos hablantes y probado sobre un conjunto diferente de hablantes. Esta tesis aborda el problema de segmentación del habla incluyendo la independencia del hablante, sin embargo, el proceso de entrenamiento para llevar a cabo la segmentación será evitado.
Sistemas de grandes vocabularios para usarse en computadoras personales han tendido a ser dependientes de hablantes para una mayor exactitud. Aunque muchos sistemas han sido independientes de hablante (al menos han sido entrenados en diferentes hablantes y probados en otros), muchos de ellos funcionan muy pobremente en hablantes que no son nativos del lenguaje objeto.
Otro importante descriptor, es si la tarea es reconocer habla aislada o habla contínua. El primer tipo de tarea es reconocer palabras en aislamiento (delimitadas por silencio) y es en general menos difícil que reconocer habla contínua, en la cual los límites de palabras no son tan aparentes. Los algoritmos de segmentación del habla implementados en esta tesis tratarán con el habla continua.
El tamaño del vocabulario también introduce otro parámetro. En general, un vocabulario de 20, 000 palabras es más difícil de manejar que un vocabulario de 10 palabras.

Esto es parcialmente debido a que hay una gran variabilidad en la acústica asociada con cada tipo de sonido del habla, pero también por que el vocabulario contiene muchas más palabras confusas unas con otras.
El estilo del habla tiene una fuerte influencia en la dificultad del reconocimiento del habla. De hecho, el habla conversacional es extremadamente difícil de transcribir. El habla que es leída de un texto preexistente es comparativamente fácil. Fluidez del habla en el diálogo entre máquina-humano es de dificultad media; un usuario motivado tenderá a hablar más claramente, pero el uso del habla fluida será más difícil de reconocer que habla leída. Algunas de las características incluidas en el estilo del habla natural es la amplia variabilidad en la tasa del habla, incremento en la disfluencia tal como muchas pausas y falsos inicios, y una gran variabilidad en el esfuerzo vocal.
Las condiciones de grabado también juegan un papel importante en determinar la dificultad de un reconocedor automático del habla. La grabación puede tener un rango de ancho de banda desde el incluido en micrófonos de alta calidad hasta el de los micrófonos de celulares. El canal telefónico tiene típicamente un ancho de banda menor a 4 kHz, en el cuál significa que las consonantes de alta frecuencia son más difíciles de distinguir, tales como /f/ y /s/, debido a que su mayor concentración de energía se encuentra arriba de los 4 kHz.
La telefonía del habla también introduce otros retos. El rango de hablantes que tienen acceso al habla telefónica tienen una gran variabilidad que es observada en bases de datos de laboratorios. Hay también una gran variabilidad en el ruido de fondo, y se debe tratar la distorsión de canales como ecos, cruce de hablantes, diferencias en las características espectrales del manos libres, y el canal de comunicación en general. Esas fuentes de variabilidad son particularmente un problema para teléfonos celulares.

3.3. Enfoques Segmentales para el Reconocimiento

Las unidades lingüísticas más comunes para el RAH han sido palabras; en donde programación dinámica es usada para encontrar la mejor coincidencia con la plantilla de referencia asociada con palabras candidatas. Así que, las distancias locales no fueron asociadas con ninguna otra clase lingüísticamente definida que palabras. Sin

embargo es notado que las palabras son típicamente elementos estructurados, consistiendo de unidades sub-palabras que son comunes en muchas otras palabras. En principio, esta estructura puede y debería ser usada para mejorar la efectividad de alguna cantidad finita de datos de entrenamiento.
Los sistemas de reconocimiento del habla basados en la programación dinámica, como los que usan unidades de sub-palabras principalmente han sido usados para sistemas estadísticos. Sin embargo en los 1970s y 1980s un gran número de sistemas fueron construidos haciendo uso extensivo del conocimiento fonético-acústico, alguno de los cuales usaron clasificación estadística y enfoques deterministas. En tales casos, unidades de sub-palabras fueron usadas.
Probablemente los mejores sistemas conocidos fueron esfuerzo del grupo de Ron Cole en CMU(Carnegie Mellon University) [12], OGI(Oregon Graduate Institute of Science & Technology) [13] [14] y recientemente en la Universidad de Colorado [15], y trabajos relacionados en MIT por Victor Zue [16] y colegas. Estos sistemas incorporaron explícito conocimiento fonético-acústico para segmentar el habla separadamente en segmentos fonéticos, y después clasificarlos, siendo los fonos o fonemas las clases a ser identificadas. Incorporando unidades más pequeñas que las palabras, los parámetros aprendidos fueron compartidos entre muchas palabras. Incorporando explícitas reglas de decisión para ejemplos fonéticos específicos, los diseñadores no se limitaron a definir una simple métrica de distancia para todos los tipos de decisiones.

3.3.1. Importancia de Unidades de Sub-palabras

La mayoría de los sistemas de reconocimiento del habla que han reportado algún éxito hasta el momento han sido probados en vocabularios pequeños. Aunque el tamaño del vocabulario no representa directamente la dificultad del proceso de reconocimiento, existen una serie de problemas que se presentan cuando el tamaño del vocabulario se incrementa, como los que se han mencionada anteriormente. Otra dificultad que podemos mencionar es el aumento en la complejidad de la búsqueda, mientras que en vocabularios pequeños es posible realizar búsquedas menos complejas. En vocabularios pequeños es posible modelar cada palabra, y almacenar sus parámetros. Por otro lado, cuando el tamaño del vocabulario aumenta, se hace imposible el entrenamiento de cada palabra explícitamente, debido a que tanto el entrenamiento como la cantidad de almacenamiento serán en un momento insuficientes. En esta

situación, se tiende a utilizar otras unidades del habla, generalmente, en forma de sub-palabras. El uso de sub-palabras generalmente conduce al degradamiento en la ejecución, debido a que no capturan los efectos de coarticulación de la misma manera que lo hacen los modelos de palabras. Algunas de las unidades de reconocimiento más importantes son: palabras, sílabas, difonos, y fonemas. Los fonemas son las unidades más pequeñas para el reconocimiento del habla, y representan una posibilidad para tratar con los problemas antes mencionados respecto al tamaño de vocabulario.
Recientemente, en RAH desearíamos ser capaces de definir un conjunto de categorías para el reconocimiento de patrones estadísticos; un conjunto obvio debería ser algún tipo de sistema fonético. Una de las cuestiones abiertas en RAH es que el apropiado conjunto de fonos existe para la clasificación. La opción del conjunto de fonos generalmente depende del diccionario usado; un diccionario provee una tabla de búsqueda de palabras para su pronunciación fonética.
El conjunto de fonos del idioma Inglés TIMIT [17] usa un inventario de 61 fonos. Algunos diccionarios, tal como el diccionario CMU, usan representaciones fonéticas con solo 40 clases. El hecho es que teniendo pocas clases frecuentemente se hace más fácil discriminar entre ellas puesto que habrá mas muestras por clase en promedio, pero más clases permitirán distinciones fonéticas finas que pueden incluir algo de información contextual.
Algunos sistemas, como los de IBM, usan categorías completamente manejadas por datos de unidades de sub-palabras.; en el caso de IBM, estas categorías fueron llamadas *fenones,* las cuales fueron modelos derivados de técnicas de agrupamiento estadístico. Típicamente fueron un número algo más grande de fenones que el número de fonos usados en tales sistemas, fue usado un número aproximado de 200. Otros sistemas de RAH usan categorías auto organizadas, dividiendo fonos en tres o tantas sub-unidades; esto es, mientras la definición inicial de la unidad de sub-palabra pueda venir de una definición humana como la establecida en una base de datos tal como TIMIT (i.e, fonos).

3.3.2. Fonos y Fonemas

Las palabras son una unidad natural para modelar en un sistema de RAH, particularmente puesto que hay muchas aplicaciones para la cual palabras aisladas son una adecuada forma de entrada. Incluso en el habla contínua, usar palabras completas

como una unidad linguística fundamental permite el modelado acústico de palabras en un contexto específico de los sonidos usados. Sin embargo, es muy desgastante el entrenamiento de datos para después ignorar lo común entre sonidos de diferentes palabras. Así que, las unidades de sub-palabras son frecuentemente empleadas en grandes vocabularios de RAH. Adicionalmente, para el habla expresada naturalmente, la pronunciación puede variar considerablemente, haciendo muy útil definir unidades lingüísticas pequeñas para el modelado de los RAH. Muchos RAH dividen palabras en unidades llamadas fonos o fonemas.

Fonos

Lingüistas han categorizado muchos de los sonidos de los lenguajes en palabras, dentro de segmentos llamados fonos. Aunque no todos los lingüistas estan de acuerdo en la identificación de estos fonos, los fonéticos en general tienen algún sistema para codificarlos. Los fonos no son necesariamente las unidades más pequeñas para describir los sonidos, pero ellos representan una base de sonidos que pueden ser usados para describir lenguajes.

Fonemas

Puesto que el conjunto de fonos estan diseñados para cubrir todos los lenguajes, dicho conjunto podría ser muy grande. Cada lenguaje escogerá el uso de solo un subconjunto de ellos. El conjunto de categorías únicas que los lenguajes usan son llamados los fonemas del lenguaje. Dos sonidos son considerados como parte de diferentes fonemas si ellos hacen una distinción entre dos palabras; estas palabras son llamadas pares mínimos. Las palabras *pares y mares* son léxicamente distintas; de esto podemos concluir que en Español, /m/ y /p/ son diferentes fonemas. En general, escribimos fonemas entre diagonales, para distinguirlos de fonos.
Habrá algunos casos en los cuales los sonidos en un lenguaje no son utilizados en otros. Por ejemplo, el fono inglés [th] no es usado en español, aun que personas que tienen problemas de dicción frecuentemente lo sustituyen por el fonema [z] o [d]. En otros casos, diferentes fonos serán posibles dado el mismo fonema. En este caso, los fonos son llamados alófonos de los fonemas. En español, por ejemplo, en las palabras *eclipse* y *eclibse*, tenemos a los alófonos [p] y [b] del mismo fonema /p/.

3.3.3. Otras Unidades de Sub-palabras

Fonemas, fonos, y sub-fonos agrupados, y versiones dependiente de contexto de éstas unidades han sido estructuras dominantes para los RAH en la década pasada. Sin embargo, otras unidades han sido consideradas y potencialmente tienen un número de ventajas significantes. Algunos investigadores han sugerido, que el principal elemento en la intelegibilidad del habla no es la clasificación del estado constante de los sonidos del habla (los cuales son infrecuentes en el habla fluida natural), pero en cierto grado la clasificación de las transiciones entre fonemas. Una unidad que encaja bien en esta perspectiva es el difono, el cual es típicamente definido como la extensión de la mitad de una región estado constante a la mitad de la proxima. Los difonos son muy comunes actualmente en síntesis del habla comercial, pero han sido usados solo ocasionalmente para el reconocimiento del habla.

Otra unidad de interés en la comunidad de los RAH ha sido la sílaba (y algunas veces unidades de media sílaba llamadas demisílabas). El inicio de las sílabas parece ser más fácil de detectar acústicamente que los fonos, y todas las sílabas parecen tener restricciones de estructura que pueden potencialmente ser usados en un sistema de RAH. Investigadores del reconocimiento del habla, quienes trabajan con lenguajes distintos al Inglés Americano, han frecuentemente incorporado el uso de sílabas o unidades relacionadas con la sílaba. Ejemplos incluyen trabajos considerables en Japonés, Chino, Alemán y Español.

Pocos proyectos de investigación estan actualmente explorando sílabas para el inglés, aun que el trabajo es controversial; los detractores ven el inglés como un caso pobre de estructura basada en sílabas, puesto que hay un rango de complejas sílabas en inglés, y el tiempo de las sílabas es complicado por patrones de estrés. Sin embargo, una examinación de patrones conversacionales han mostrado que los tipos silábicos extremadamente simples y patrones de tiempo representan la mayor parte del habla conversacional fluida. En general, las sílabas son asociadas con los contornos de energía y pitch que son de 150-250 ms de longitud.

Para propósitos fonológicos, las sílabas son frecuentemente divididas en tres partes: el inicio, el núcleo y la *coda* (cola). El núcleo es el componente mínimo, y, como el nombre lo indica, es el centro de la sílaba. El núcleo esta formado normalmente por vocales. El inicio típicamente es material consonántico que precede al núcleo (en una sílaba que inicia con vocal no hay otro inicio que el mismo núcleo), mientras que la

cola es el material siguiente. En la palabra inglesa spat, el inicio es [sp], el núcleo es [a], y la coda es [th].

3.3.4. Frases

En muchos sistemas de RAH actuales, los modelos acústicos son estructurados como palabras (los cuales son típicamente compuestos por unidades de sub-palabras), y las palabras son agrupadas juntas en una completa oración por usar modelos del lenguaje (típicamente modelos estadísticos simples). Ordinariamente, no hay otra unidad acústica más grande que la palabra. Sin embargo, algunos trabajos inician sobre la incorporación de estructuras de frases en el reconocimiento del habla. Acústicamente, las frases aparentan tener alguna coherencia en términos de contorno de energía y pitch, esencialmente corresponden a una secuencia de sílabas que pueden incluir multiples palabras; hay también comúnmente distinción por las roturas de baja energía entre las frases (i.e. silencios). Los límites de frases también frecuentemente sirven para indicar un cambio de tema o de un límite sintáctico (tal como el inicio de una cláusula). Sin embargo, actualmente la principal aplicación para estructuras de frases es como parte de la sintaxis de un lenguaje natural que es algunas veces utilizado por sistemas de lenguaje hablado para algún dominio limitado, como por ejemplo, el Sistema de Información de Viajes Aéreos.

3.4. Resumen

En este capítulo se ha remarcado la importancia de hacer uso del habla como el medio natural por excelencia para la comunicación hombre-máquina, así mismo, se ha hecho una revisión de las diversas aplicaciones del RAH y el impacto que tienen en diversos ámbitos de la sociedad.
Se revisaron de forma breve, aquellos factores que afectan la calidad de la señal del habla, y cómo repercute en los procesos de segmentación y reconocimiento de la misma. Se han expuesto los distintos enfoques segmentales, y las alternativas del uso de unidades de sub-palabras para minimizar los problemas de reconocimiento, y capacidades tecnológicas para el entrenamiento y almacenamiento. El uso de las unidades mínimas del habla (fonemas) se considera como una alternativa promete-

dora para generalizar y simplificar el reconocimiento de las palabras de un lenguaje. En esta tesis se ha considerado la detección de límites fonéticos para contribuir al desarrollo de los reconocedores del habla basado en unidades en forma de sub-palabras, eliminando restricciones como la independencia de texto y hablante.

Capítulo 4

El Proceso de Segmentación del Habla

Las diversas unidades de habla como frases, palabras, sílabas, difonemas o fonemas pueden ser consideradas como objetos del habla, por lo tanto estas unidades pueden ser clasificadas utilizando técnicas de reconocimiento de formas [18].
A diferencia de otras áreas donde los objetos son entidades individuales, en las unidades de habla articulada no existe una separación obvia entre ellas, y es difícil conocer con precisión donde inicia y termina una entidad. Es claro concluir que un proceso de segmentación es necesario, para dividir una oración de habla en algún tipo de unidades.

4.1. Preproceso y Segmentación

Como se ha expuesto en apartados anteriores, un problema de Reconocimiento de Formas debe obtener algún tipo de representación de los objetos del universo externo. Estos objetos se encuentran presentes en alguna forma física, sobre las que algún tipo de sensor (cámara, micrófono, etc.) obtiene información y las convierte generalmente en señales eléctricas. Estas señales son sometidas a ciertos procesos para ser convertidas en algún tipo de codificación numérica comprimida. Al conjunto de procesos que codifican la señal se les suele llamar como *preproceso*. El proceso previo necesario y crucial en el reconocimiento de habla es llamado *segmentación*, que es aplicado para delimitar los distintos sub-objetos que pueden encontrarse en el objeto que se

está considerando, o simplemente para separar el objeto de interés de algún fondo en el que está inmerso.
En el habla, el universo físico de los objetos a reconocer está constituido por ondas de presión producidas por el aparato fonador humano. Los objetos externos de este universo lo constituyen las diferentes formas acústicas del habla. Según el objetivo y/o la frase de reconocimiento de que se trate, estos objetos pueden ser alguna de las unidades de habla antes mencionadas, así como formantes, energía, amplitud, intensidad, etc. La interpretación de los objetos en el universo físico es concebida por algún sensor o transductor, en el caso del habla el micrófono. Por ello, la parte inicial de todo subsistema de preproceso de la señal vocal estará siempre constituida por: un *micrófono* que convertirá la presión ejercida en el aire a señal eléctrica, y un *amplificador* que tiene como función elevar a niveles manejables la señal proporcionada por el micrófono.

