ESCUELA SUPERIOR POLITECNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

"Diseño y Simulación de un Sistema de Reconocimiento del Habla, Aplicando el Algoritmo de Alineamiento Temporal Dinámico en Matlab"

TESIS DE GRADO

PREVIO A LA OBTENCION DEL TITULO DE:

Ingeniero en Electrónica y Telecomunicaciones

PRESENTADO POR:

Manuel Ignacio Paredes Vera

GUAYAQUIL – ECUADOR

Año: 2009

ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

"Diseño y Simulación de un Sistema de Reconocimiento del Habla, Aplicando el Algoritmo de Alineamiento Temporal Dinámico en Matlab"

TESIS DE GRADO:

Previo a la obtención del Título de:

INGENIERO EN ELECTRÓNICA Y TELECOMUNICACIONES

Presentado por:

Manuel Ignacio Paredes Vera

GUAYAQUIL – ECUADOR
Año: 2009

AGRADECIMIENTO

A Dios por la fuerza y fe para conseguir los objetivos propuestos,

a mis padres Manuel Paredes, Lupita Vera Soto,

familiares, mi novia Alexandra Saavedra Ordóñez ,

amigos como Diana Jaramillo por su apoyo incondicional,

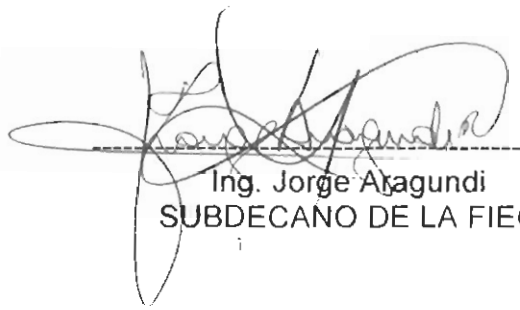
a los profesores que me ayudaron en mi formación profesional,

y a mi director de tesis, Ing. Ronald Ponguillo por su guía y consejos.

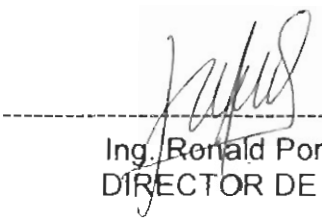
DEDICATORIA

PADRES
FAMILIARES
AMIGOS

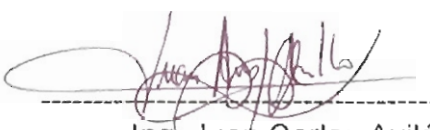
TRIBUNAL DE GRADUACION



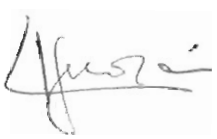
Ing. Jorge Aragundi
SUBDECANO DE LA FIEC



Ing. Ronald Ponguillo.
DIRECTOR DE TESIS



Ing. Juan Carlos Avilés
VOCAL PRINCIPAL



Ing. Carlos Jordán
VOCAL PRINCIPAL

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de este Trabajo Final, nos corresponde exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL"

(Reglamento de Graduación de la ESPOL)

Manuel Paredes Vera

RESUMEN

El presente trabajo consiste en el diseño de un sistema de reconocimiento del habla con restricciones tales como reconocimiento de palabras aisladas y monolocutor. Este documento está dividido en cinco capítulos.

En el capítulo 1 muestra la justificación del proyecto, el alcance del mismo y una definición formal sobre los sistemas de reconocimiento del habla: antecedentes, problemas, clasificaciones y técnicas aplicadas en el mismo.

En el capítulo 2 se describe todo el proceso de la producción del habla en el ser humano con el fin de conocer los elementos que actúan en él y de esta manera construir un modelo digital que es posteriormente utilizado en el diseño del sistema de reconocimiento.

En el capítulo 3 se analiza a la señal de habla desde el punto de vista articulatorio y acústico. Para conocer las características más importantes de una señal de habla.

En el capítulo 4 se realiza el diseño del sistema de reconocimiento del habla. Utilizando los conocimientos de los capítulos anteriores. Aquí se definen las partes que constituyen este sistema y todos los algoritmos aplicados en cada uno.

En el capítulo 5 se efectúa la evaluación del sistema, mediante los experimentos realizados al mismo. Los datos obtenidos serán analizados para posteriormente realizar las conclusiones.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS

ASR Automatic speech recognition

VAD voice activity detector

MFCC Mel-frecuency cepstral coefficient

DTW Dynamic time warping

INTRODUCCIÓN

En la actualidad, existe un creciente interés en el desarrollo de interfaces que permitan la interacción entre el hombre y la máquina, siendo la comunicación táctil (se usan el tacto y la vista) la más utilizada. Una forma nueva y natural de ésta interacción es por medio de la voz, el sistema que se desea implementar utiliza como control la voz (oído para verificación), algo necesario en situaciones donde las manos y vista estén ocupados o peor aun se tenga alguna discapacidad que impida su uso.

Ejemplos de algunos sistemas controlados por medio de voz son el marcado telefónico, la selección de servicios mediante palabras claves, entre otros. El uso de este tipo de tecnología puede crear la posibilidad de que la población en general (incluyendo personas con discapacidad física) puedan usar las computadoras, las telecomunicaciones y equipo para el manejo de transacciones, mensajes, información y de control de varios dispositivos.

Un sistema de reconocimiento de voz es una herramienta computacional capaz de procesar y reconocer la información contenida en la señal de voz. En este proceso, las palabras pronunciadas son adquiridas como señales eléctricas a través de un micrófono, luego son digitalizadas para finalmente ser interpretadas

por el procesador del sistema, el cual mediante algoritmos matemáticos extrae patrones o parámetros característicos de esta señal, con la finalidad de realizar una clasificación y reconocimiento de éstos ya sea para realizar una conversión de voz a texto, o interactuar con la computadora mediante comandos de voz.

El presente trabajo abarca el desarrollo de una aplicación para el Reconocimiento de la voz utilizando la Matlab versión 7.3.0

CAPÍTULO 1

1. ESPECIFICACIÓN DEL TEMA

Cuando pensamos en un sistema de reconocimiento automático del habla (RAH), la primera imagen que se nos viene a la mente es la conversión de la voz a un flujo de palabras que deben de ser almacenadas y luego comparadas. En este capítulo se explica el problema a solucionar a través de la ejecución de este proyecto y muestra la magnitud del problema de reconocimiento automático del habla con el fin de definir una solución diseñada.