4.1.1. Preproceso

Las funciones del preproceso es codificar la señal en algún tipo de representación manejable, y el filtrado, restauración y/o realce.
La fase de preproceso se encarga de recibir una señal a partir de los sensores, muestrearla en el tiempo o espacio y representarlas de manera comprimida mediante una *codificación*, conocida también como digitalización o cuantificación. Cumpliendo ciertos requisitos en el proceso de muestreo, una señal puede ser reconstruida a partir de su conjunto finito de muestras:
Si una señal es representada por una función $f(t)$ cuya transformada de Fourier es nula fuera del intervalo de las frecuencias [-w,w]. Entonces f(t) puede ser reconstruida a partir de sus muestras si estas son tomadas al menos en intervalos de $1/(2w)$ segundos.
Las técnicas de filtrado, suavizado, restauración o realce suelen usarse para eliminar ruidos, compensar la degradación ocasionada por la compresión y mejorar la calidad de representación del objeto.
Las técnicas de suavizado se utilizan para reducir el ruido presente en el objeto codificado. Este proceso a su vez es llevado a cabo por técnicas de filtrado invariante en el tiempo o espacio. Como ejemplo, podemos citar el suavizado por medias, que consiste en reemplazar el valor de cada punto por el promedio de sus puntos alrededor

localizados en cierto intervalo.
En contraste con el suavizado, podemos hacer que la señal tenga un mayor realce o diferenciación como la que se logra con el filtro pre-énfasis. Esta técnica tiene como objetivo resaltar intervalos de frecuencia frente a otros que contengan menos información relevante.

4.1.2. Segmentación

Este proceso es necesario si en la representación de los objetos externos existen subobjetos a reconocer, los cuales pueden estar estructuralmente relacionados entre sí. El objetivo de la segmentación consiste en descomponer la representación del objeto en elementos más simples, de tal forma que cada parte pueda ser reconocida individualmente.
La segmentación puede presentarse desde la perspectiva de separar un objeto del fondo que lo rodea. En este caso la segmentación consiste en detectar *bordes* o *fronteras*; este es un proceso complicado, que afecta en buena o mala medida a los procesos posteriores.

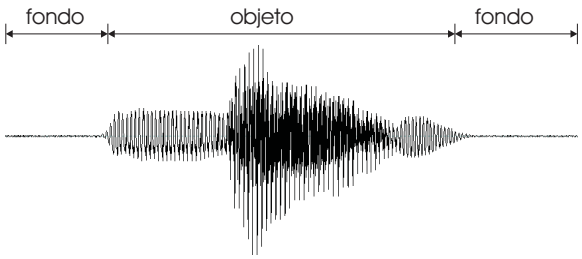

Figura 4.1: El objeto y el fondo que lo rodea

En el caso de la individualización de multiples objetos, la complejidad puede llegar a ser extraordinariamente grande, y en ocasiones, pueden combinarse de forma cooperativa la segmentación y reconocimiento en un proceso único.

4.2. Segmentación y Detección de Límites

En el RAH, la dificultad de esta tarea depende del tipo de objetos acústicos que se traten de aislar. Sería conveniente utilizar objetos relacionados con categorías lingüísticas correspondientes a la forma de percepción oral como frases, palabras,

sílabas o fonemas. Sin embargo, la presencia física de estos objetos suele ser muy compleja y relacionada con la gran cantidad de conocimientos a priori sobre la fonología, léxico y gramática de la lengua. Por esta razón, se suelen introducir otros objetos cuya presencia en la señal de habla es más directa y exigen menos conocimientos de la lengua; algunos de estos objetos son los siguientes:

- *Segmentos de la señal vocal o silencio:* separación de palabras o frases del fondo que las rodea.
- *Segmentos sonoros o sordos:* segmentos de señal que contienen la frecuencia fundamental y los que no.
- *Segmentos vocálicos o consonánticos:* relacionados con picos o valles de la energía dependiente del tiempo de la señal vocal.
- *Microfonemas:* elementos acústicos obtenidos directamente del sub-muestreo temporal de alguna representación paramétrica del habla.
- *Difonemas:* intervalo comprendido entre los puntos centrales de dos segmentos estacionarios de señal. Idealmente se asimilan a parejas de semifonemas(mitades de fonemas).
- *Segmentos pseudosilábicos:* intervalo comprendido entre los puntos centrales de dos vocales consecutivas.
- *Pseudofonemas:* elementos asimilables más o menos directamente con los fonemas de una lengua.

No solamente la dificultad de segmentación para cada uno de estos objetos es diferente, sino que la calidad o consistencia de los resultados obtenibles es también variable. Así, por ejemplo, una definición de segmentos sordos/sonoros es especificable con cierta presición mediante características físicas conocidas de la señal vocal. Y los resultados de una segmentación ideal pueden ser bastante consistentes, mientras que una definición al nivel acústico de difonema es más imprecisa y que la segmentación realizada sea perfecta con respecto a la definición, los resultados suelen ser bastante inconsistentes. En la Figura 4.2 se muestra el grado de inconsistencia en relación con la dificultad de segmentación en las unidades de habla mencionadas.

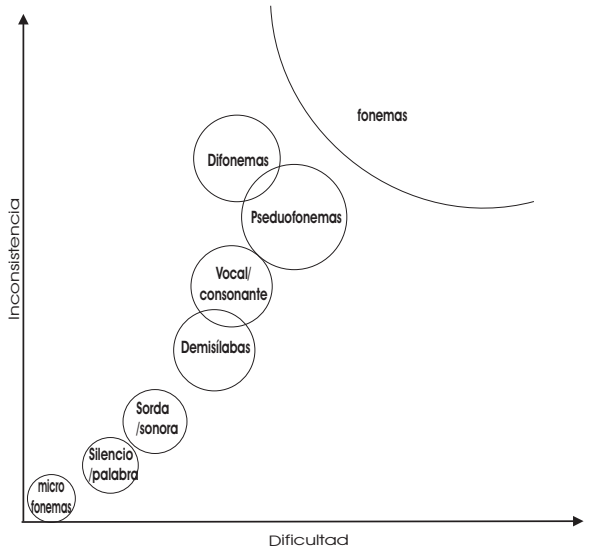

Figura 4.2: Las unidades de habla con sus respectivos grados de inconsistencia y dificultad de segmentación

Aunque la segmentación fonética ha sido un objetivo primordial de las investigaciones del Reconocimiento Automático de Habla, en la actualidad sigue siendo un problema abierto para el cual no existen soluciones satisfactorias. La mayoría de las técnicas utilizadas han sido de naturaleza heurística, en las que se han ido empleando diversos tipos de parámetros *ad hoc* para ir resolviendo problemas. Aunque algunos de estos métodos han llegado a dar resultados en casos específicos, la naturaleza de las técnicas empleadas ha imposibilitado su formalización y con ello la extensión o mejora de los resultados.

Dentro de las técnicas heurísticas cabe destacar aquellas que utilizan el grado de estabilidad de la señal como parámetro principal de la segmentación. Al asumir la señal de habla como un proceso localmente estacionario, los resultados de cualquier tipo de parametrización dependiente del tiempo sólo son representativos de aquellos segmentos en los que la señal es suficientemente estacionaria en el interior de las ventanas de análisis correspondientes. Si se establecen como categorías fonéticas o pseudofonemas las distintas clases posibles de segmentos estacionarios de la señal de habla, el grado de estabilidad proporciona un método relativamente formal de segmentación cuyos resultados pueden ser satisfactorios. Otra técnica relacionada con lo que se acaba de exponer, recurre a la aplicación de un modelo markoviano basado en la definición estadística de cierta similitud o disimilitud entre los vectores de parámetros extraídos en ventanas de análisis consecutivas, así como en las estadísticas de duración de los

segmentos a obtener.
Otro grupo de técnicas son aquellas en las que la segmentación no se lleva a cabo como un proceso separado del reconocimiento, sino que se realiza en cooperación con y como subproducto de este. La mayoría de estas técnicas recurren a una segmentación microfonética previa, seguida de un reconocimiento de los microfonemas con respecto a cierto conjunto de categorías acústicas. Una técnica de este grupo recurre a la definición de una "función de similitud"de los microfonemas extraídos de la señal vocal con categorías o prototipos microfonéticos. Dicha función se suele definir como el máximo de las funciones discriminantes de las categorías correspondientes. La segmentación pseudofonética se basa en localizar los máximos y mínimos locales de esta función: los máximos representan los centros de los pseudofonemas y los mínimos las fronteras entre pseudofonemas contiguos.
Las técnicas anteriores asumen que las características de las señales de habla presentan un grado de uniformidad en las fronteras de los fonemas. En la práctica, los fonemas reales del habla raramente presentan esta uniformidad asumida; en particular, la mayoría de las consonantes se manifiestan como sucesión de segmentos de características distintas, algunas de las cuales son esencialmente de naturaleza no estacionaria. Estas consideraciones han llevado a muchos autores a renunciar a cualquier tipo de segmentación fonética y a intentar el acceso a unidades lingüísticamente superiores como difonemas y sílabas directamente a partir de la cadena microfonética. Dicho acceso se puede fundar en métodos de reconocimiento de formas basado en la teoría de la decisión, sintácticos, o en métodos de inteligencia artificial.

4.2.1. Detección de Bordes

Aunque los fonemas han sido los objetos acústicos en los que se han centrado mayormente las investigaciones sobre segmentación, hay otros objetos cuyo intereses prácticos es considerable. Se trata de las palabras o frases aisladas. Al problema de aislar las palabras del silencio(fondo) que los rodea, se les conoce como "detección de fronteras o bordes", y su interés práctico radica en la existencia de métodos eficientes de reconocimiento global de palabras aisladas pertenecientes a cierto diccionario de prototipos.
La detección de fronteras suele abordarse en dos etapas diferenciales: detección burda y detección fina. La detección burda conviene llevarla a cabo en combinación con el

proceso de adquisición de la señal de habla.
Un algoritmo simple y eficaz para la detección burda de bordes, válido si el ruido no es excesivo, utiliza umbrales de tiempo y energía (o amplitud) de la señal: la superación de estos umbrales indica la presencia de palabra. El umbral de tiempo resulta necesario debido a que no es suficiente afirmar que hay palabra cuando hay energía, por que podría considerar algún ruido espurio como una palabra. Una palabra presenta siempre un mínimo de energía durante un mínimo de tiempo y viceversa, la ausencia de energía no implica fin de la palabra, puesto que se terminaría al encontrar el silencio que forma parte de un fonema plosivo (/p/,/t/,/k/, etc.). Una zona de silencio real tiene un mínimo de duración. Por otra parte, es necesario considerar también como parte de la palabra el tramo de señal inmediatamente inferior o posterior al momento del paso por el umbral de energía, y en caso contrario podrían perderse los principios y/o finales de palabras constituidos por señal de energía inferior al umbral, como la contenida en algunos sonidos como los nasales y fricativos. Para realizar la detección burda en tiempo real, se requiere un *buffer* que permita trabajar a un *micrófono abierto*, es decir, con tiempos de silencio iniciales indefinidamente largos. En este *buffer* se va almacenando cíclicamente la señal, con lo que en cada momento t se dispone de la señal adquirida desde el instante t-Tb hasta t, dependiendo Tb del tamaño de la memoria física. De esta manera, en el momento que se detecta el principio eficaz de la palabra, se dispondrá del segmento de señal inmediatamente anterior en el que está contenido el principio real de la palabra. El tamaño del buffer dependerá del tamaño del segmento inicial anterior al principio eficaz que sea necesario conservar. Para palabras o frases en español se puede utilizar para este segmento una duración de 200 msec aproximadamente.
La detección burda asegura que la palabra o frase pronunciada quedará contenida en su totalidad dentro de los límites detectados, lo que resulta suficiente en muchos casos. Sin embargo, un procedimiento más detallado(detección fina) conducirá no solo a la mejora de resultados en una etapa posterior de reconocimiento, si no un indudable ahorro de memoria requerida para el almacenamiento de la señal y tiempo a emplear en los tratamientos posteriores. La detección fina se suele basar en los parámetros extraídos para tratamientos posteriores, aunque también se puede llevar a cabo mediante parámetros ex profeso para esta tarea, típicamente amplitud y densidades de cruces en cero. A partir del resultado proporcionado por la detección burda, el algoritmo debe buscar de atrás hacia adelante, a partir del momento en que se

cruzó inicialmente el umbral de amplitud, el punto donde los parámetros extraídos indican la existencia de silencio. Para el fin de palabra, también de atrás hacia adelante a partir del final burdo, se procede de forma análoga hasta encontrar el momento que deja de haber silencio. La existencia o no de silencio determina a partir de los parámetros utilizados, y de la decisión de su existencia depende de la la naturaleza de éstos. Un método simple consiste en utilizar alguna *medida de similitud* disponible para los vectores de parámetros empleados, y establecer un patrón de silencio con su correspondiente umbral de similitud. Con estas premisas, la existencia o no de silencio corresponderá a valores superiores o inferiores al umbral de similitud.

4.3. Enfoques de la Segmentación Automática del Habla

En esta sección se hace una breve revisión de los dos grandes enfoques existentes para la segmentación automática del habla. Uno de estos enfoques consiste en algoritmos que realizan la segmentación con el apoyo de información adicional a la onda del habla, como el conocimiento de la secuencia de fonemas. Estos sistemas requieren datos de entrenamiento segmentados manual o automáticamente. La segmentación se lleva a cabo mediante un método denominado alineación forzada, utiliza la información previamente extraída de la señal a segmentar, y efectúa varias iteraciones, comparando los puntos de segmentación obtenidos con los puntos de segmentación reales hasta alcanzar los resultados esperados.
El otro enfoque toma solo la señal de habla como entrada, sin tener un conocimiento previo de la secuencia de fonemas contenidas en la señal en cuestión; estos algoritmos localizan las instancias de tiempo donde se encuentran los límites, basándose en la detección de puntos donde hay un alto grado de variación en la forma de onda de la señal. Como se menciona anteriormente (Capítulo 1), los algoritmos desarrollados se encuentran en la categoría de los no supervisados, que toman únicamente como entrada la señal de habla, y no requieren de ningún tipo de entrenamiento.

4.4. Resumen

Se ha explicado la importancia del proceso de segmentación como factor crucial en el Reconocimiento Automático del Habla. Para facilitar el proceso de segmentación, se utilizan técnicas de preproceso, las cuales pueden ser suavizados o filtrados aplicados a la señal del habla. Por otro lado, se utilizan esquemas de codificación que emulan el sistema auditivo, realizando el proceso de muestreo, filtrado y cuantificación, dando como resultado vectores de valores espectrales de la señal en proceso. La señal del habla puede ser segmentada en distintos objetos que van desde la obtención de frases, palabras, hasta fonemas. La segmentación fonética, sigue siendo un problema abierto , debido a que no existen soluciones satisfactorias. Actualmente, la mayoría de los enfoques de segmentación requieren de procesamiento complejo (Modelos Ocultos de Markov, Redes Neuronales) donde frecuentemente es requerido un entrenamiento. Métodos sobresalientes en segmentación fonética, como [8] [19], no requieren entrenamiento alguno, aunque emplean una fase de post-procesamiento para la selección de puntos de segmentación, donde una matriz que contiene los brincos detectados en cada banda de frecuencia, es recorrida y analizada en ventanas, obteniendo los puntos de segmentación definitivos de la señal del habla procesada.
En esta tésis el problema de la segmentación se abordará tratando la disimilaridad existente entre frames, puesto que las variaciones espectrales son prominentes en los límites fonéticos y en donde una alta disimilaridad refleja dichas variaciones. Como se ha citado previamente, se utiliza el enfoque independiente de texto para tratar el problema de la segmentación, sin requerir algún tipo de entrenamiento.

Capítulo 5

Trabajos Relacionados

Este capítulo tiene como objetivo hacer una revisión de trabajos enfocados al proceso de segmentación, así como las restricciones con las que fueron diseñados y probados. Cada trabajo es descrito de manera breve, remarcando sus características importantes.

5.1. Segmentación Automática de Fonemas por Aplicación de Reconocimiento Vocal

En este trabajo [4] Mayora propone un algoritmo para segmentación automática de sílabas. La segmentación se hace sobre una sola palabra, sub-dividiéndolo en fonemas a partir de la sílaba encontrada. Este algoritmo se basa en la individualización por el "decremento de sonoridad", que verifica un punto de menos acentuación en el pasaje entre dos sílabas de una palabra. Lo remarcable de este trabajo es que la segmentación está basada en características en dominio del tiempo como lo es la intensidad y la frecuencia fundamental F0.

5.1.1. Síntesis del lenguaje basado en sílabas

El uso de la sílaba como base para el reconocimiento de la palabra, garantiza una mejor coarticulación entre la unidad de la palabra respecto a la técnica basada en fonemas y una mejor calidad del lenguaje sintetizado; esto es debido al hecho que el principal parámetro prosódico (F0) está estrechamente asociado a ellos. Por

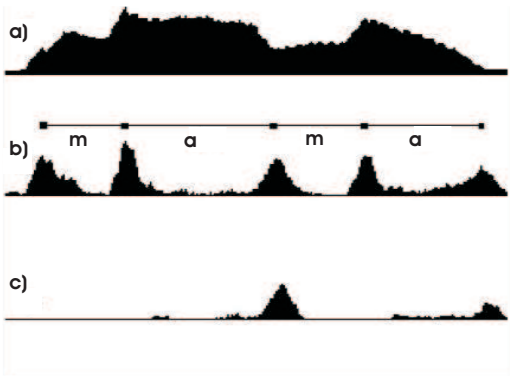

Figura 5.1: (a)Intensidad acústica, b)Variación acústica, c)Decremento de sonoridad

otro lado, este enfoque debe afrontar el problema del manejo de una base de datos de segmentación significativamente amplio y tratar de modo eficaz los efectos de coarticulación entre sílabas adyacentes.

5.1.2. Decremento de Sonoridad

El enfoque basado en sílabas resulta ser una opción para la segmentación, especialmente si el análisis es procesado en el dominio del tiempo. La característica a la que se recurre para dividir la palabra en sílabas es el "decremento de sonoridad", que se presenta una importante atenuación de la intensidad acústica al pasar de una sílaba a otra.

El uso de la intensidad en el dominio del tiempo facilitará la detección de la posición exacta de las fronteras o límites.

5.1.3. Algoritmo

División en Sílabas

El algoritmo de segmentación en sílabas se desarrolla utilizando como punto de partida los instantes de inicio y fin de la señal vocal. La idea base es delimitar la señal usando una función que calcula los cruces en cero sobre la elocución.

División en fonemas

Para ampliar la posibilidad de uso del algoritmo en ambiente de software, donde el reconocimiento del lenguaje se apoya en fonemas, se desarrolla una nueva función que consiste en dividir cada sílaba en fonemas. Para apoyar la segmentación en fonemas se crea una base de datos que contiene , por cada una de las diez palabras sobre la cual se efectúan las pruebas, la siguiente información:

- Número de sílabas presentes en la palabra
- Número de fonemas por cada sílaba
- Porcentaje de longitud de cada sílaba dentro de la palabra
- Porcentaje de longitud de cada fonema dentro de cada sílaba

5.1.4. Observaciones

Este algoritmo realiza una segmentación de palabras en sílabas y posteriormente los segmentos silábicos en fonemas, haciendo uso de características temporales como lo es la intensidad y el cruce en cero. Aunque es un enfoque novedoso, este algoritmo es dependiente de vocabulario por que solo es probado con señales de habla correspondientes a números en italiano, es dependiente de texto por que necesita de información adicional a la señal de habla como el número de sílabas y fonemas en cada palabra.

5.2. Segmentación Fonética de Habla Contínua Usando un Perceptrón Multicapa (MLP)

En esta sección se expone una breve revisión de este método propuesto por Suh y Lee en [1], el cual consiste en un perceptrón multicapa (Multi Layer Perceptron) para la segmentación de fonemas. Se adopta el problema de detección de límites como un tipo de clasificación de patrones. Este método se respalda del hecho que los enfoques basados en MLP han mostrado una notable capacidad de discriminación no lineal. Este enfoque es muy interesante, ya que consiste en representar y clasificar con

características bien definidas los patrones de los que son límites fonéticos y los que no lo son.

5.2.1. Arquitectura MLP para la Segmentación en Fonemas

Esta arquitectura consiste básicamente de tres partes: Un preprocesador, el MLP, y el post-procesador.
El pre-procesador extrae las características, y tiene a su vez dos estados. El primer estado corresponde a la extracción de las características de cada frame de habla, y el segundo estado lo que hace es re-extraerlas como la diferencia de frames adyacentes. Las características son aquellas derivadas de la Transformada Rápida de Fourier (FFT). Se extraen características del orden 44 de cada frame, y el tamaño en tiempo de cada frame es de 25 msec con un desplazamiento de 10 msec.
Puesto que las variaciones de las señales son más prominentes en los límites fonéticos, estas variaciones son tomadas como buenos indicadores para la segmentación. Se utiliza el concepto de *inter-frame* para denotar a las diferencias de características entre frames adyacentes. Las características inter-frame son normalizadas entre -1 y +1 para ser usadas en el MLP.
El segmentador de fonemas basado en MLP, tiene una capa oculta y una capa de entrada. Las 176 características de cuatro inter-frames consecutivos son usados como datos de entrada por haber demostrado un buen desempeño en los experimentos. La capa de salida tiene un solo nodo que decide si el actual frame es límite fonético o no. En la capa oculta una función sigmoide es usada como una función de activación.
En el post-proceso, las posiciones de los límites fonéticos son decididos usando el valor de salida del MLP. Cuando la salida del MLP es mayor a cierto umbral, se indica que las características del inter-frame representan un límite fonético.

5.2.2. Algoritmo de Aprendizaje

Se hace uso de una modificación del método de propagación hacia atrás que converge mucho más rápido. La salida tiene un valor de +1 en un límite fonético y -1 o valor similar en cualquier otra posición. Características de cuatro frames consecutivos son aplicadas al MLP, y son desplazadas por un frame,para aprender todos los casos de los patrones de habla de entrada. La tasa de aprendizaje es asignada a 0.0005,

y los valores de peso inicial son aleatoriamente generados en el rango de -5.0E-7 a 5.0E-7 para todos los casos.

5.2.3. Experimentos

Se hace uso de una base de datos Koreana de habla leída, expresada por un hablante femenino y etiquetada manualmente por expertos fonéticos. Se utilizaron 156 señales de habla, muestreadas a 16 kHz. Este método presenta un tasa de correcta detección de fronteras del 87 % con una tasa de inserción del 3.4 %. Aunque presenta un desempeño eficiente, este método fué probado sobre un solo hablante de sexo femenino, y requiere de un gran número de características (144) para la detección de fronteras, así como requiere de mucho tiempo de entrenamiento.

5.3. Método para la Segmentación de Fonemas Independiente de Texto

Existe una gran variedad de métodos enfocados a la segmentación del habla en fonemas, sin embargo, la mayoría de los reportados en el estado del arte presentan alguna restricción, como ya se ha descrito en éste capítulo. El método propuesto en [8] que se explica a continuación evita algún tipo de restricción como dependencia de hablante, de texto y de vocabulario dentro de las más importantes. El preproceso está basado en un análisis perceptual de banda crítica, y el fenómeno de sobre segmentación es tratado.

5.3.1. Preprocesamiento

Aunque el habla es producida por ondas no estacionarias de sonido, es posible hacer un análisis de características *quasi-estacionarias* por medio de pequeñas ventanas subsecuentes, donde el tamaño comúnmente no es mayor a 20 msec, consiguiendo una adecuada descripción de la onda del habla.
La señal del habla es descompuesta en una secuencia de *frames* de 20 msec (320 muestras) con sobrelapamiento de 10 msec (160 muestras). Las muestras son pesadas en ventanas de Hamming para evitar distorsiones espectrales. Cada frame es entonces pasado por un análisis basado en percepción, que incluye resolución de banda crítica,

pre-énfasis de igual sonoridad y compresión de intensidad del espectro de Fourier. Se extraen entre 15 y 19 características por cada frame, donde cada característica cuantifica la energía espectral contenida en un cierto intervalo de frecuencia.
El preproceso obtiene vectores de características por secuencia de tiempo por cada frame. La colección de características es denotada por:$\{x_i[n]\}_{i=1}^{k}$, donde n representa a cada frame.

5.3.2. Detección de Límites Fonéticos

Dados los vectores de características por cada frame, el algoritmo procede a detectar cambios rápidos y significativos de las características acústicas sobre la secuencia de tiempo, conocidos como brincos *(jumps)* en este trabajo.
Se computa sobre cada secuencia de tiempo $\{x_i[n]\}$ una función *brinco* definida como sigue:

$$\jmath_i^a[n] = \left| \sum_{m=n-a}^{n-1} \frac{x_i[m]}{a} - \sum_{m=n+1}^{n+a} \frac{x_i[m]}{a} \right| \tag{5.3.1}$$

donde n denota al frame que se computa por la función brinco, a representa el número de frames previos y posteriores en relación al frame n. En otras palabras, la función brinco obtiene una diferencia absoluta entre las medias de $\{x_i[n]\}$ calculada sobre frames adyacentes al frame n, sobre la secuencia de tiempo. Si $\{x_i[n]\}$ presenta un cambio significativo, $\{\jmath_i^a[n]\}$ presentará un "pico"para el mismo frame. Un pico es definido como válido si excede un cierto umbral b.
Este algoritmo tiene un proceso fundamental para combinar, en una única indicación de límite fonético, los eventos brinco que son detectados alrededor del mismo frame n en las k distintas secuencias de tiempo; el proceso es llamado "ajuste"(fitting) y fue introducido para colocar el límite de segmentación en la mitad de un grupo de brincos *quasi-simultáneos*, dentro de un intervalo genérico de frames. La serie de picos finales son almacenados en un vector llamado acc[m], donde m vá desde 1 hasta M, siendo M el número de picos detectados en la señal. Un parámetro c es utilizado para ajustar el ancho de los sub-intervalos, en los cuales el algoritmo busca por el punto medio de los brincos *quasi-simultáneos*.

5.3.3. Experimentos

Para probar este algoritmo se usaron 480 oraciones representativas de 48 hablantes(24 masculinos y 24 femeninos) de la base de datos TIMIT, muestreadas a 16 kHz. Este conjunto de sentencias involucra a 17930 fronteras.
El algoritmo es capaz de detectar el 73.56 % de las transiciones fonéticas con aproximadamente 0 % en la tasa de inserciones.

5.4. Integración de Segmentación Independiente del Lenguaje y Modelado Basado en Fonemas Dependientes del Lenguaje

El algoritmo propuesto en [19] es una modificación del método para la segmentación de fonemas independiente de texto, expuesto en la sección anterior, logrando una ligera mejora en el desempeño.

5.4.1. Modificación al Algoritmo Original

En la modificación al algoritmo original, en lugar de considerar todos los picos, un límite fonético es colocado en cada frame solo si el pico es mayor que un umbral identificado como *thresh1*. Este umbral es definido como sigue:

$$Thresh1 = \frac{max(acc[m]) + min(acc[m])}{2} \tag{5.4.1}$$

El pico es detectado en el punto m donde acc[m] es mayor que el valor del umbral Thresh1. Otra gran modificación es la introducción de otro paso de post segmentación, en el cual el segundo nivel de segmentación es llevado a cabo. Este segundo nivel de segmentación es aplicado a aquellas secuencias de señal segmentadas que no empalmen con los fonemas correspondientes al Tamil. Además, algún segmento de señal que tenga una duración mayor que la máxima duración permitida por los fonemas del Tamil(20 msec), es considerado como patron de sílabas y para una futura segmentación.
Un nuevo umbral(thresh2) fue adoptado el cual estuvo fijo a la mitad del umbral thresh1. En el segundo nivel de segmentación, picos de segmentación que se encuen-

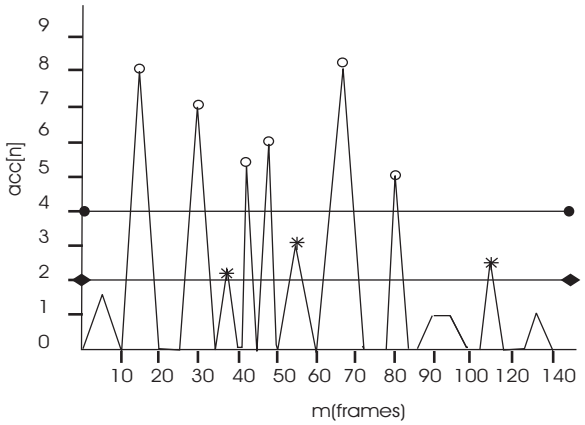

Figura 5.2: Secuencia de puntos de segmentación, después del Nivel 1 de segmentación denotadas por O y puntos incluidos por el Nivel 2 de segmentación denotada por *

tran dentro del rango de thresh1 y thresh2 fueron detectados y los segmentos silábicos fueron segmentados. En la figura 5.2 se muestran picos detectados usando los umbrales thresh1 y thresh2.

5.4.2. Experimentos

Una colección de 200 señales de habla muestreadas a 16 kHz fueron usadas, representativas de 20 hablantes (10 masculinos y 10 femeninos).
El algoritmo fue probado con distintos valores para los parámetros *a,b y c*. Los valores *a=2, b=0.3 y c=7* fueron con los que mejores resultados se obtuvieron. El algoritmo fue capaz de detectar correctamente 75.8 % de transiciones fonémicas. El desempeño del algoritmo con dos niveles de segmentación presenta un incremento del 2.3 % respecto al algoritmo de un solo nivel de segmentación.

5.5. Resumen

En este capítulo se ha hecho una revisión del estado del arte en cuanto a segmentación del habla se refiere. La segmentación del habla en sub-palabras no es una tarea sencilla, ya que se debe tratar con los problemas vistos en el capítulo anterior. Se han propuesto métodos que incluyen restricciones importantes como la dependencia de texto, de hablante y vocabulario.

Se analizaron los métodos [8] y [19] que han sido probados sin las restricciones antes mencionadas y considerando el factor de sobre-segmentación, reportando resultados más realistas. Estos métodos han sido probados en el idioma Inglés y Tamil respectivamente. En esta tésis, los métodos desarrollados se probarán bajo condiciones similares, utilizando distintas representaciones del habla en el dominio del tiempo y de frecuencia, y haciendo uso de membresias difusas para obtener mayor detalle de estas representaciones. Los métodos se probarán tanto en el idioma Inglés como el Español, puesto que en este último hasta el momento no se ha reportado algún método de segmentación fonética con independencia de texto, del hablante y vocabulario sobre habla contínua.

Capítulo 6

Características Acústicas del Habla

En este capítulo se describe el análisis de características acústicas tanto en el dominio del tiempo como en el dominio de frecuencia. Este análisis consiste en detectar patrones que puedan presentar las características acústicas en los puntos donde existen límites fonéticos.

6.1. Características en el Dominio del Tiempo

El objetivo de analizar características en el dominio del tiempo es conocer la utilidad de algunas de ellas en el proceso de segmentación.
Las características comúnmente utilizadas en el dominio del tiempo son la energía, intensidad, frecuencia fundamental y cruces en cero, por mencionar algunas. Ciertamente, las características acústicas en el dominio del tiempo cuantifican los fenómenos físicos producidos al hablar, como puede ser por ejemplo la presión ejercida en el aire (amplitud) por las ondas sonoras. Estos fenómenos físicos se cuantifican teniendo valores escalares en un instante de tiempo t.
Se hizo el análisis sobre métricas de la intensidad, amplitud y energía; aunque estas medidas son relacionadas, no presentan exactamente los mismos comportamientos en los límites fonéticos.
Este proceso es puramente observatorio, donde se ubican los puntos de segmentación (establecidos por expertos fonéticos) sobre la gráfica de cada característica.

6.1.1. Frecuencia fundamental

La frecuencia fundamental también conocida como *pitch* (F0) es la frecuencia de vibración de un cuerpo, siendo en la producción del habla el cuerpo vibratorio las cuerdas vocales. El número de oscilaciones de la presión del aire cada segundo, determinan el *pitch* del sonido, lo cual físicamente correlacionado es frecuencia.
Aunque la F0 no fue utilizada en el proceso de análisis acústico en el dominio del tiempo, si lo fue en el proceso de extracción de la intensidad de la señal de habla.