1.1 Definición del problema

La comunicación oral es sin duda una de las capacidades más fascinantes del ser humano. Ya que habilita a los humanos a intercambiar información de una manera directa, simple y efectiva (1). En la actualidad el ser humano a mas de comunicarse con sus semejantes se comunica con máquinas por medio de una interfaz táctil. Pero existen situaciones en donde las manos y vista están ocupadas o puede existir una discapacidad que impide su uso. En estos casos es necesario una interfaz alternativa entre el hombre y la maquina, por lo que surge la necesidad de dotar a las máquinas con la capacidad de recibir mensajes orales.

La presentación de esta tesis tiene como propósito la realización de un sistema de reconocimiento automático del habla utilizando el simulador matlab en su diseño e implementación final.

Adicionalmente la tesis mediante su contenido y pruebas de experimentación desarrolladas puede tomarse como una referencia, para solventar dudas para el desarrollo futuro de una aplicación real.

1.2 Reconocimiento automático del habla

El reconocimiento automático del habla es el procedimiento por el cual se convierte una señal de voz, capturada por un micrófono, en un conjunto de símbolos. Dichos símbolos son una representación paramétrica de cada palabra entrenada por el usuario, la cual es almacenada y finalmente comparada para así conseguir una comunicación entre el usuario y la máquina, este tipo de interfaz entre el hombre y la máquina tiene una amplia gama de aplicaciones:

- Acceso a sistemas de información automáticos,
- ayuda a minusválidos.
- traducción automática.
- Operaciones comerciales o bancarias automáticas.
- control oral de sistemas, etc.

En la figura 1.1 se muestra un ejemplo de aplicación de un sistema de reconocimiento del habla, donde a una persona se comunica hacia un computador.

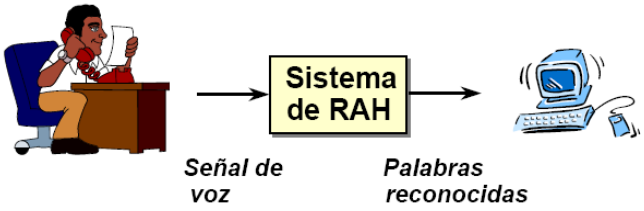


Figura 1.1 Ejemplo de aplicación de un sistema RAH.

1.2.1 Antecedentes

A continuación una breve reseña histórica sobre los avances del reconocimiento del habla desde sus inicios hasta la actualidad

En 1870 Alexander Graham Bell

Quiso construir un sistema o dispositivo de ayuda con el habla a las personas con problemas auditivos y el resultado de su trabajo fue el teléfono.

En 1880 Tihamir Nemes

Solicitó un permiso para desarrollar y patentar un sistema de transcripción automática que identificara secuencias de sonidos y los imprimiera como texto. Pero fue rechazado como “Proyecto no realista”

30 anos después AT&T Bell Laboratorios

Construyó la primera máquina capaz de reconocer voz, basada en tablas de los 10 dígitos en ingles.

A mediados de la década de los 60

La mayoría de los investigadores reconoció que era un proceso más difícil de lo que había pensado en un principio y comenzaron a trabajar en sistemas más específicos (2), como por ejemplo, los sistemas dependientes de locutor. Trabajan con un flujo discreto de habla, con espacios y pausas entre palabras y un vocabulario pequeño, menor o igual a 50 palabras.

A principios del año 1970

Se produjo el primer producto de reconocimiento de voz, el VIP100 de Threshold technology Inc. Este reconocedor podía utilizar un vocabulario pequeño, dependiente del locutor, y reconocer palabras discretas. Esta empresa ganó el "US National Award" en 1972.

Luego nació el interés de ARPA del departamento de defensa norteamericano. El proyecto financiado por ARPA buscaba el reconocimiento de habla continua, de vocabulario grande. Impulsó que el trabajo de los investigadores se dirigiera

más al entendimiento del habla. Los sistema empezaron a incorporar módulos de análisis léxico o conocimiento léxico, análisis sintáctico, de la estructura de las palabras, análisis semántico y análisis pragmático que analiza la intención.

En los ’80

Se aplicaron los conceptos de dynamic time warping. Se produce un importante cambio de paradigma de comparación de plantillas hacia el modelado estadístico/probabilístico como un gran avance de aproximación al reconocimiento de voz.

A mitad de los ’80

Se hizo masiva una técnica que revolucionó el campo de reconocimiento se trata de los modelos ocultos de Markov o HMM que obtuvo excelentes resultados en el modelado de señales de voz y virtualmente indispensable hoy en día. Se reintroduce el uso de redes neuronales(ANN) que habían vencido algunos obstáculos de tipo conceptual y de recursos necesarios para su implementación. También se comenzó a experimentar con reconocimiento continuo de vocabularios largos independientes del locutor.

En los ´90

Se comenzó a hacer énfasis en interfaces de lenguaje natural, y recuperación de la información en grandes documentos de voz, continuó la investigación de reconocimiento continuo en vocabularios grandes y a usarse masivamente a través de redes telefónicas, también en el estudio de sistemas en condiciones de ruido.

Antes y durante la mitad de los ´90

Se dio la investigación de sistemas híbridos HMM-ANN, que también han dado excelentes resultados siendo la excelencia de los motores de reconocimiento de voz de hoy en día.

Desde la última década hasta nuestros días

La mayor parte de la comunidad científica, está siguiendo las pautas establecidas por los dos programas americanos DARPA relacionados con el problema del reconocimiento automático de habla continua (Continuos Speech Recognition) y el problema de la compresión del lenguaje hablado (Spoken Language Understanding). Los resultados del trabajo de los últimos años han sido unas técnicas basadas en la probabilidad para resolver el problema del habla. Las técnicas más reconocidas son: modelos ocultos de Markov o

modelos de Markov con redes neuronales, el alineamiento temporal dinámico, los modelos de lenguaje estocásticos, y la teoría de decisión bayesiana. (3)

1.2.2 Problemas

Algunos de los problemas presentes en el reconocimiento automático de habla son inherentes a la complejidad del sistema. Otros, sin embargo son debidos a las limitaciones de los sistemas automáticos y al hecho de que no conocemos perfectamente el proceso de la comunicación oral. Los principales problemas que dificultan el reconocimiento automático de habla son:

Variabilidad:

Dos señales acústicas correspondientes a una misma cadena de palabras nunca son idénticas, incluso en el caso en que sean pronunciadas por el mismo locutor y en las mismas condiciones. La variabilidad aumenta cuando cambian las condiciones físicas o psíquicas del locutor (variabilidad intralocutor), o cambia el locutor (variabilidad interlocutor) o el entorno en el que se adquiere la señal de voz (cambia el micrófono, las características acústicas del local, etc).