6.1.2. Amplitud

La amplitud es la cantidad de presión que ejercen las ondas sonoras sobre las partículas de algún medio elástico, y es cuantificada en unidades absolutas correspondientes a la distancia entre la posición de reposo y el punto máximo de desplazamiento. La amplitud puede ser cuantificada desde el punto de vista físico en Pascales (Pa), o desde el punto de vista perceptivo en decibeles (dB). En nuestro análisis se utiliza la amplitud desde el punto de vista físico.
Muchos de los límites fonéticos, generalmente, se encuentran por cambios significa-

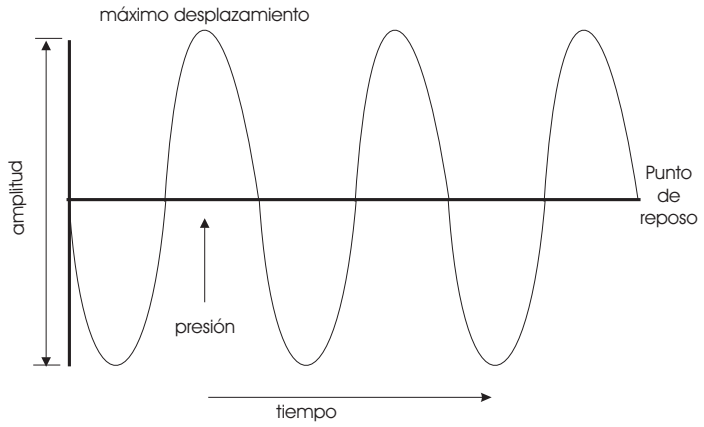

Figura 6.1: Amplitud del movimiento

tivos sobre valores de la amplitud. Sin embargo, existe un gran número de cambios poco significativos, y solo en algunos de ellos se presentan límites fonéticos. Es evidente que hacer una adecuada detección de límites sobre cambios poco significativos no es una tarea fácil, sobre todo si se quieren evitar las inserciones. En la figura 6.2 se muestra una señal de habla expresada por un hablante femenino con sus respectivos

límites fonéticos (•) mapeados sobre los valores de la amplitud.

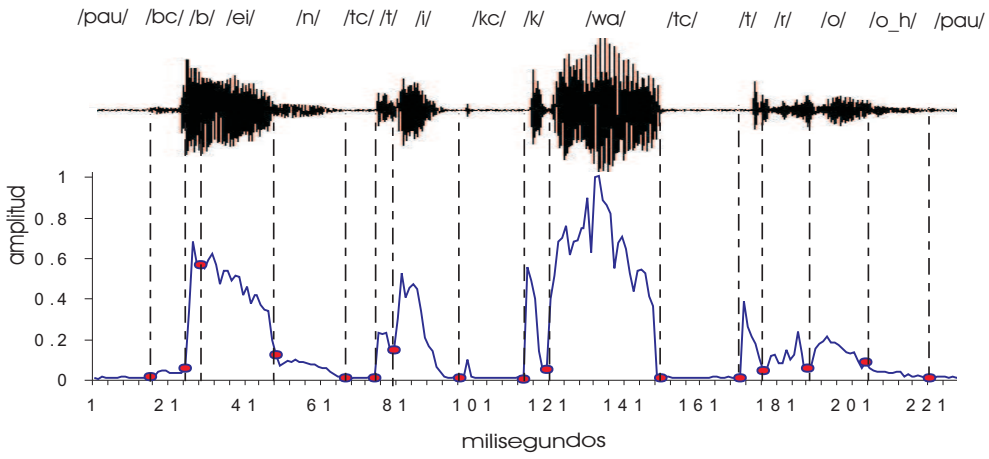

Figura 6.2: Límites fonéticos sobre valores de la amplitud

Se puede observar que los cambios significativos (inflexiones pronunciadas en la gráfica) no puede ser considerados un patrón determinante en la segmentación, puesto que muchos de estos cambios pueden presentarse por factores como cambios de entonación, inadecuada pronunciación, *clicks* (ocasionados por abrir los labios al inicio de una oración), respiros del habla, o ruidos de fondo.

6.1.3. Energía

La señal de habla puede representarse de distintas formas, una de ellas es la energía. Si $x(t)$ es la amplitud del sonido dada en Pascales (Pa), la energía es definida en general como:

$$E = \sum_{n=-\alpha}^{\alpha} x^2(t) \tag{6.1.1}$$

La energía es obtenida por la sumatoria de los cuadrados de la amplitud de la señal, y es cuantificada en Pa^2 s. La energía puede ser utilizada para distinguir conjuntos vocálicos de consonánticos, así como para la detección de presencia o ausencia de voz sobre señales de alta calidad por medio de umbrales.

Puesto que la energía es obtenida a partir de la amplitud, se observan resultados simi-

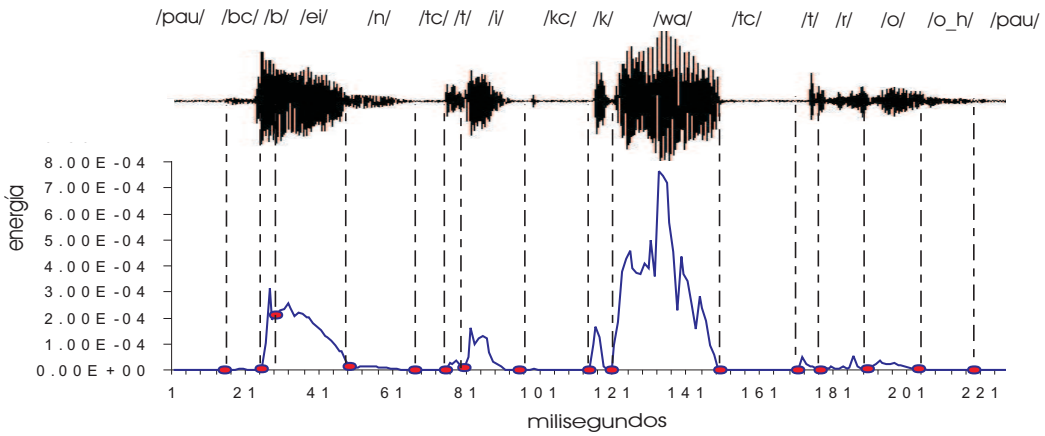

Figura 6.3: Límites fonéticos sobre valores de la energía

lares entre ambas características en relación con el mapeo de límites fonéticos, como se puede observar en la Figura 6.3. Sin embargo, los valores de amplitud presentan cambios más notorios que los presentados en los valores de la energía.

6.1.4. Intensidad

La intensidad, al igual que la energía son métricas obtenidas a partir de la amplitud. Cuando el sonido es producido, la energía de la fuente es irradiada sobre el area donde se propagan las ondas sonoras. La intensidad se define como el flujo de energía por unidad de area por unidad de tiempo.

Tanto la amplitud como la energía de una señal de habla contienen un gran número de pequeños cambios que rodean a los límites fonéticos, de tal forma que se podrían reflejar en el desempeño de la segmentación como falsos límites (inserciones). Para reducir el inconveniente de las inserciones, una posible solución es aplicar un suavizado de medias sobre los valores de estas características del dominio del tiempo. Por otro lado, aplicando el suavizado podría ocasionar la pérdida de algunos límites fonéticos. La intensidad tiene la característica de no presentar tantas variaciones en el tiempo por tener baja sensibilidad a los cambios de amplitud, en comparación con la sensibi-

lidad de la energía. En la Figura 6.4 se muestran los límites fonéticos presentes en la intensidad, de acuerdo a ésta, la intensidad es una métrica que presenta cambios gra-

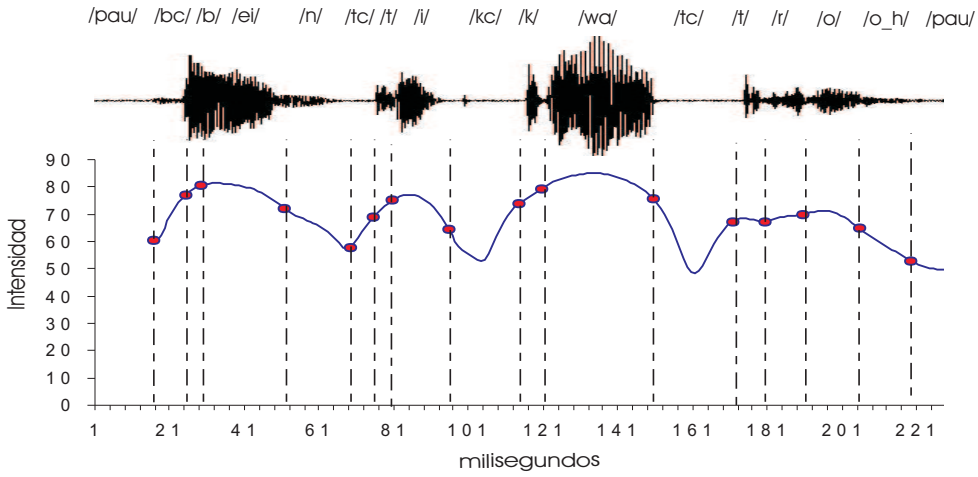

Figura 6.4: Límites fonéticos sobre valores de la intensidad

duales ligeramente perceptibles, que aparentemente, complica la detección de límites fonéticos (desde el punto de vista gráfico). Sin embargo, presenta en la segmentación una efectividad similar, e incluso mejor que la amplitud y energía en el dominio del tiempo. En el **Capítulo 7** se muestra el desempeño en el proceso de segmentación por cada una de las características aquí analizadas.

6.2. Características en el Dominio de Frecuencia

El sonido puede tener distintas codificaciones tanto en el dominio del tiempo como en el dominio de frecuencias. La señal de habla contiene concentraciones de energía en diferentes frecuencias generadas por el tracto vocal, por lo tanto, se pueden extraer cuantificaciones de estas energías, obteniendo una representación más elaborada del sonido. Estas representaciones del sonido vienen dadas en vectores de características por unidad de tiempo, y se utilizan técnicas basadas en bancos de filtros para obtenerlas. Algunas codificaciones exitosas en el dominio de frecuencias son MFCC, PCBF, espectros de Mel, basados en el sistema de percepción auditiva humano.

Las codificaciones antes mencionadas, se basan en investigaciones de fisicoacústica, donde es conocido que el sistema auditivo humano lleva a cabo un análisis de frecuencia del sonido de entrada por un banco de filtros. Así que en un intento por querer emular este sistema humano, varios extractores de características han sido estudiados. La extracción de características por banco de filtros es ilustrada en la Figura

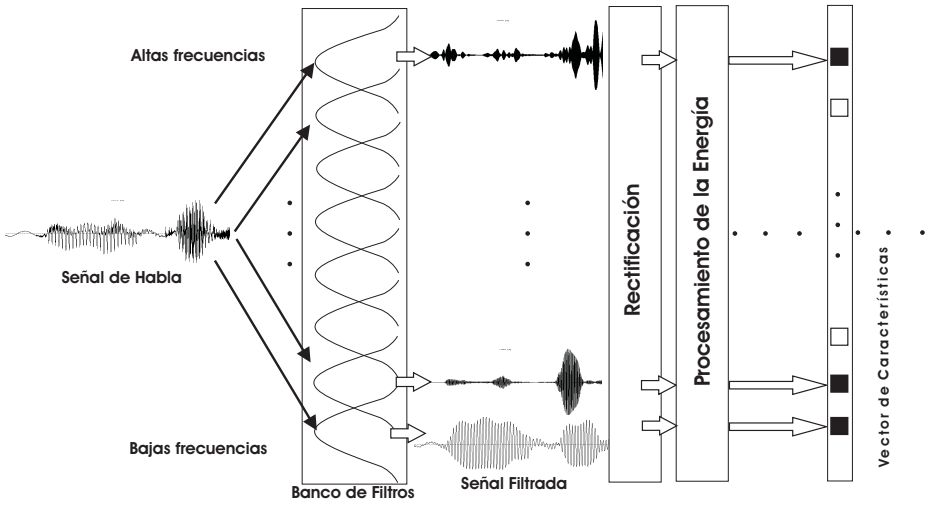

Figura 6.5: Diagrama de un extractor de características con bancos de filtros

6.5. Una señal de habla *S(n)* es pasada a través de un banco de filtros pasabanda. Generalmente, los filtros individuales se encuentran traslapados en frecuencia, y están espaciados de acuerdo a una escala de frecuencia no uniforme, tal como la escala Bark o Mel basadas en experimentos perceptuales.
En el proceso de segmentación, con vectores de características tenemos más información para poder detectar límites fonéticos, así como un mayor procesamiento para su extracción y manejo.

6.2.1. Espectros Mel

Los vectores de espectros de Mel se obtienen como resultado de pasar una señal a través de un banco de filtros. Cada espectro en el vector es el resultado de filtrar el espectro de entrada a través de un filtro individual, siendo el vector de longitud igual al número de filtros.
Los filtros triangulares son centrados sobre los ejes de frecuencias dispuestas en la escala no lineal de Mel; esta escala fue inicialmente sugerida por Stevens y Volkman

en 1940 [20]. El banco de filtros emula las bandas críticas perceptuales, acentuando las bajas frecuencias. Los bordes de los filtros coinciden con los ejes de frecuencias adyacentes. Un modelo común para la relación de frecuencias en Mel y escalas lineales es como sigue:

$$frec.Mel = 2595 * log_{10}(1 + \frac{frec.lineal}{700}) \tag{6.2.1}$$

Puesto que se obtienen características vectoriales (cuantificaciones de energía por frecuencia) por cada frame, es un tanto complicado analizar las transiciones entre fonemas. Cada frame con sus respectivos vectores puede ser enfocado como un objeto, donde los espectros en cada frecuencia representarán los atributos del mismo. Las transiciones pueden ser detectadas aplicando medidas de distancia entre dichos frames(objetos). En la Figura 6.7 se observa la descomposición de una señal de habla

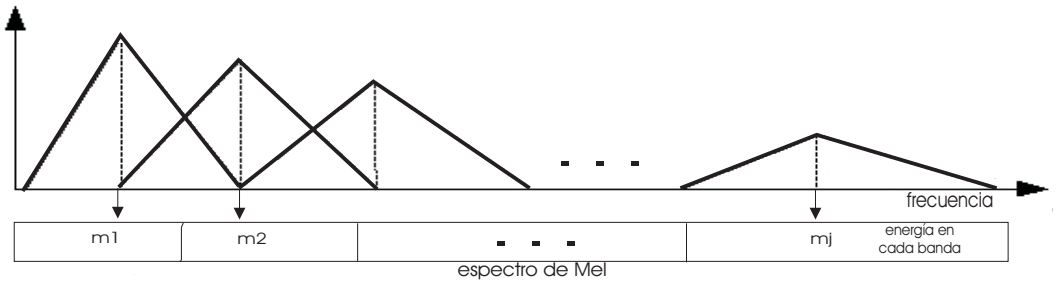

Figura 6.6: Banco de fltros Mel

en sus distintas cuantificaciones de energía. Se observa que se obtiene por cada filtro la intensidad correspondiente a cierto intervalo de frecuencia. Así como la intensidad, energía y amplitud son características correlacionadas, la intensidad se encuentra de alguna manera correlacionada con los espectros Mel, puesto que al promediar estos espectros en un instante t es posible obtener la intensidad en ese mismo instante t.

6.2.2. Coeficientes Cepstrales de Frecuencia Mel

Es una de las representaciones del sonido más utilizadas en reconocimiento automático del habla e ingeniería de audio. El proceso inicia aplicando una ventana de Hamming para evitar distorsiones espectrales, descomponiendo la señal contínua

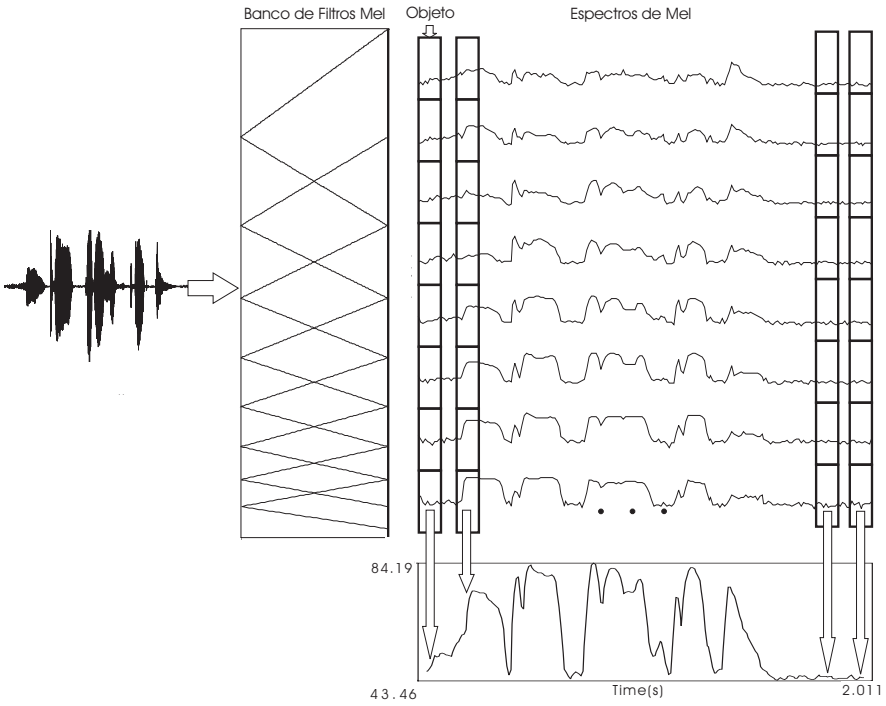

Figura 6.7: Descomposición de una señal de habla por Filtros de Mel

en tramas, donde típicamente estas tramas suelen ser de 20 o 30 ms durante la cual la señal se asume ser estacionaria. Las tramas traslapadas son comunes. Las características MFCC son derivadas de la magnitud espectral de la Transformada Rápida de Fourier aplicando un banco de filtros Mel. El logaritmo de la energía de cada filtro es calculado y acumulado antes de ser aplicada una Transformada de Coseno Discreta para producir los vectores de características de MFCC.

El análisis *cepstral* denota un inusual tratamiento en el dominio de la frecuencia como si fuera en el dominio del tiempo [21]. El cepstrum es una medida de periodicidad de la frecuencia de respuesta. La unidad de medida en el dominio cepstral es en segundos, pero indican las variaciones espectrales de frecuencias. Los MFCC son una elegante forma compacta para la representación de formas espectrales. En la parte superior de la Figura 6.8 se observa la onda de habla, de la cual se obtuvieron medidas físicas tales como la amplitud, intensidad y energía que se presentan en el dominio del tiempo para llevar a cabo la segmentación fonética del habla. Estas medidas físicas del habla permiten obtener información de la señal presente en el medio de propagación, sin tener que utilizar esquemas codificados como los antes mencionados.

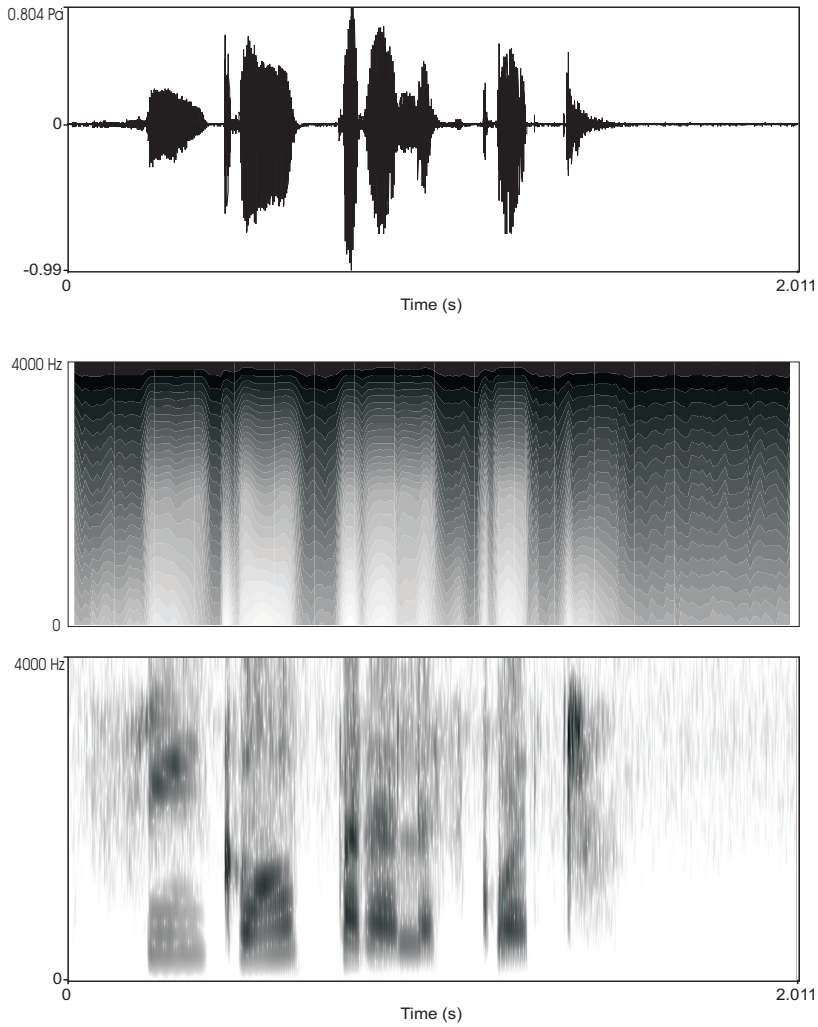

Figura 6.8: Forma de onda, MFCC, y Espectrograma

En la parte inferior de la Figura 6.8, se observa el espectrograma, el cual representa las variaciones de frecuencia y la intensidad de la onda de habla con respecto al tiempo. La representación espectrográfica tiene la desventaja de contener una gran cantidad de información que podría ser difícilmente tratable. Para minimizar este problema se han propuesto esquemas de codificación espectrales y cepstrales como los antes mencionados. Estos esquemas de codificación del habla tienen el objetivo de extraer información relevante de la señal y evitar la redundancia de la misma (compresión de audio).

En el centro de la Figura 6.8 se observa la representación en MFFC's correspondien-

te a la misma onda de habla. Se puede ver que la intensidad en grises es invertida respecto al espectrograma, puesto que precisamente, el proceso de los MFCC's consiste en obtener la transformada inversa de Fourier de la magnitud logarítmica de los espectros , de ahí que reciben el nombre de cepstrums.

6.3. Resumen

En este capítulo se analizaron distintas representaciones del habla en el dominio del tiempo y frecuencia. Se observó que características que requieren procesamiento simple para su extracción y manejo como la amplitud, energía o intensidad en conjunto con membresias difusas ofrecen un detalle similar a los vectores de características en el dominio de frecuencia. Los límites fonéticos se presentan en las inflexiones de la curva representada por los valores en el tiempo de estas características. La mejor característica del dominio del tiempo es la intensidad, debido a su baja sensibilidad a los cambios de energía en la señal del habla.
Los espectrogramas representan la señal del habla en dominios de frecuencia, conteniendo información redundante y de difícil tratamiento. Esquemas de codificación en forma de vectores de espectros o *cepstrums* son extraídos a partir de la señal de habla en un procesamiento de muestreo y filtrado de la misma, reduciendo la información a ser tratada. Los vectores de características con más éxito en el RAH fueron analizados; los coeficientes cepstrales en las frecuencias de Mel (MFCC) representan de forma compacta la señal del habla en el dominio de frecuencia y son uno de los esquemas con mejores resultados en el proceso de reconocimiento del habla. Los espectros de Mel cuantifican la intensidad en cada intervalo de frecuencia, a partir de las cuales es posible reconstruir la intensidad en el dominio del tiempo. Los algoritmos descritos en el próximo capítulo hacen uso de las características analizada, donde se discuten los resultados.

Capítulo 7

Algoritmos de Segmentación

7.1. Introducción

En este capítulo los algoritmos de segmentación propuestos serán descritos a detalle. El primer grupo de algoritmos de segmentación está basado en medidas físicas sobre la señal de habla tal como intensidad, amplitud y energía. Por otro lado, se implementó un algoritmo de segmentación que hace uso de características espectrales, probado en distintas codificaciones del sonido como MFCC(Mel Frequency Cepstral Coefficients) y espectros Mel. Los métodos propuestos son totalmente independientes de texto puesto que, en el proceso de segmentación, no hacen uso del etiquetado u otro tipo de información adicional asociada a la señal de habla, la única entrada utilizada es la expresión de habla continua a ser segmentada. La segmentación no es realizada en tiempo real. El proceso se realiza con independencia de hablante, ya que en los experimentos se han considerado una gran variedad de hablantes.
El objetivo de los algoritmos es encontrar las instancias de tiempo donde exista una transición entre fonemas, con una tolerancia de $\pm$ 20 msec. La estrategia se enfoca en la detección de cambios espectrales o de medidas físicas del habla. Con el objetivo de obtener un mayor detalle de transiciones vagas entre fonemas, se hace uso de membresias difusas utilizadas en una función de disimilaridad. Teóricamente, se presume que los cambios espectrales notorios ocurren entre las transiciones de un fonema a otro; sin embargo, esto no es totalmente cierto, puesto que las oraciones de habla tratadas son expresadas con naturalidad donde algunas transiciones no están claramente definidas, especialmente al finalizar las oraciones.

7.2. Bases de Datos

Esta sección tiene por objeto dar una breve descripción de las bases de datos utilizadas en los experimentos.

7.2.1. Base de Datos DIMEx100

El corpus fué grabado por 100 hablantes, donde cada uno grabó 50 frases individuales en adición a 10 frases que fueron grabadas por los 100 hablantes, teniendo en total 6000 frases. El corpus fué grabado en un estudio de sonido en CCADET, UNAM, con un formato de muestreo mono a 16 bits, y una tasa de muestreo de 44.1 *kHz*.
Los hablantes fueron seleccionados entre los 16 a 36 años de edad, con estudios superiores a secundaria de la ciudad de México. Un grupo aleatorio de hablantes en la UNAM(investigadores, estudiantes, maestros y trabajadores) fué seleccionado, con una edad promedio de 23.82 años; el 87 % son no graduados y el 82 % nació y vivió en la ciudad de México. También, 18 personas de otros lugares residiendo en la ciudad de México participaron en las grabaciones. En total el corpus contiene el 49 % de oraciones expresadas por el sexo femenino, y un 51 % de oraciones expresadas por el sexo masculino. Aunque el español de México tiene muchos dialectos (del norte de la región, costa del Golfo y Península de Yucatán por mencionar algunos), el dialecto de la ciudad de México representa la variedad hablada por la mayoría de la población en el país.

7.2.2. Base de Datos TIMIT

La base de datos de habla DARPA TIMIT fué diseñada para proporcionar datos fonético-acústicos del habla para el desarrollo y evaluación de sistemas de reconocimiento automático del habla. Consiste de expresiones de 630 hablantes que representan la mayoría de los dialectos del Inglés americano. Las regiones dialécticas incluidas son: New England, Northern, North Midland, South Midland, Southern, New York City, Western y Army Brat.

Cada hablante grabó 10 oraciones del siguiente tipo: oración de calibración dialéctica (2 por hablante), oración variante contextual aleatoria (3 por hablante) y oración fonéticamente compacta (5 por hablante).
Cada directorio de sentencias contiene 3 archivos incluyendo: archivo de forma de onda con una tasa de muestreo de 16 *kHz*, un archivo con transcripción fonética, y un archivo con transcripción ortográfica.

7.2.3. Datos Experimentales

Se utilizaron en la fase de pruebas 240 señales de habla con un total 11195 límites fonéticos, correspondientes a 30 hablantes (15 masculinos y 15 femeninos) extraídas del corpus DIMEx100 [22] con oraciones en español. Con la intención de tener una prueba más completa, los mismos algoritmos son también probados con el corpus TIMIT con oraciones en inglés, del cual se han extraído 544 señales expresadas por 68 hablantes(34 masculinos y 34 femeninos).

7.3. Evaluación de Desempeño

Los algoritmos desarrollados y explicados en esta sección, son evaluados en términos de las tasas de correcta detección y sobre-segmentación. Un límite fonético es definido como "correctamente detectado"si tiene una distancia de $\pm$ 20 *ms* al "verdadero límite"[8] [23]. Los límites fonéticos localizados por los algoritmos se comparan contra los límites fonéticos de las bases de datos (que son definidos por expertos fonéticos). La medida de desempeño se definen como sigue:

$$P_c = 100 \cdot (\frac{S_c}{S_t}) \tag{7.3.1}$$

Donde P_c es el porcentaje de segmentación correcta, y S_c es el número de puntos de segmentación detectados correctamente, y S_t es el número de los puntos reales de segmentación.
Usando solamente la expresión 7.3.1 no podría medirse la calidad del proceso de segmentación. Puesto que el algoritmo podría tener como salida un gran número de límites detectados, incrementando, ficticiamente, la probabilidad de detectar los verdaderos puntos de segmentación, con la desventaja de haber introducido un número

no esperado de puntos erróneos de segmentación [23]. El fenómeno previamente citado es conocido como *sobre-segmentación*, y puede ser cuantificado como sigue:

$$D = 100 \cdot (\frac{S_d}{S_t} - 1) \tag{7.3.2}$$

Donde D es el porcentaje de sobre-segmentación y S_d es el número de puntos totales detectados. Esta medida nos permite cuantificar el desempeño de manera estricta, puesto que a medida que se vaya alcanzando un alto porcentaje de correcta segmentación, existirá un margen de error cade vez más reducido, expresado en términos de sobre-segmentación.

7.4. El Filtro Pre-énfasis

El filtro pre-énfasis es utilizado en la fase de preproceso de todos los algoritmos implementados.

La función del filtro pre-énfasis es obtener un sonido de alta cuesta espectral, haciendo más notorios los cambios espectrales que se presentan en las transiciones de fonemas.

Una frecuencia F debe proporcionarse, a partir de ella la cuesta espectral será incrementada en 6 dB/octava. El factor pre énfasis α es procesado como:

$$\alpha = exp(-2\pi F \Delta t) \tag{7.4.1}$$

Donde Δt es el periodo de muestreo del sonido. El sonido resultante y es obtenido con:

$$y = x_i - \alpha x_{i-1} \tag{7.4.2}$$

Donde cada muestra x del sonido es cambiada, a partir de la última muestra.

En la parte inferior de la Figura 7.1 se puede observar que la intensidad de la señal de habla preenfatizada define de una mejor manera los cambios existentes a lo largo de la señal de habla, resultando en un contorno de intensidad mejor definido, mientras que la intensidad sin preénfasis no define muchos de los cambios contenidos en la señal de habla, lo cual es relevante en el proceso de segmentación.

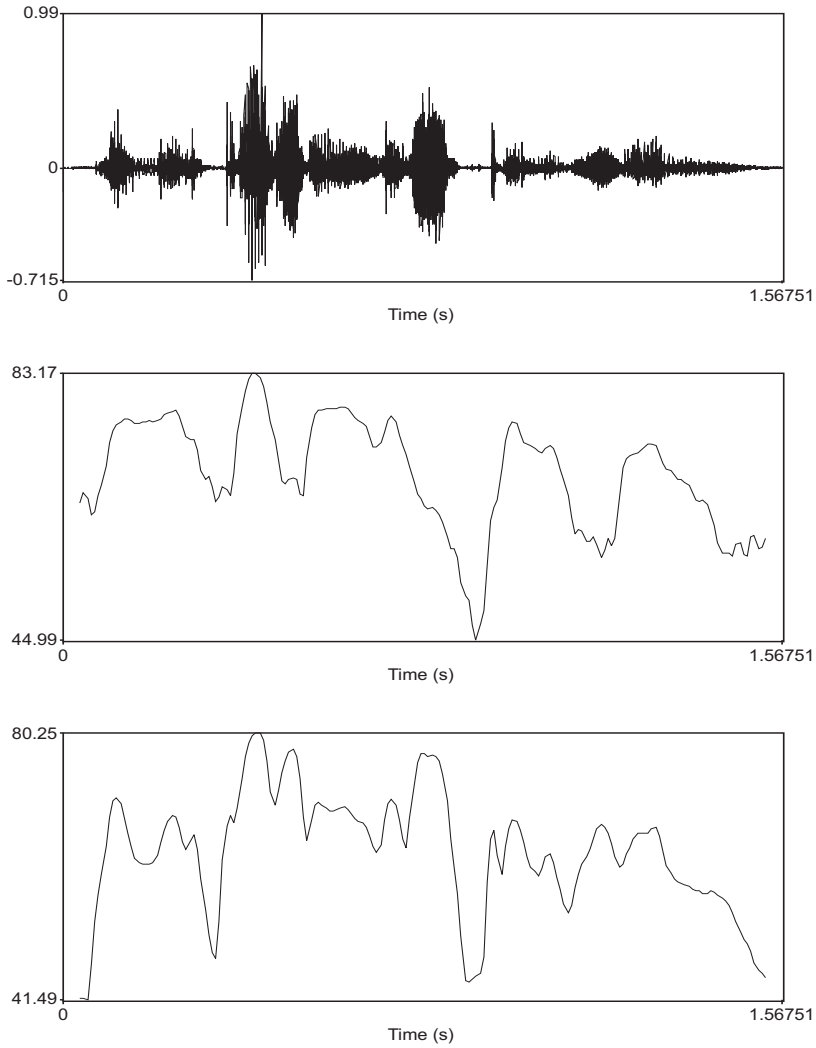

Figura 7.1: Intensidad con y sin pre-énfasis

7.5. Medidas Difusas Utilizadas en Algoritmos de Segmentación

Dado un conjunto difuso A, se definen ciertas magnitudes medibles del conjunto, que se conocen como **medidas difusas**. Una de las principales magnitudes es la **difusión**. Si llamamos C al conjunto discreto de valores k en los que $\mu_A(k) > 0$, la difusión indica la distancia de A al conjunto discreto C. Para efectos de segmentación, la magnitud de difusión mide el grado de pertenencia que tiene un frame de la onda del habla a los conjuntos Alto, Medio y Bajo de medidas físicas o características

espectrales del habla, que se explican a detalle en los algoritmos desarrollados.
Por otro lado, la distancia entre dos objetos representados por sus membresias difusas a conjuntos definidos se pueden definir utilizando diversas medidas. Se definen las medidas más frecuentes como sigue:

7.5.1. Distancia Euclidiana

Mide la distancia que existe entre un punto $X(X_1, ..., X_n)$ y un punto $Y(Y_1, ..., Y_n)$, que desde el punto de vista de reconocimiento de patrones se consideran objetos X y Y con sus respectivos atributos. En el proceso de segmentación los objetos comparados corresponderán a frames (f) adyacentes (o separados por un frame de distancia) de la señal de habla y los atributos serán sus respectivas membresias a los conjuntos difusos definidos.

$$\delta(f,f\text{-}2) = (\sum (\mu_A(f) - \mu_A(f\text{-}2))^2)^{1/2} \tag{7.5.1}$$

Donde $A = \{Alto, Medio, Bajo\}$

7.5.2. Distancia Manhattan

La función de distancia de Manhattan, obtiene la distancia que debería ser recorrida de un elemento a otro si una ruta en escuadra es seguida. La distancia entre dos elementos es igual a la suma de las diferencias de sus componentes.

$$\delta(f,f\text{-}2) = \sum \left|\mu_A(f) - \mu_A(f\text{-}2))\right| \tag{7.5.2}$$

7.5.3. Distancia de Correlación Pearson

La correlación de Pearson mide la similaridad en forma entre dos perfiles. La formula para obtener la distancia de correlación de Pearson es:

$$\delta = 1 - r \tag{7.5.3}$$

Donde:

$$r = \frac{Z(u) \cdot Z(v)}{n} \tag{7.5.4}$$

es el producto punto de los valores-Z de vectores u y v. Los valores-Z de u son construidos al substraer de u la media y su desviación estándar.