Ruido:

En las aplicaciones reales, los sistemas deben operar en entornos ruidosos. Esto disminuye la precisión de los reconocedores. El ruido afecta de dos formas: por una parte el ruido se superpone a la señal de voz, enmascarándola y alterando las características de ésta. Por otra parte, en presencia de ruido los locutores alteran su forma de hablar.

Continuidad:

No existe separación entre las palabras que componen una frase cuando ésta es pronunciada de forma natural. Los fonemas aparecen también concatenados. Además, la pronunciación de un fonema se ve afectada por los fonemas anterior y posterior debido al efecto de coarticulación.

Multiinteractividad:

En el proceso de percepción y comprensión del habla existen varios niveles interrelacionados que aportan informaciones complementarias. En el nivel acústico se analizan las características acústicas de la señal de voz. En el nivel fonético se determinan las unidades correspondientes a los sonidos elementales. En el nivel sintáctico se aplican reglas a las unidades fonéticas, descartándose combinaciones no permitidas. En el nivel semántico se llega a la

comprensión del mensaje transmitido eliminando interpretaciones absurdas. Todos estos niveles presentan fuertes interacciones entre sí y cada uno de ellos aporta una información útil al proceso de reconocimiento. En la actualidad no existen formalismos eficientes que permitan la total integración de las informaciones correspondientes a los distintos niveles.

Volumen de datos a procesar:

Debido a los problemas anteriores es necesario almacenar y procesar gran cantidad de información. Esto hace necesarios sistemas automáticos de gran capacidad y velocidad para desarrollar aplicación.

1.2.3 Clasificación

Debido a los problemas que aparecen en el reconocimiento del habla, a los sistemas se les suelen imponer restricciones. Con estas restricciones se intenta controlar la variabilidad o simplificar la interacción entre los distintos niveles de reconocimiento, con el objeto de mantener el rendimiento del reconocedor dentro de límites tolerables. Las restricciones se refieren principalmente a los siguientes aspectos:

Usuarios:

En cuanto al espectro de usuarios para los que está preparado el sistema, los reconocedores de voz se clasifican, por orden de dificultad en:

- **Monolocutor:** el sistema está entrenado por un solo locutor y está preparado únicamente para reconocer voz emitida por este.
- **Multilocutor:** el sistema se entrena por varios locutores y el reconocimiento se realiza con estos mismos locutores.
- **Independiente de locutor:** el sistema está entrenado con un número suficientemente grande de locutores de modo que funciona aceptablemente para cualquier locutor.

Ruido:

Con respecto al ruido, los sistemas aplican distintas técnicas para su robustecimiento, siendo optimizados para las condiciones en las que han de operar. Las técnicas para el robustecimiento son variadas y van desde la utilización de representaciones poco sensibles al ruido hasta la adaptación de la señal de voz a las condiciones de referencia, limpiando la señal contaminada. Otra posibilidad es entrenar el sistema con voz adquirida en entornos similares a aquellos en los que debe operar.

Por la forma de afrontar el ruido, los sistemas se pueden clasificar globalmente en robustos y no robustos dependiendo de que apliquen o no alguna técnica de compensación o robustecimiento. Una clasificación más precisa requerirá considerar las técnicas aplicadas y tener en cuenta el rendimiento de los sistemas en distintas condiciones, para distintos niveles y para distintos tipos de ruido.

Forma de pronunciación:

En cuanto a la forma de pronunciación nos encontramos con tres tipos de sistemas:

- **Palabras aisladas:** las palabras son pronunciadas con pausas entre ellas de 0.1 a 0.5 segundos, de modo que una palabra no se ve afectada por la que la precede o sigue. Además, existen técnicas para la localización del principio y final de la señal acústica correspondiente a la palabra, lo que reduce la complejidad del reconocimiento. Por ello, son los más sencillos.

- **Habla continua:** las palabras son pronunciadas sin pausas, de manera natural, existiendo una fuerte coarticulación entre los diferentes sonidos extremos de las palabras, lo que dificulta su reconocimiento.

- **Habla espontánea:** el locutor habla con una naturalidad total, con dudas, posibles repeticiones, paradas, etc. La dificultad del reconocimiento es máxima. Hay otros factores que contribuyen a definir la complejidad de un sistema. Por ejemplo, la existencia de una gramática que restrinja el número de frases permitidas.

Vocabulario:

El tamaño del vocabulario es determinante en la complejidad de la tarea y suele condicionar las técnicas aplicadas para el reconocimiento. Usualmente se habla de sistemas de pequeño vocabulario cuando no excede las 100 palabras, de vocabulario mediano si no llega a 1000, y de gran vocabulario si sobrepasa las 1000 palabras. En los sistemas de vocabulario reducido, la unidad de reconocimiento suele ser la palabra. Con vocabularios superiores es necesario utilizar unidades menores, como la sílaba o el fonema (o fonemas dependientes de contexto), con el objeto de mantener un número de unidades reducido.

Representación de las restricciones

Una vez definidos las restricciones que puede tomar un sistema de reconocimiento, es posible representar estas limitaciones mediante un diagrama de Kiviat. En la figura 1.2 se muestra el diagrama de Kiviat de las restricciones que puede poseer un sistema de reconocimiento automático del habla.

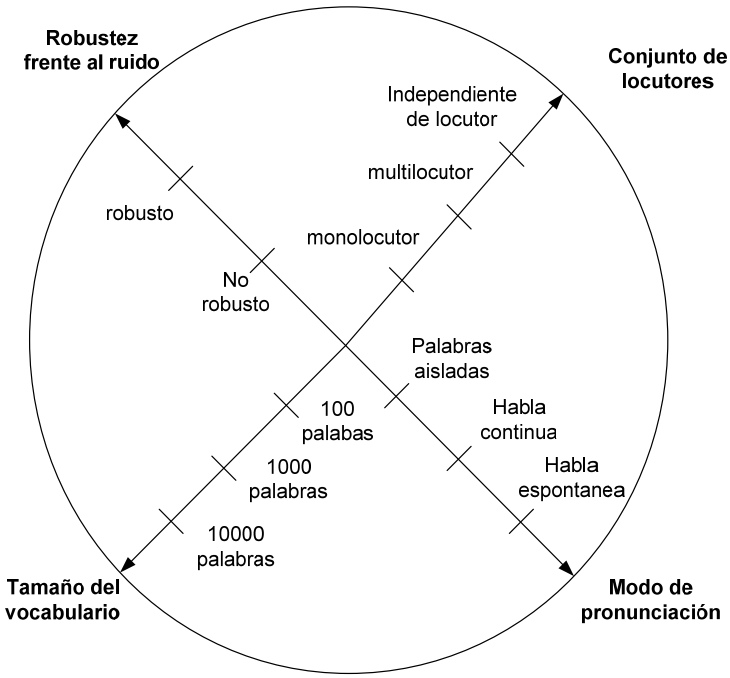


Figura 1.2: Diagrama de Kiviat de las restricciones de los sistemas de reconocimiento del habla.