La fórmula 7.5.3 nos proporciona la similaridad entre dos objetos, sin embargo, para detectar límites fonéticos se ha utilizado un enfoque basado en la disimilaridad, puesto que en las transiciones entre fonemas, generalmente, las diferencias son prominentes. En esta tésis se utiliza una modificación de la distancia de correlación Pearson para obtener la disimilaridad entre frames de la señal de habla:

$$\psi(f) = \frac{\sum \mu_A(f)}{n} \tag{7.5.5}$$

donde n es el número de conjuntos difusos y ψ es la media aritmética de las membresias del frame bajo análisis (f y *f-2*).

$$\sigma(f) = \Sigma(\mu_A(f) - \psi(f))^2 \tag{7.5.6}$$

donde $\sigma(f)$ es la varianza del frame procesado con respecto a sus memebresías.

$$Z(f) = \frac{\psi(f)}{\sigma(f)} \tag{7.5.7}$$

$$\delta(f,f\text{-}2) = \frac{Z(f) \cdot Z(f\text{-}2)}{n} \tag{7.5.8}$$

Basandose en experimentos, se hicieron ligeras modificaciones de la función original 7.5.3, donde se aplica la varianza en lugar de la desviación estandar, y no se aplica el complemento para obtener la distancia, en su lugar se utiliza 7.5.8 para determinar la disimilaridad.

7.5.4. Distancia de Chebyshev

La distancia de Chebyshev es calculada según la expresión:

$$\delta = max_{i=1..d}\Big\{|X_i - Y_i|\Big\} \tag{7.5.9}$$

esto es, el valor absoluto de la máxima diferencia entre atributos individuales.
La distancia de Chebyshev es una métrica muy atractiva computacionalemente frente a la distancia Euclidiana.

7.6. Algoritmos de Segmentación con Características en el Dominio del Tiempo

En esta sección se detallan los algoritmos de segmentación desarrollados con las representaciones del habla en el dominio del tiempo antes expuestas. Una de las ventajas de estas representaciones es el procesamiento escalar por unidad de tiempo, y que se cuantifican sin ningún tipo de procesamiento complejo como el requerido en las representaciones del dominio de frecuencias.

7.6.1. Algoritmo básico de segmentación con detección de bordes

Este algoritmo hace uso de la amplitud, intensidad y la energía para llevar a cabo la segmentación. El algoritmo toma como entrada la señal de habla para pasarla por el filtro pre-énfasis a partir de una frecuencia F.
En primera instancia, el algoritmo debe dividir la señal en silencio/sonido, en el cual la energía es comparada contra un umbral establecido, donde el silencio es detectado si el segmento de la señal bajo análisis se encuentra bajo el umbral definido, en caso contrario el segmento de la señal será algún tipo de sonido [18],[9]. Esta primera fase del algoritmo determina los límites entre pausas y sonidos, procesando en la siguiente fase aquellos segmentos en los cuales hay presencia de sonido.
La segunda fase consiste en detectar las diferencias significativas en la amplitud, intensidad o energía de la señal de habla, para esto se emplea 7.6.1 que fue también utilizada en [8] en el dominio de frecuencias.

$$\delta = \left| \sum_{m=n-a}^{n-1} \frac{x_i[m]}{a} - \sum_{m=n+1}^{n+a} \frac{x_i[m]}{a} \right| \tag{7.6.1}$$

Donde a es el número de frames considerados antes y después del frame bajo análisis, y n hace referencia al frame bajo análisis.
El objetivo de esta función es obtener la diferencia absoluta de las medias de los a valores de la amplitud previos y posteriores respecto al frame bajo análisis. De esta manera se obtienen máximos locales (picos) que representan una máxima diferencia entre los valores de la amplitud. Si denotamos como v_t a los valores representados en la Figura 7.3, los candidatos a límites fonéticos (presentes como máximos locales)

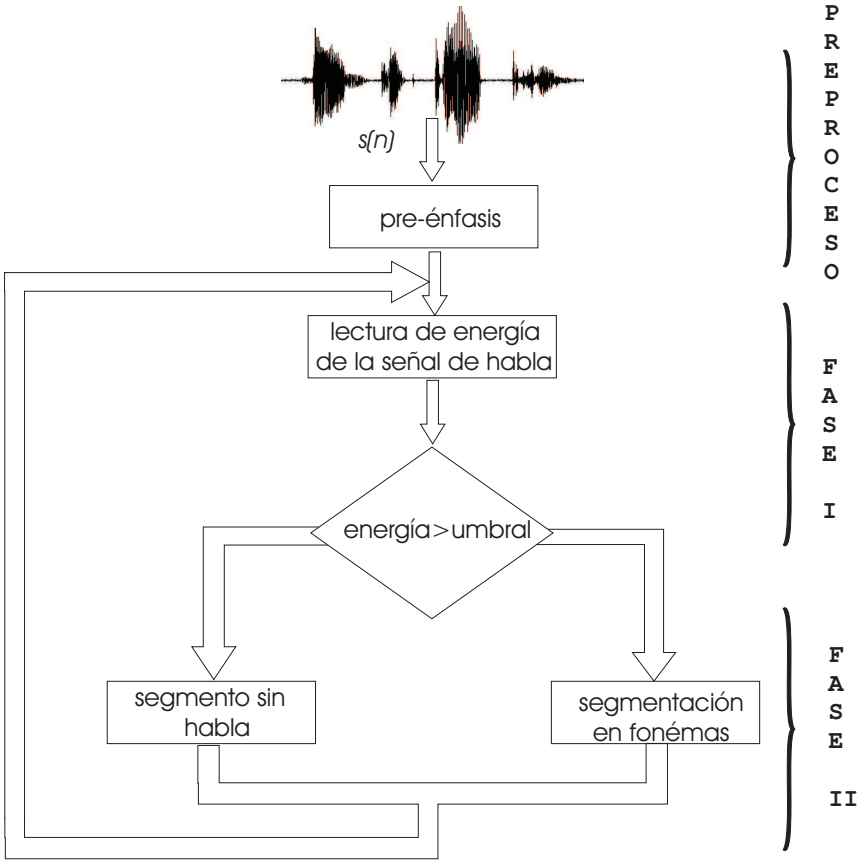

Figura 7.2: Diagrama del algoritmo de segmentación con detección de bordes

deberán cumplir con las siguientes condiciones:

- $v_t > v_{t-1} \ldots v_{t-b}$ y $v_t > v_{t+1} \ldots v_{t+b}$
- $\left|v_t - v_{t-b}\right| > 0{,}019$ y $\left|v_t - v_{t+b}\right| > 0{,}019$

Este algoritmo utiliza frames de 4 *ms* en la fase de detección de bordes(silencios/sonidos). En la segunda fase correspondiente al procesamiento del segmento de sonido, el tamaño de los frames dependerá de la característica usada para segmentar.

Experimentos

Se realizaron experimentos previos para detectar bordes sobre las bases de datos TIMIT y DIMEX, obteniendo una energía de 135 E^{-9} Pa^2 en ambas para llevar a

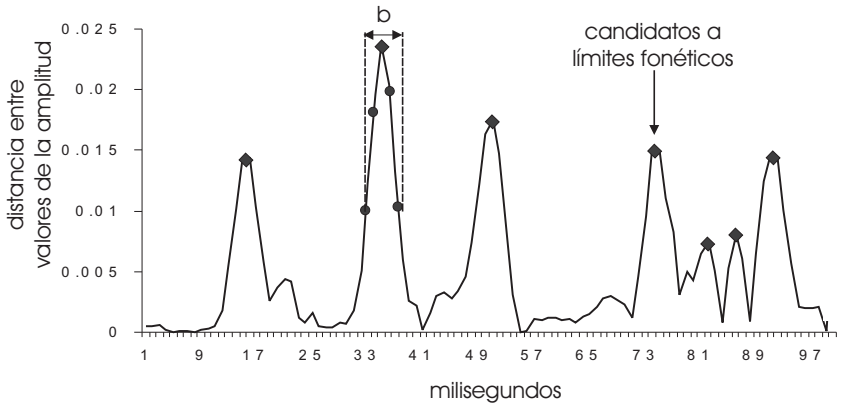

Figura 7.3: Diferencias absolutas de las medias de la amplitud, y sus respectivos parámetros

cabo esta detección. Este nivel de energía será utilizado como umbral en los siguientes experimentos para detectar bordes(silencio/sonidos). En la Tabla 7.1 se muestran los parámetros utilizados con cada característica.

Característica	**a**	**b**	**c**	**frames en ms.**
Intensidad	2	4	2.9	5
Amplitud	2	4	0.035	10
Energía	2	4	0	10

Tabla 7.1: Parámetros empleados en cada una de las características sobre el corpus DIMEx100

El desempeño en términos de tasa de detecciones correctas y tasa de inserciones se muestra en la Tabla 7.2.
El uso de la intensidad y la amplitud en este algoritmo generan una tasa de correcta segmentación aproximada al 70 % siendo la intensidad ligeramente más eficiente que la amplitud. Por otro lado, el uso de la energía proporciona un desempeño inferior, puesto que la tasa de inserciones es muy alta.

Característica	**Sd**	**Sc**	**% correcta detección**	**% inserciones**
Intensidad	11225	7843	70.21	0.49
Amplitud	11426	7898	70.70	2.26
Energía	12448	8071	72.25	11.44

Tabla 7.2: Desempeño del algoritmo usando distintas características sobre el corpus DIMEx100

Las pruebas sobre el corpus TIMIT se realizaron con los mismos parámetros

mostrados en la Tabla 7.1, y el desempeño del algoritmo con cada una de las características físicas se observa en la Tabla 7.3. En las pruebas de ambos corpus se han observado comportamientos similares, donde resalta ligeramente el desempeño del algoritmo que utiliza la intensidad para llevar a cabo la segmentación. Este algoritmo

Característica	Sd	Sc	% correcta detección	% inserciones
Intensidad	20547	15775	76.30	-0.48
Amplitud	21448	14852	71.93	3.87
Energía	23667	14519	70.32	14.62

Tabla 7.3: Desempeño del algoritmo usando distintas características sobre el corpus TIMIT

tiene la desventaja de ser altamente vulnerable a la calidad de la señal. El desempeño de la segmentación es directamente dependiente de la correcta separación de habla y silencio (detección de bordes), una señal con ruido puede afectar considerablemente esta separación y en consecuencia la correcta segmentación en fonemas. Sin embargo, como se menciona arriba, el mismo umbral fue utilizado en ambos corpus con tasas de muestreo, idiomas y etiquetados diferentes. A pesar de ser un enfoque simple, proporciona resultados aceptables, tomando en consideración que hace uso reducido de información con 2 ó 4 valores cada 20 ms cuando hay presencia de voz.

7.6.2. Algoritmo con Medidas de Distancias Difusas de la Intensidad

Se desarrolló un nuevo algoritmo difuso para la segmentación de habla continua con independencia de texto. El enfoque usa sólo la intensidad, teniendo una reducción importante de información y reglas simples en el proceso para la detección de límites fonéticos. En la fase de preproceso se aplica el filtro pre énfasis, y en el proceso de segmentación se usa una estrategia basada en una métrica de distancia con membresias difusas normalizadas. El método alcanza 77.54 % de correcta segmentación con una exactitud de 20 ms, y una tasa de sobre segmentación de 0 %.
El uso de esquemas de codificación ha sido citado como una restriccion en [23], puesto que la señal de habla no es procesada como se presenta en el medio ambiente, sino que sufre un proceso de transformación. Algún tipo de esquema de codificación de habla como los previamente citados fueron evitados en este algoritmo, en su lugar, una cantidad reducida de información como la obtenida de la intensidad es aprovechada.

Para llevar a cabo la segmentación de fonemas, se hace uso de una métrica de distancia difusa entre frames y un conjunto de simples reglas para detectar distancias significativas de intensidad, las cuales fueron tratadas como cambios fonéticos en habla contínua. Tomamos ventaja de membresias difusas de la intensidad para obtener detalles en casos donde las diferencias entre frames son vagas.

Segmentación Fonética

La señal pre enfatizada a partir de los 50 Hz es usada para obtener la intensidad. Usando un pitch mínimo de 93 Hz, y frames de 3 ms sin traslapamiento fueron usados (por haber obtenido mejor resultado en experimentos).
Para cada señal, la intensidad maxima es obtenida, y como intensidad mínima intensidad se utiliza un valor constante de 25 dB para establecer el espacio difuso. Se calcula el promedio entre la máxima y mínima intensidad, para obtener la mitad del espacio difuso. Se asignaron tres funciones triangulares difusas, representando los términos linguísticos correspondientes a baja, media y alta intensidad fueron asignadas.
Las membresias de los conjuntos difusos fueron obtenidas, y luego ellas fueron normalizadas como sigue:

$$\lambda = max(M) \tag{7.6.2}$$

Donde M es el grupo de membresias difusas obtenidas de los frames comparados. La máxima membresia difusa, denotada como λ es obtenida. Entonces $\mu_i = \mu_i/\lambda \forall \mu_i \epsilon\ M$ es aplicado. Desde que nuestra estrategia es basada en diferencia de frames para detectar límites fonéticos, las membresias difusas normalizadas son usadas en (6.1.4)(distancia euclidiana), y denotaremos a los valores obtenidos como v.

$$\delta(f_t, f_{t-2}) = \sqrt{\Sigma(\mu_A(f_t) - \mu_A(f_{t-2}))^2} \tag{7.6.3}$$

El enfoque es simple, puesto que detecta máximos locales sobre los valores v, los cuales indican diferencias significativas entre frames comparados, y por lo tanto, la presencia de un límite fonético. Las reglas usadas para detectar máximos locales son las siguientes:

1. $v_t > v_{t-1}$ y $v_t > v_{t+1}$
2. $v_t >$ 46,8 dB

3. $v_t > \phi$

La condición 1) es usada, por que un valor v en el tiempo t es tratado como máximo local si es mayor que los valores previo y siguiente v sobre la secuencia de tiempo. La condición 2) es usada, por que los valores v_t con baja intensidad, generalmente no son representativos de límites fonéticos (aunque en pocos casos se presentan). La condición 3) es usada para seleccionar máximos locales, los cuales son representativos de cambios significativos.
Las últimas dos condiciones son usadas para evitar inserciones, aunque algunos límites fonéticos son descartados por ellas, por otro lado muchos puntos detectados inválidos son rechazados, resultando en un desempeño competitivo. Un aspecto a mejorar de este algoritmo, es evitar la condición de tomar sólo aquellos segmentos de duración mayor a 0.021 *ms*, aunque esta disminuye la sobre-segmentación no distingue aquellos segmentos válidos menores a 0.021 *ms*. Las condiciones impuestas en el algoritmo sin duda alguna ayudan a reducir la sobre segmentación, sin embargo un pequeño número de segmentos válidos son también sacrificados.