Hacia el centro del diagrama de Kiviat de la figura 1.2, nos encontramos con los sistemas más restrictivos, mientras que los más flexibles son los que cubren una superficie mayor del diagrama encontrándose sus características en la periferia.

A medida que progresa la investigación en las tecnologías del habla, es posible el desarrollo de sistemas cada vez menos restrictivos siendo posible que estos trabajen de manera eficiente en condiciones acústicas desfavorables, con mayores vocabularios, permitiendo una pronunciación más espontánea y cubriendo un rango mayor de locutores.

A continuación se muestran dos aplicaciones de reconocimiento automático reales que operan en la actualidad:

Dictáfono:

El dictáfono es una aplicación cuyo fin es la conversión de habla a texto a manera de dictado, sus restricciones son las siguientes: permite el reconocimiento de grandes vocabularios, admite voz como si se estuviera leyendo. El sistema puede ser considerado monolocutor, y las restricciones relativas al ruido son bastante severas. En la figura 1.3 se muestra el diagrama de las restricciones del dictáfono.

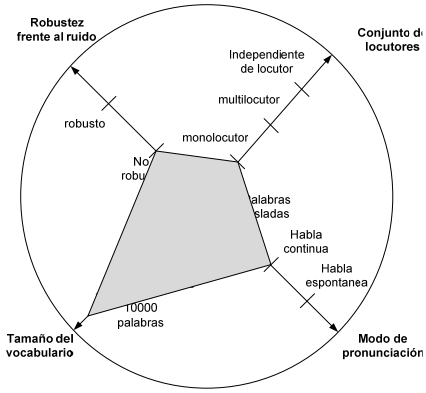


Figura 1.3: Diagrama de Kiviat de las restricciones de un dictáfono

Aplicación telefónica:

Las aplicaciones telefónicas basadas en reconocimiento automático de voz se encuentran en el extremo opuesto. El objetivo es proporcionar un servicio a un conjunto amplio de locutores en condiciones acústicas frecuentemente desfavorables. Aunque hacen uso de técnicas de reconocimiento de habla continua, los vocabularios son relativamente reducidos y las gramáticas utilizadas son muy restrictivas. En la figura 1.4 se muestra el diagrama de las restricciones de las aplicaciones telefónicas.

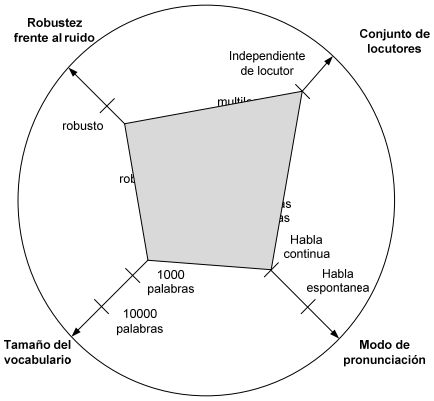


Figura 1.4: Diagrama de Kiviat de las restricciones de las aplicaciones telefónicas.

En la actualidad no es posible diseñar reconocedores eficientes que eviten simultáneamente todas las restricciones. En este sentido, el reconocimiento humano lleva una enorme ventaja sobre el reconocimiento automático. En la figura 1.4 se muestra el diagrama de Kiviat del reconocimiento humano.

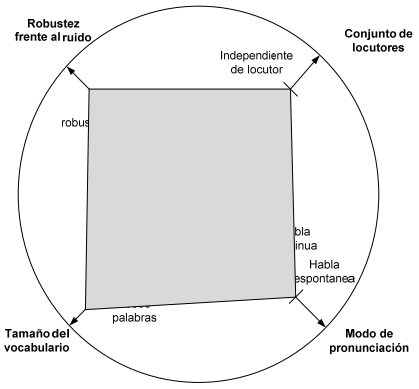


Figura 1.5 Diagrama de Kiviat del reconocimiento humano

1.2.4 Estructura

Aunque los sistemas de reconocimiento posean diferentes restricciones poseen en general una misma estructura que puede ser modelada mediante un diagrama de bloques. En la figura 1.5 se muestra la estructura simplificada de los sistemas de reconocimiento del habla.

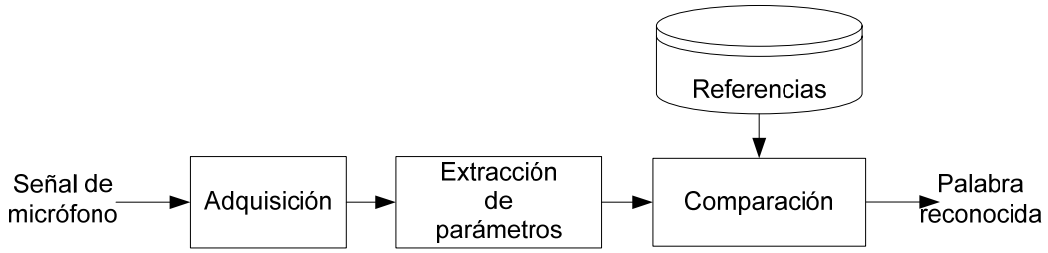


Figura 1.6: Estructura general de un sistema de reconocimiento de voz.

A continuación se definen la función de cada bloque del sistema de reconocimiento del habla.

Adquisición:

Realiza las operaciones necesarias para obtener la señal digital de voz: la onda acústica es convertida en una señal eléctrica analógica que es amplificada, filtrada, muestreada, codificada y realiza la detección de actividad de voz. para finalmente entregar a su salida solo la señal de voz.

Parametrización o de representación:

Este bloque tiene por objetivo representar la señal de voz de una forma adecuada para el reconocimiento. A la salida de este bloque la señal queda usualmente representada como una secuencia de vectores, cada uno de los cuales contiene básicamente información espectral correspondiente a un segmento de voz.

Comparación

Los sistemas RAH disponen de un conjunto de referencias que dependiendo de la arquitectura del reconocedor pueden ser parámetros extraídos de fonemas, silabas o palabras.

El bloque de comparación tiene como objetivo contrastar las referencias con la señal de habla a reconocer, generando así a su salida un texto que contiene el nombre de la referencia más probable (4).

1.2.5 Técnicas aplicadas al reconocimiento del habla

Las técnicas aplicadas al reconocimiento del habla se refieren específicamente al algoritmo presente en la etapa de comparación. Estos algoritmos gobiernan las etapas internas de la comparación que son el entrenamiento (generación de

referencias) y reconocimiento (comparación). A continuación se dará una breve introducción sobre los tres algoritmos más utilizados en la actualidad.

Alineamiento temporal dinámico

La técnica de alineamiento temporal dinámico o DTW (Dynamic time warping en ingles) es uno de los algoritmos aplicados en los sistema de reconocimiento automático del habla más antiguos e importantes. Si bien esta técnica aun se continúa utilizando, tiene un número de limitaciones que restringen su uso a sistemas de vocabularios pequeños ya que en sistemas de mayor tamaño, el número de plantillas a generar y el coste computacional de las búsquedas son intratables.

Consiste en comparar el patrón de entrada a reconocer con la serie de referencias almacenadas, la comparación incluye una alineación temporal no lineal. Lo cual produce una deformación del eje de tiempo de las señales de referencia almacenadas, con el objetivo de asemejarlas a la señal a reconocer, de esta manera la señal de referencia que haya experimentado la menor deformación temporal será la más probable a ser escogida como reconocida.

Modelos ocultos de Markov

Los modelos ocultos de Markov o HMM (Hidden Marcov models), representan otro enfoque alternativo, que es el de adoptar un modelo estadístico para cada una de las palabras del vocabulario de reconocimiento, proporcionando los mejores resultados hasta la fecha tanto para el reconocimiento de habla aislada como continua y para independencia del locutor.

Utiliza un algoritmo de alineamiento no lineal conocido como Algoritmo de Viterbi, capaz de alinear la secuencia de vectores de entrada con el conjunto de referencias estadísticas (HMM) que representan las palabras del diccionario, en forma de la probabilidad de que esa secuencia observada sea observada o generada por los distintos modelos ocultos de Markov.

Redes Neuronales

Los modelos computacionales basados en redes neuronales surgieron hace relativamente bastante tiempo, pero se abandonó su estudio por no disponer de algoritmos eficientes de entrenamiento. Ahora ya no existe esa dificultad, y se ha demostrado ampliamente su enorme potencia computacional.

Los sistemas de reconocimientos basados en redes neuronales pretenden, interconectando un conjunto de unidades de proceso o neuronas en paralelo de forma similar que en la mente humana, obtener prestaciones de reconocimiento similares a las humanas, tanto en tiempo de respuesta como en tasa de error (5).

1.3 Alcance del proyecto

En la tabla 1 se muestra las restricciones y la técnica aplicada en la comparación del sistema de reconocimiento automático del habla que será implementado en matlab.

Tabla I: Restricciones del sistema RAH a implementar

Restricciones	***Sistema a desarrollar***
Usuarios	Monolocutor
Ruido	No robusto
Forma de pronunciación	Palabras aisladas
Vocabulario	Pequeño
Técnica aplicada para comparación	Alineamiento temporal dinámico (DTW)

La selección del alineamiento temporal dinámico (DTW) como técnica aplicada en la comparación del sistema a implementar es debido a que el DTW tiene un gran desempeño sobre sistema de las mismas restricciones. Las restricciones del sistema a implementar también son graficadas en la figura 1.6 mediante un diagrama de Kiviat.

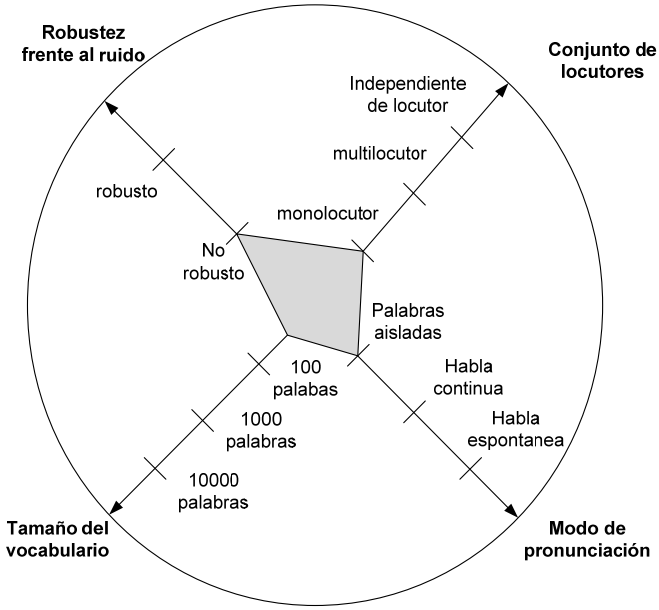


Figura 1.7: Diagrama de Kiviat de las características del reconocedor de voz implementado en esta tesis.

En el caso del tamaño del vocabulario que puede reconocer el sistema RAH depende los experimentos que se le realizan al sistema.

CAPÍTULO 2

2. MODELO DE PRODUCCIÓN DEL HABLA

El sistema de reconocimiento del habla requiere de la extracción de información proveniente de la señal de habla, por lo que es necesario comprender como esta se produce. En este capítulo se analiza al aparato fonador con el fin de construir un modelo digital que permita nos conocer la información que contiene una señal de habla, para que posteriormente sea extraída o eliminada alguna información redundante o innecesaria.

2.1 Análisis de la producción de habla

El aparato fonador es el conjunto de órganos del cuerpo humano, que tiene como objetivo la generación de sonidos para la comunicación. Estos sonidos por si solos carecen de significado, aunque si estos son utilizados de manera adecuada por el ser humano pueden expresar pensamientos o ideas. Estas ideas al ser manifestadas de manera acústica se las conoce como habla.

Si cogemos dos hojas de papel, las aproximamos, de tal manera que estén juntas y soplamos a través de ellas, conseguiremos producir un movimiento de ambas hojas y con ello un sonido (6). Pues bien, Explicando de manera sencilla, se puede definir que la producción y emisión de los sonidos por el aparato fonador, son consecuencia de un flujo de aire proveniente de los pulmones, que al pasar por las cuerdas vocales, que inicialmente se encuentran cerradas, son obligadas a moverse y así producir una vibración, estas vibraciones generan un sonido que atraviesa las cavidades superiores, las cuales modulan formando sonidos resultantes, siendo nuestro cerebro y sistema de percepción capaces de interpretar estos sonidos como palabras de un lenguaje definido.