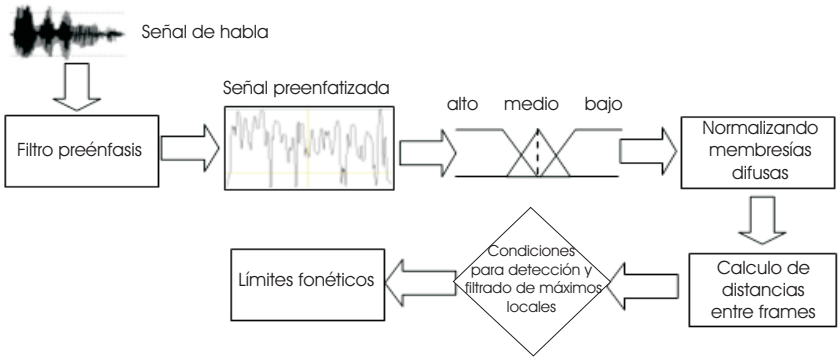

Figura 7.4: Diagrama a bloques del algoritmo con membresias difusas de la intensidad

Experimentos

En estos experimentos hemos incluido habla contínua expresada naturalmente, con independencia de hablante y de texto. El algoritmo fue probado con 544 señales de habla muestreadas a 16 kHz de la base de datos American English DARPA-TIMIT, correspondientes a 68 hablantes (34 hombres y 34 mujeres) de todas las regiones dialécticas. La segmentación fonética del algoritmo fue comparada con los límites fonéticos reales obtenidos de la transcripción asociada con las señales de habla, hecha

manualmente por expertos fonéticos. El subconjunto de oraciones tiene un total de 20647 límites fonéticos.
En estos experimentos se detallan las modificaciones realizadas al algoritmo aquí presentado, explicando el impacto de dichas modificaciones en los resultados.
En primera instancia, se muestran resultados en la Tabla 7.4 de este algoritmo sin el uso de membresias difusas, en donde se puede observar un desempeño pobre. Se opta

Tamaño	ϕ	**Sd**	**Sc**	**% detección**	**% inserción**
5 msec	1.120	21103	14316	69.33	2.20
3 msec	0.085	20708	14791	71.63	0.67

Tabla 7.4: Desempeño del algoritmo sin utilizar membresias difusas

por utilizar membresias difusas de tres conjuntos difusos (alto, medio y bajo) para tener información detallada sobre la intensidad. Inicialmente, se utilizaron frames de 5 ms sin hacer el normalizado de membresias difusas, aplicando la fórmula 7.6.4 para obtener la distancia de frames adyacentes. Con estas cualidades, el algoritmo obtuvo un desempeño aceptable, pero por debajo de los últimos reportados en el estado del arte.

$$\delta(f_t, f_{t-1}) = \sqrt{\Sigma(\mu(f_t) - \mu(f_{t-1}))^2} \tag{7.6.4}$$

En la tabla 7.5 se muestra el desempeño del algoritmo de segmentación con distintos tamaños de frames. Se observa que cuando se disminuyó el tamaño del frame el desempeño fue ligeramente incrementado. Típicamente el tamaño de frames es de 20

Tamaño	ϕ	**Sd**	**Sc**	**% detección**	**% inserción**
5 msec	0.040	20946	15046	72.87	1.44
4 msec	0.036	21039	15353	74.35	1.89
3 msec	0.032	20910	15485	74.99	1.27

Tabla 7.5: Desempeño del algoritmo, con distintos tamaños de frames y sus respectivos parámetros

ms, sobrelapados en 10 ms cuando se utilizan vectores de características en la secuencia del tiempo en el proceso de segmentación. El uso de vectores de características implica tener un conjunto de ellas por cada frame (15 o más características), que incluso en algunos métodos del estado del arte, estas son agrupadas para efectos de segmentación. En nuestro algoritmo, aunque el tamaño del frame es muy pequeño, sólo se extraen valores escalares en cada uno de ellos, teniendo en promedio menos características (6.66) en cada 20 ms que las extraídas en trabajos previos.

Una ligera mejora tanto en la tasa de detecciones correctas como en la tasa de inser-

ϕ	**Sd**	**Sc**	**% detección**	**% inserción**
0.050	20642	15567	75.39	-0.02

Tabla 7.6: Desempeño del algoritmo con membresias difusas normalizadas

ciones cuando se aplica un normalizado de las membresias difusas. Los resultados se muestran en la Tabla 7.6.

Hasta este punto, los resultados mostrados se han obtenido utilizando la fórmula

ϕ	**Sd**	**Sc**	**% detección**	**% inserción**
0.0992	20668	15796	76.50	0.10

Tabla 7.7: Desempeño del algoritmo con frames no adyacentes

7.6.4. Puesto que la distancia obtenida entre frames adyacentes no detectaba algunos límites fonéticos, debido a que su significativa diferencia no aparece en frames adyacentes, se procede a obtener la distancia de frames no adyacentes (esto es, frames separados por un frame intermedio). Usando la función expresada en 7.6.3 se produce una importante mejora en el desempeño del algoritmo, que se muestra en la Tabla 7.7. Esta modificación a la métrica de distancia, resulta en un incremento arriba del 1 % en las correctas detecciones, manteniendo el porcentaje de la tasa de sobre segmentación cercana al 0 %. El espacio difuso se había establecido en el rango de la intensidad máxima y mínima relativa a cada señal, sin embargo, modificando la intensidad mínima a 25 dB se obtuvieron mejores resultados. Se realizaron pruebas con

ϕ	**Sd**	**Sc**	**% detección**	**% inserción**
0.0855	20666	16010	77.54	0.09

Tabla 7.8: Desempeño del algoritmo utilizando intensidad mínima de 25 dB

distintas métricas, obteniendo los resultados que se muestran en la Tabla 7.9. Aunque se obtiene un desempeño similar entre las funciones aplicadas, la función expresada en la fórmula 7.5.1 obtiene ligeramente un mejor desempeño, puesto que presentó la menor tasa de inserciones y la mejor tasa de correctas detecciones de límites fonéticos. Durante el desarrollo del algoritmo de segmentación se hizo el compromiso de calidad y rapidez, por lo que se experimentó con métricas de distancia de bajo costo computacional en comparación con otras como la métrica de Mahalanobis. Como se observa

en la Tabla 7.10, los resultados son similares excepto al utilizar la métrica de Pearson. Por otro lado, este algoritmo se ha experimentado sobre las mismas señales de habla

distancia	ϕ	**Sd**	**Sc**	**% detección**	**% inserción**
Chebyshev	0.0596	20642	15966	77.32	-0.02
Manhatan	0.1185	20705	15990	77.44	0.28
Pearson	0.3500	20867	13441	65.09	1.06
Spearman	0.1170	20671	16012	77.55	0.11

Tabla 7.9: Desempeño del algoritmo utilizando distintas medidas de disimilaridad sobre el corpus TIMIT

del corpus DIMEx100 utilizadas en el algoritmo anterior, con las mismas medidas de disimilaridad de la Tabla 7.9. A diferencia de los resultados obtenidos utilizando el corpus TIMIT, sobre el corpus DIMEx100 se obtiene un desempeño menor, uno de los factores que influye en este aspecto es el tipo de etiquetado. Sin embargo, se

distancia	ϕ	**Sd**	**Sc**	**% detección**	**% inserción**
Chebyshev	0.028	11230	8286	74.015	0.31
Manhatan	0.088	11314	8311	74.23	1.06
Pearson	0.5	11492	7740	69.13	2.65
Spearman	0.062	11354	8335	74.45	1.42
Euclidiana	0.039	11198	8264	73.81	0.026

Tabla 7.10: Desempeño del algoritmo utilizando distintas medidas de disimilaridad sobre el corpus DIMEx100

observó un comportamiento similar al obtenido en el corpus TIMIT con las mismas medidas de disimilaridad en el sentido que obtienen entre ellas un desempeño similar en la segmentación, ver Tabla 7.9, aunque se obtuvo ligeramente un mejor resultado utilizando la medida de Chebyshev, puesto que al utilizarla se presenta la mayor tasa de detecciones correctas y una de las tasas más bajas de inserciones. La diferencia es despreciable en los resultados por aplicar las distintas medidas de disimilaridad, exceptuando la medida de Pearson que en ambos corpus promovió un resultado pobre.

7.7. Algoritmos de Segmentación con Características Vectoriales

En esta sección se describen de forma breve los algoritmos para la segmentación en fonemas que hacen uso de características vectoriales del habla, cabe mencionar que se

ha utilizado el mismo enfoque que el empleado en la segmentación con características en el dominio del tiempo. Las características utilizadas en los experimentos son los espectros de Mel, espectros Bark y los MFCC.

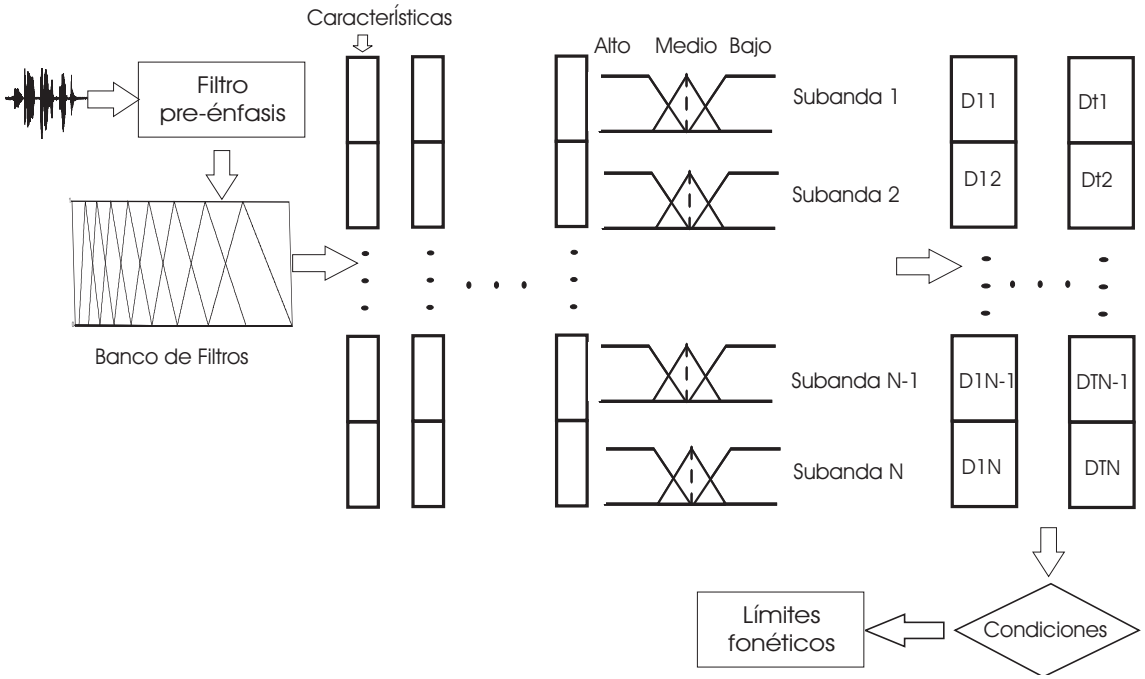

Figura 7.5: Diagrama a bloques del algoritmo con membresias difusas en sub-bandas

La señal de habla es pasada por el filtro pre-énfasis. Como se ha explicado en el capítulo anterior, los vectores obtenidos por unidad de tiempo son tratados también como objetos cada uno. Para cada intervalo de frecuencia se define un espacio difuso con el objetivo de tener un mayor detalle de cada cuantificación espectral(espectros o cepstrums). El espacio difuso es definido obteniendo el espectro máximo y mínimo de cada sub-banda, con conjuntos traslapados entre sí en un 50 %. Para cada espectro o cepstrum se obtienen las membresías *Alto, Medio y Bajo* relativas al intervalo de frecuencia en el que residen. Para obtener una representación cuantitativa de la existencia o no de un cambio espectral entre frames, se lleva a cabo una sumatoria de las distancias correspondientes a cada sub-banda para un instante de tiempo t y de esta manera determinar el grado de disimilaridad entre frames comparados. Las operaciones para obtener el grado de disimilaridad sobre los valores vectoriales de

frames comparados se definen por:

$$\delta_{t-1_i}(f_{t_i}, f_{t-2_i}) = \sqrt{\sum (\mu_A(f_{t_i}) - \mu_A(f_{t-2_i}))^2} \tag{7.7.1}$$

$$\delta_{t-1}(f_t, f_{t-2}) = \sum_{i=1}^{N} \delta_i(f_{t_i}, f_{t-2_i}) \tag{7.7.2}$$

Donde N es el número de sub-bandas de las caracterísicas. De igual manera que los algoritmos previamente explicados, se hace una comparación entre frames separados por un frame de distancia, definiendo la distancia en ese frame intermedio.
El resultado de la sumatoria de distancias es analizada para determinar la existencia de límites fonéticos bajo un grupo de condiciones que se definen como sigue:

1. $v_t > v_{t-1}$ y $v_t > v_{t+1}$

2. $v_t > \phi$

El diagrama del algoritmo se muestra en la Figura 7.5.

7.7.1. Experimentos

Se efectuaron pruebas sobre el corpus TIMIT y DIMEx100, en donde se obtiene mejores resultados en el esquema de espectros de Mel que en cepstrum de Mel para la representación de la señala de habla. El número de filtros presentados en la Tabla 7.11

Característica	Filtros	ϕ	Sd	Sc	% detección	% inserción
Espectros Mel	8	1.36	20630	15796	76.50	-0.08
Espectros Bark	8	1.67	20621	15568	75.40	-0.12
MFCC	3	3.1	21142	15317	74.18	2.39

Tabla 7.11: Desempeño del algoritmo con características vectoriales sobre el corpus TIMIT

obedecen al mejor resultado presentado en los experimentos, donde se puede observar que para efectos de segmentación se obtiene una adecuada definición utilizando tan solo 3 filtros.
La información correspondiente a las pruebas realizadas sobre el corpus DIMEx100 se ve en la Tabla 7.12. El algoritmo difuso basado en intensidad presenta resultados

Característica	Filtros	ϕ	Sd	Sc	% detección	% inserción
Espectros Mel	12	2.82	11260	8924	79.89	0.80
Espectros Bark	12	1.2	11206	8602	76.83	0.09
MFCC	4	10.5	10906	1995	71.41	-2.58

Tabla 7.12: Desempeño del algoritmo con características vectoriales sobre el corpus DIMEx

similares a los resultados utilizando los espectros de Bark, sin embargo, los experimentos que hacen uso de espectros de Mel presentan el mejor de los resultados de las pruebas, alcanzando aproximadamente el 80 % de correcta segmentación con una tasa de inserciones aproximada al 0 %. En ambos corpus se obtiene un comportamiento similar, en el que se destaca la eficiencia de los espectros de Mel al segmentar.

7.8. Conclusión y Discusión

En este capítulo se hizo descripción detallada de los algoritmos de segmentación implementados. En cada uno de ellos se aborda el problema de la segmentación en un enfoque simple, consistiendo en medidas de disimilaridad aplicadas a lo largo de la señal del habla a segmentar. Se hace uso de diversas características acústicas del habla presentes en el dominio del tiempo como la energía, amplitud e intensidad de manera independiente, nunca antes probadas de manera individual en la segmentación fonética. A diferencia de métodos como [1], que hacen uso de varias características (44) en el dominio del tiempo y frecuencias simultaneamente para efectuar la segmentación, en los algoritmos aquí propuestos se utilizan a lo mucho dos características de este tipo. La ventaja de estos algoritmos es que promueven un procesamiento simple, tanto en la extracción de dichas características como en su procesamiento, sin tener la necesidad de hacer uso de algún tipo de codificación del habla más elaborada como vectores de características.

Se utilizaron membresías difusas sobre valores escalares de amplitud, intensidad o energía para obtener un mejor detalle de ellos, además de una normalización difusa de las membresias obtenidas previa a la aplicación de alguna medida de disimilaridad. Este simple procesamiento incrementa el desempeño de la segmentación en un 6 % aproximadamente con respecto al uso directo de valores escalares , que fué probado sobre el corpus TIMIT.

Por otro lado, también se implementa un algoritmo que hacen uso de esquemas de

codificación de la señal del habla, dada en vectores de características por secuencia de tiempo como espectros de Bark, espectros Mel y coeficientes cepstrales en frecuencia de Mel (MFCC). En base en experimentos realizados el mejor de estos esquemas de codificación son los espectros de Mel.