En la figura 2.1 se muestra la sección sagital de la parte superior del cuerpo de un ser humano, en el cual se observan los órganos del aparato fonador.

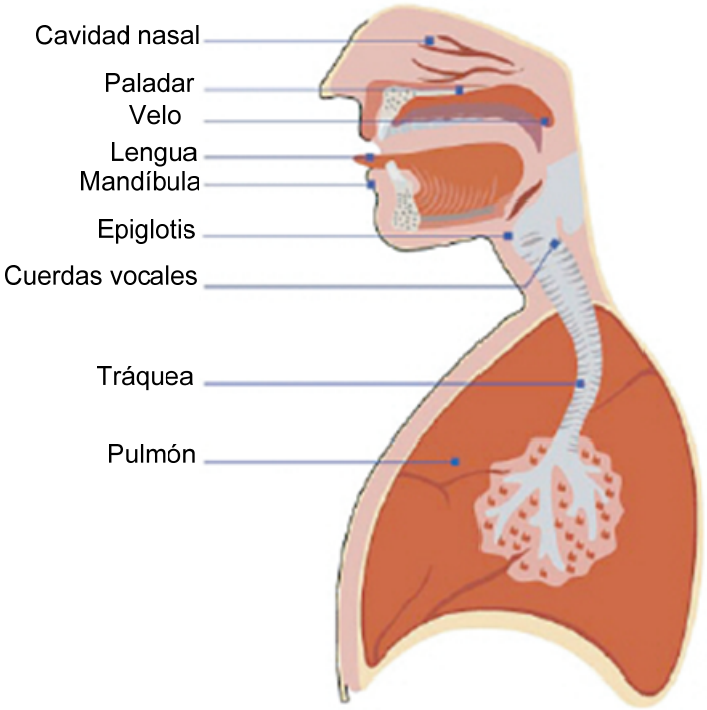


Figura 2.1: Aparato fonador del ser humano (7).

2.1.1 Partes principales del aparato fonador

Los órganos que conforman al aparato fonador se pueden dividir en tres grupos de órganos diferenciados por la función que ellos desempeñan en la producción de sonidos:

- Órganos de respiración
- Órganos de fonación
- Órganos de articulación

2.1.2 Órganos de respiración

Los órganos de respiración están principalmente representados por los pulmones, la tráquea y el diafragma. El proceso de respiración es el combustible para la generación de sonidos ya que suministra aire a los pulmones para que luego estos provean de manera controlada el flujo de aire a través de la tráquea. Los valores de presión del aire expulsado por los pulmones deben de ser del orden de 4 cm H_2O para sonidos muy suaves hasta aproximadamente de 20 cm H_2O para sonidos muy fuertes y de altas frecuencias (8). En la figura 2.2 se muestran los órganos de respiración.

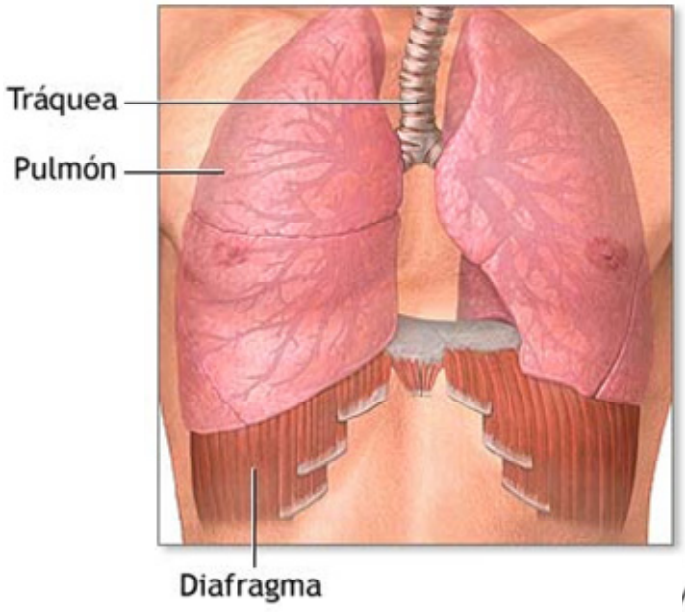


Figura 2.2 Vista frontal de los pulmones [9].

A continuación se describe con mayor detalle a los órganos que componen el sistema respiratorio:

Tráquea: Es un tubo semiflexible, constituido por anillos cartilaginosos, que en su parte inferior se conecta a los pulmones como bronquios y en la parte superior se enlaza con la laringe.

Pulmones: Los pulmones son una masa esponjosa de una larga área. Su capacidad es de 4 a 5 litros en un adulto. Están contenidos en una cámara de aire llamada pleura, la misma que está contenida a su vez lateralmente por las costillas e inferiormente por el diafragma.

Diafragma: El diafragma es un músculo en forma de domo, ubicado en la parte inferior de las costillas y separa al tórax del abdomen

2.1.3 Funcionamiento del ciclo respiratorio

El ciclo respiratorio tiene dos fases, la primera se produce cuando el diafragma se contrae y desciende el domo del diafragma, por consecuencia el aire entra a los pulmones, esta parte del ciclo respiratorio se denomina “inhalación”. Una vez que los pulmones se han llenado, comienza la parte del ciclo respiratorio

correspondiente a la “exhalación”. Al salir el aire de los pulmones, el diafragma se eleva y se relaja. En la figura 2.3 se muestran las dos fases antes mencionadas.

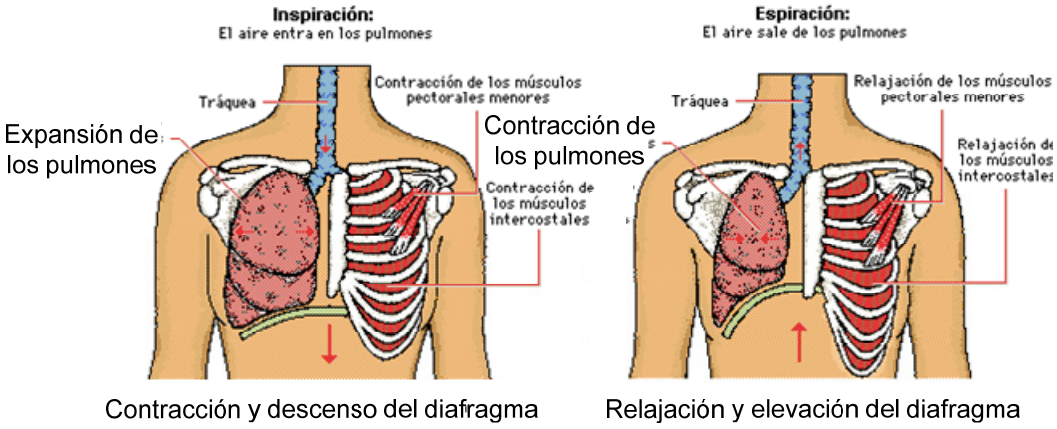


Figura 2.3: fase de inhalación y exhalación (10),

La importancia de este ciclo respiratorio para la generación de voz es porque la voz o habla se produce en el cambio de fase de inhalación a exhalación y generalmente consiste inhalaciones largas o cortas, así como exhalaciones controladas.

2.1.4 Órganos de fonación

Los órganos de fonación, están situados en la parte superior de la tráquea e inferior a la laringe. Está formada por tres cartílagos (cricoides, tiroides y

aritenoides), por un conjunto de músculos, y por las cuerdas vocales. Los dos primeros contienen y controlan a las cuerdas vocales. Éstas últimas constituyen la fuente de generación de sonidos, también cierran la tráquea para proteger el tracto pulmonar de objetos y permitir la formación de presión dentro del tórax y el abdomen. En la figura 2.4 se muestra mediante un corte horizontal los órganos de fonación.

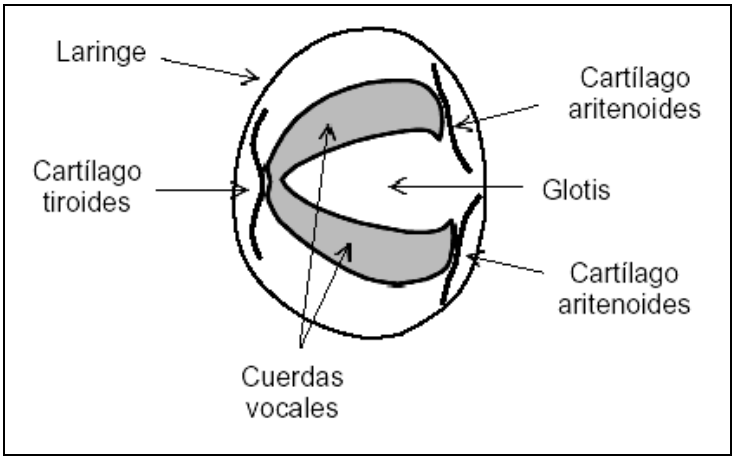


Figura 2.4: La laringe vista desde la perspectiva superior

Cuerdas vocales

Las cuerdas vocales son un tejido sólido con cuatro dobleces entre el frente y la parte posterior de la laringe. Cuando las partes terminales de las cuerdas están separadas, las cuerdas están abiertas, la cual es la posición para la respiración. Cuando las partes terminales están juntas, las cuerdas están cerradas y

proporcionan el sello al tracto pulmonar para la deglución. En la figura 2.5 se muestra la disposición de las cuerdas vocales.

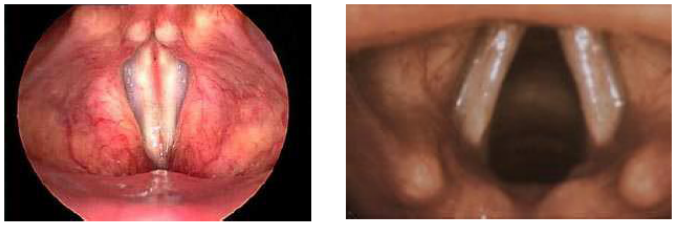


Figura 2.5: Posiciones de las cuerdas vocales “cerrada” y “abierta” respectivamente

2.1.5 Tipos de excitación

La producción de sonidos se debe a la interacción entre un flujo de aire y las cuerdas vocales, A esta interacción se le llama excitación, y es consecuencia del abrir y cerrar parcial o total de las cuerdas vocales de manera rápida y secuencial. Esta excitación puede ser de las siguientes formas:

- Fonación
- Susurro

Fonación

Este término se refiere a la oscilación de las cuerdas vocales por los movimientos de los cartílagos aritenoides. Cuando el aire es forzado a través de

las cuerdas vocales, éstas vibran. Dando a lugar a la apertura y cierre de las cuerdas vocales y seccionando el pulso de aire en pulsos cuasi-periódicos llamados pulsos glotales, estos pulsos poseen una frecuencia de vibración llamada frecuencia fundamental. Además su forma de onda es aproximadamente triangular y tiene un ciclo de trabajo del orden de 0.3 a 0.7; a como consecuencia de su forma, las altas frecuencias disminuyen su amplitud a 12 dB/octava. Por lo que su naturaleza paso-bajo proporciona un espectro con una fuerte frecuencia fundamental y progresivamente más débiles armónicas.

Los sonidos que resultan de la fonación se los llama como sonidos sonoros. En la figura 2.6 se muestra un modelo de Rosenberg que se asemeja a la onda de pulso glotal.

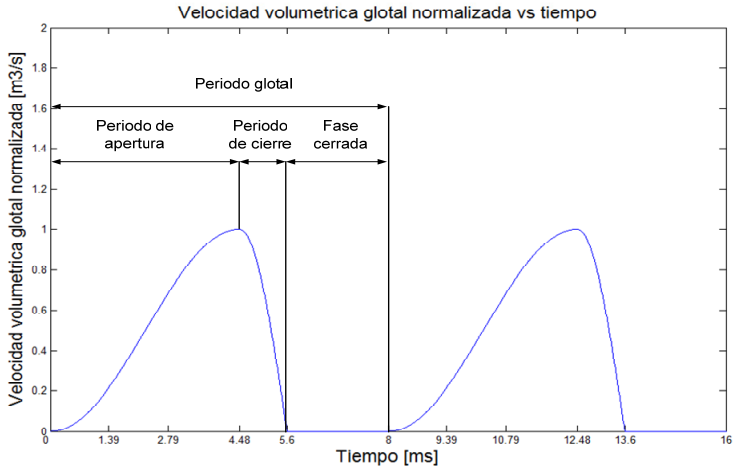


Figura 2.6: Forma de onda de un periodo de pulso glotal

En la figura 2.6 se observa un periodo de la señal de pulso glotal normalizada, donde se puede destacar que el periodo total de esta onda es de 8 ms por lo que la frecuencia o tono es 125 Hz, y su ciclo de trabajo es de 0.7.

Susurro

Los susurros se forman cuando las cuerdas vocales están juntas por el cartílago aritenoides, pero en lugar de sellar completamente la glotis existe una pequeña abertura triangular entre estos cartílagos. El aire al a través de esta apertura genera turbulencias, que ocasionan ruido de banda ancha, por lo que esta excitación puede ser modelado como un ruido con distribución gamma.