7.9. Discusión de Resultados

Se alcanzó un desempeño competitivo, utilizando características básicas en el dominio del tiempo, teniendo la ventaja de hacer uso de información mínima extraída de la señal de habla. En uno de los enfoques se hace uso de membresías difusas que ayudaron a incrementar el desempeño del algoritmo en aproximadamente 6 %. Los coeficientes cepstrales en frecuencias de Mel(MFCC) son un modelo paramétrico de la generación de voz, mientras que los espectros de Mel y Bark emulan el sistema auditivo humano. El esquema de codificación dado en MFCC un desempeño por debajo de los espectros de Bark y Mel, y con base en los resultados obtenidos, se puede concluir que la segmentación fonética del habla tiene mejor desempeño utilizando espectros, puesto que los *cepstrums* filtran la información proveniente de la señal de exitación (cuerdas vocales) que podría aportar información útil en la segmentación.
Los algoritmos que hacen uso de vectores de características presentaron ligeramente una mayor robustez al etiquetado y frecuencia de muestreo, que los que utilizan características básicas.

7.9.1. Análisis de Resultados

Se analizan los resultados en el peor y mejor caso de los corpus TIMIT y DIMEX respectivamente. Se denotan con (o) los puntos de segmentación detectados por el algoritmo en cuestion, con (*) los *puntos reales* de segmentación, y con (x) los puntos insertados. Iniciaremos analizando los mejores casos:
En la figura 7.6, se observa que la única inserción existente tiene lugar por el efecto de coarticulación, en donde el fonéma /ux/ hace resonancia sobre el fonema /w/, el cual es percibido por el algoritmo como una transición fonética (por que existe un tenue cambio espectral). En este ejemplo se puede observar que la precisión del algoritmo es similar al etiquetado real.
El mejor caso sobre el corpus DIMEX, se ve en la figura 7.7, en donde se omite un

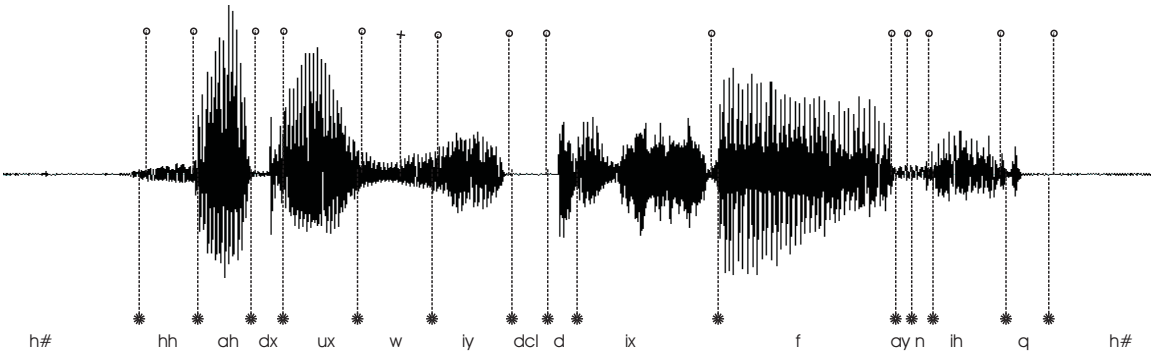

Figura 7.6: Puntos de segmentación del mejor caso TIMIT

límite fonético dentro de un diptongo en la palabra *serie*, estos son uno de los casos más difíciles de detectar. Por otro lado, el fonema /r/ es pronunciado con mucha rapidez que lo hace casi imperceptible. Estas dificultades son acentuadas cuando el locutor tiene altas tasas de habla. Las inserciones son provocadas por casos como, al pronunciar la palabra *recreación* como *rekereación*, que por la rapidez de elocución es casi imperceptible. Algunos sonidos como las vocales al terminar las frases o palabras, no son terminadas abruptamente, si no que en ocasiones se queda la resonancia que puede ser detectado como un falso límite fonético.

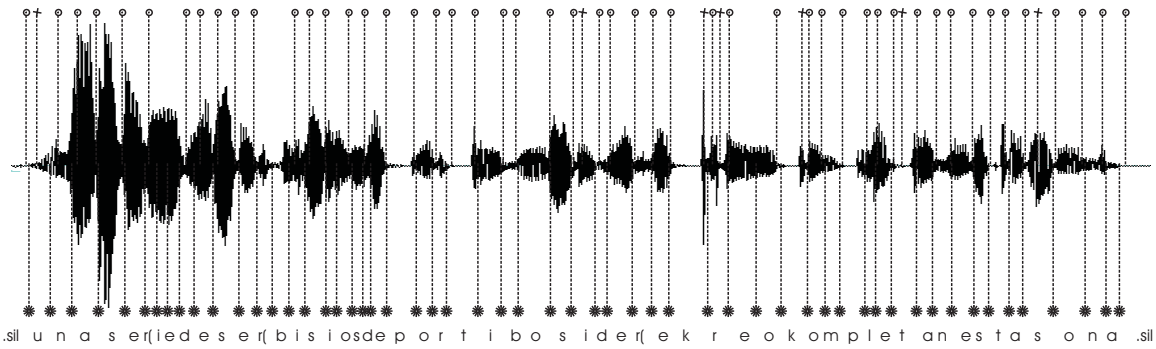

Figura 7.7: Puntos de segmentación del mejor caso DIMEX

El desempeño de los mejores casos en ambos corpus se muestra en la tabla 7.13. El desempeño es afectado principalmente por características locales del hablante como dicción, tasa de habla (rapidez) e intensidad de la misma. En la figura 7.8, se observa la señal del habla de un locutor femenino, el cual tiene una baja intensidad en la pronunciación en la mayor parte de la oración, lo cual implica que el algoritmo tenga

Base de Datos	fonemas	correctos	detectados	% detección	% inserción
TIMIT	15	14	15	93.33	0
DIMEX	56	51	58	91.07	3.57

Tabla 7.13: Parámetros de desempeño en los mejores casos en TIMIT y DIMEX

un desempeño pobre.

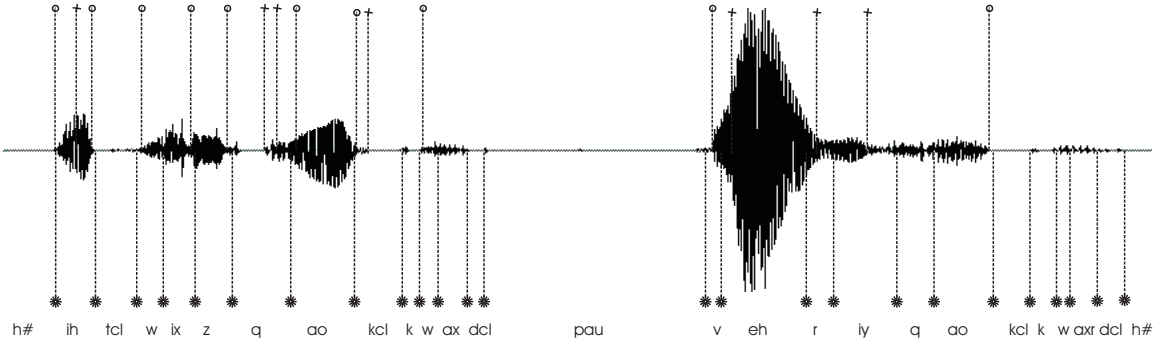

Figura 7.8: Puntos de segmentación del peor caso Timit

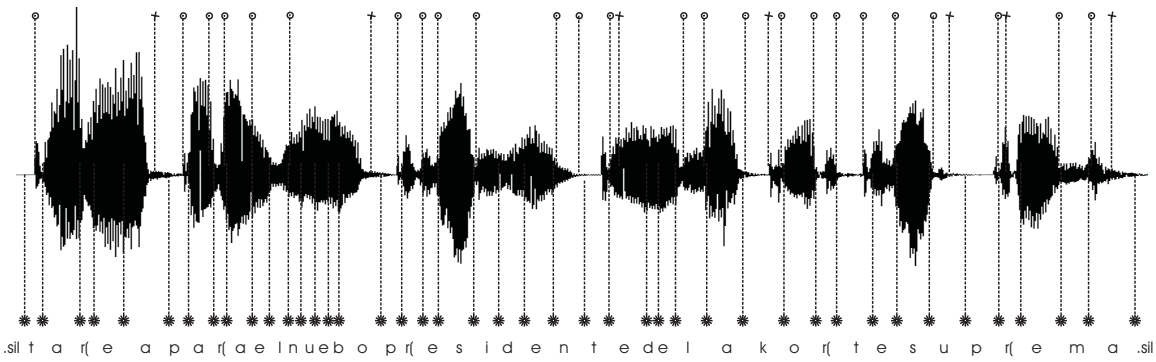

Figura 7.9: Puntos de segmentación del peor caso Dimex

El ejemplo de la señal del peor caso en el corpus DIMEX mostrada en la figura 7.9, en la cual percibe una tasa alta de habla que acentúa los problemas de coarticulación, existiendo transiciones fonéticas poco prominentes. El desempeño de los peores casos en ambos corpus se muestra en la tabla 7.14. La naturalidad del habla agrega dificultad tanto al proceso de reconocimiento como al de segmentación. Puesto que se utiliza una amplia gama de hablantes en los experimentos, los algoritmos propuestos se encuentran en la clasificación de los independientes de hablante.

Base de Datos	fonemas	correctos	detectados	% detección	% inserción
TIMIT	25	10	17	40.00	-31.99
DIMEX	42	25	32	59.52	-23.81

Tabla 7.14: Parámetros de desempeño en los peores casos en TIMIT y DIMEX

Capítulo 8

Conclusiones y Perspectivas

Los objetivos de esta tesis de Maestría han sido alcanzados satisfactoriamente, explorando distintas representaciones del habla e idiomas distintos.
En esta tesis se desarrollaron nuevos métodos para la detección de límites fonéticos con independencia del hablante, vocabulario y habla expresada naturalmente utilizando características en el dominio del tiempo y frecuencias respectivamente. Los métodos son independientes de texto, puesto que no hacen uso a *priori* de información adicional presente en la señal del habla para la segmentación. Los métodos desarrollados no requieren de ningún tipo de entrenamiento para llevar a cabo la segmentación de fonemas y han sido aplicados tanto a expresiones de habla inglesa como española. Se ha utilizado el enfoque basado en medidas de disimilaridad para la detección de cambios espectrales, dando resultados aceptables.
Las membresias difusas permiten tener un mayor detalle de las características en el dominio del tiempo, equiparandose al desempeño de los esquemas de codificación vectoriales del habla, reduciendo la cantidad de información extraída de la señal, el proceso de extracción de la misma y el tiempo de segmentación. Por otro lado, empleando características basadas en el sistema auditivo humano (espectros y cepstrums), las membresias difusas logran una mejoría mínima en el desempeño de la segmentación. Los métodos fueron probados utilizando 64 hablantes de distintas regiones del idioma inglés, y 30 hablantes del idioma español, esto es, independencia de hablante con resultados competitivos. Las dificultades identificadas al segmentar son ocasionadas por el habla expresada de forma natural, donde principalmente al terminar frases y oraciones se hace una inadecuada pronunciación de palabras, que el

ser humano llega a interpretar para precisarlas. Por otro lado, el habla natural afecta el proceso de segmentación, debido a que la resonancia de fonemas puede transmitirse a los fonemas adyacentes. Otros factores del habla natural que influyen en la segmentación son la intensidad y tasa de habla, dado que afectan las transiciones fonéticas. Una de las dificultades ajena al habla natural es la union de fonemas vocal-vocal, silencio-fricativas, fricativas-silencio.
La mayoría de los trabajos reportados en el estado del arte han experimentado con una serie de restricciones, como dependencia de hablante, texto, vocabulario y sentencia reducidos, por lo que el problema de segmentación fonética ha sido escasamente probado sin estas restricciones. En esta tésis se han probado los algoritmos sin las restricciones mencionadas, y en diferentes idiomas.

8.1. Trabajo Futuro

El desempeño de los métodos propuestos se evaluó en términos de detecciones correctas y sobre-segmentación. Sin embargo, el desempeño se ve afectado por el tipo de etiquetado y conjunto de fonemas seleccionados, y sin duda, los errores que genera el etiquetado humano.
Existen errores sistemáticos que se deben considerar para minimizar el impacto en el desempeño de los algoritmos, por ejemplo, el manejo de milisegundos en el algoritmo requiere de una alta precisión, puesto que la pérdida de milésimas en las variables manejadas, podría de manera acumulativa afectar la precisión, y en consecuencia afectar en la comparación entre instancias de tiempos de límites fonéticos obtenidos por los algoritmos y las instancias de tiempo existentes en las bases de datos para determinar su validez.
El uso de umbrales adaptativos que permitan discriminar activamente entre límites válidos y puntos de inserciones, podrían ser utilizados. El algoritmo puede ser probado con un proceso de corrección de inserciones a *posteriori* a la segmentación, reconociendo los límites de las inserciones, re-segmentando segmentos candidatos a ser divisibles como uniones vocal-vocal, llevando el método a un multi nivel de segmentación. Otros esquemas de codificación de la señal del habla pueden ser probados.

Apéndice A

Apéndice

A.1. Publicaciones

On the Processing of Fuzzy Patterns for Text Independent Phonetic Speech Segmentation. 11^{th} *Iberoamerican Congress on Pattern Recognition*. LNCS 4225, pp. 437−445, 2006. Springer-Verlag Berlin Heidelberg 2006.

Bibliografía

[1] Suh Y. and Lee Y. Segmentation of continuous speech using multi-layer perceptron. *IEEE Trans. Speech and Audio Proc.*, 1999.

[2] Fernández L. *Aportaciones a la Mejora de los Sistemas de Reconocimiento*. PhD thesis, Universidad de Vigo, 2001.

[3] Hansen J. Pellom B. Automatic segmentation of speech recorded on unknown noisy channel characteristics. *Speech Communication*, 1998.

[4] Mayora O. Segmentazione automatica di fonemi per aplicazioni di riconoscimento vocale. Technical report, Università di Genova, 2000.

[5] Bernard E. Cole R. Hu Z., Schalwyk J. Speech recognition using syllabe-like units. *ICSLP '96*, 1996.

[6] Korhonen P. Unsupervised segmentation of continuous speech using vectorautoregressive modeling. Master's thesis, Helsinki University of Technology, 2004.

[7] Dalsgaard P. Petek B., Andersen O. On the robust automatic segmentation of spontaneous speech. *Proceedings of ICSLP '96*, 1996.

[8] Esposito A. Aversano G. New text-independent method for phoneme segmentation. *IEEE Midwest Symposium on Circuits and Systems*, 2001.

[9] Vidal E. Casacuberta F. *Reconocimiento Automático del Habla*. Marcombo Boixareu, 1987.

[10] Reyes C. and Bandler W. Una introducción al reconocimiento automático del habla. Technical Report 93-041, Dept. of CS University of Tallahasee, Florida, 1992.

[11] Katagiri S. *Handbook of Neural Networks for Speech Processing.* Artech House Inc., 2000.

[12] Philips M. Brill S. Pilant A. Specker P. Cole R., Stern M. Feature-based speaker-independent recognition of isolated english letters. *Proceeding of the IEEE International Conference on Acoustics Speech, and Signal Processing*, 1983.

[13] Cole R. Yan Y., Fanty M. Speech recognition usnig neural networks with fordward-backward probability generated targets. *Proceedings of the International Conference on Acoustic Speech and Signal Processing*, 1997.

[14] Schalkwyk J. Sutton S. Cole R. Cosi P., Hosom J. Connected digit recognition experiments with ogi toolkit's neural network and hmm-based recognizers. *In Proceedings of the 4^{th} IEEE Workshop on Interactive Voice Technology for Telecommunications Applications(IVTTA-ETWR98)*, 1998.

[15] Cole R. Hagen A., Pellom B. Children's speech recognition with application to interactive books and tutors. *Proceedings IEEE Workshop on Automatic Speech Recognition and Understanding(ASRU)*, 2003.

[16] Zue V. Cole R. Spoken language input. *Survey of the State of Art in Human Language Technology*, 1997.

[17] Lamel L. Garofolo J. Darpa timit acoutic-phonetic continuous speech corpus. Technical report, U.S Departament of Commerce, 1993.

[18] Juang B. Rabiner L. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[19] Geetha T.V Saraswhati S. and Saravanan K. Integrating language independent segmentation and language dependent phoneme based modeling for tamil speech recognition system. *Asian Journal of Information Technology*, 2006.

[20] Volkman J. Stevens S. The relation of pitch to frequency. *American Journal of Psychology*, 1940.

[21] Tukey J. Bogert P., Healy M. The quefrency analysis of time series for echoes: Cepstrum, psuedo-autocovariance, cross-cepstrum and sa phe cracking. *In Proceedings of the Symposium on Time Series Analysis*, 2006.

[22] Cuétara J. Castellanos H. López I. Pineda L., Villaseñor L. Dimex100: A new phonetic and speech corpus for mexican spanish. *Survey of the State of Art in Human Language Technology*, 1997.

[23] Esposito A. Aversano G. Automatic parameter estimation for a context-independent speech segmentation algorithm. *TSD 2002*, 2002.