Los susurros son más débiles que las fonaciones dado que implican un menor volumen de aire, y poseen mayor energía en altas frecuencias. Los sonidos resultantes del susurro se denominan sordos.

2.1.6 Órganos de articulación

Los órganos de articulación también llamados tracto vocal, están representados por la laringe, la cavidad bucal y la cavidad nasal. Su misión fundamental en la producción de sonidos, es perturbar adecuadamente el flujo de aire proveniente

de las cuerdas vocales, para dar lugar a la señal acústica expulsada a la salida de la boca y la nariz. En la figura 2.7 se muestran los órganos de articulación.

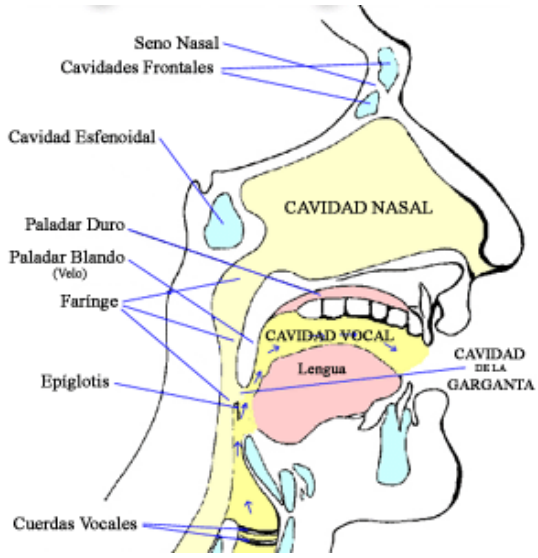


Figura 2.7: Órganos de articulación o tracto vocal [9].

A continuación se describe con mayor detalle los tres órganos antes mencionados:

La faringe

La faringe es un tubo membranoso formado por paredes musculares flexibles, su función es conectar la laringe con las cavidades bucal y nasal, suele dividirse en tres partes: faringe laríngea, faringe bucal y faringe nasal, las dos últimas

están separadas por el velo del paladar. La faringe es poco versátil en lo que se refiere a la modificación de su forma y su tamaño.

Cavidad nasal

La cavidad nasal está formada por las fosas nasales que son dos canales paralelos divididos por el tabique nasal, su función es servir como cámara de resonancia cuando el velo del paladar o paladar blando esta despegado de la pared faríngea dando paso al flujo de aire a través de esta cavidad. La cavidad nasal no puede modificar su forma o tamaño.

Cavidad bucal

Se extiende desde la faringe bucal hasta los labios, es la cavidad más versátil en lo que se refiere a la modificación de su forma y tamaño dado que posee elementos llamados articuladores, que modifican la forma interna de la cavidad bucal. En el capítulo 3 se describirán con mayor detalle la función de los articuladores.

En la figura 2.7 se muestran los articuladores de la cavidad bucal.

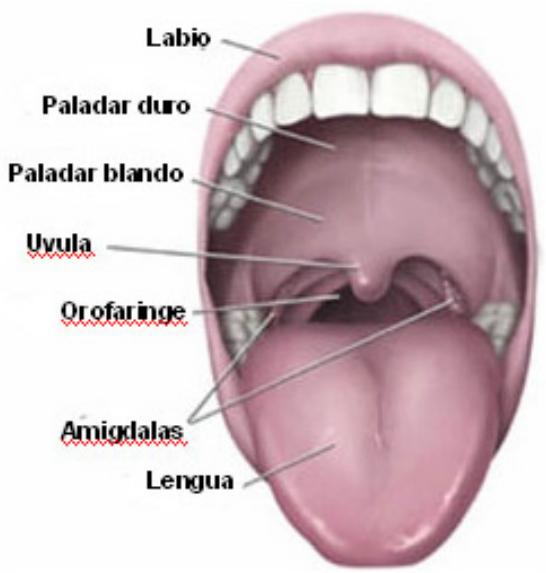


Figura 2.8: articuladores de la cavidad vocal

2.1.7 Funciones de los articuladores

El tracto vocal realiza dos funciones importantes en la producción de sonidos y son:

- Resonancia
- Radiación

Resonancia

La resonancia se define como el aumento de la amplitud de frecuencias alrededor de una banda. Este fenómeno se manifiesta sobre una onda sonora cuando esta atraviesa un tubo acústico.

El tracto vocal es considerado un tubo acústico de sección no uniforme que varia con el tiempo, esta variación de sección se debe a los movimientos de los articuladores en la cavidad bucal. En la figura 2.9 se muestra la sección del tracto vocal y la grafica del área del tracto vocal en función de su distancia a las cuerdas vocales, al pronunciar la vocal /i/.

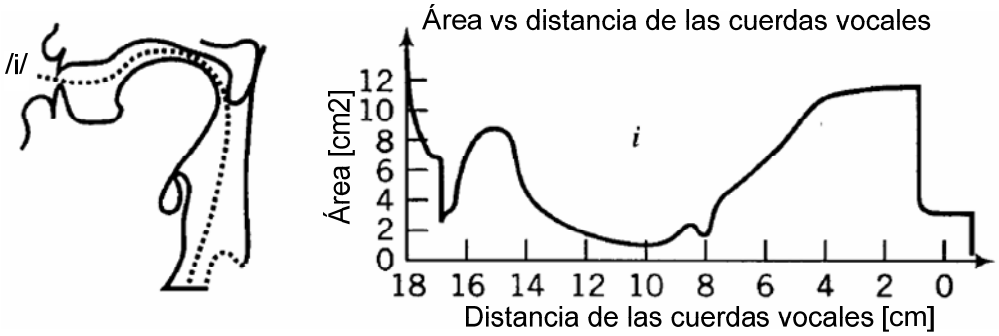


Figura 2.9: Sección del tracto vocal y la grafica del área del tracto vocal en función de su distancia a las cuerdas vocales, al pronunciar la vocal /i/ [8].

Para comprender el fenómeno de resonancia que existe en el tracto vocal es necesario ir al caso más simple, es decir considerar al tracto vocal como un tubo acústico simple o de sección constante. En este caso la frecuencia de resonancia está relacionada con la longitud del tubo por la ecuación 2.2

$$f = \frac{c_0}{4L}(2n+1), \quad n = 0,1,2,\ldots \tag{2.1}$$

.

Donde Co corresponde a la velocidad del sonido en el medio, L la longitud del tubo acústico, f la frecuencia de resonancia, y n corresponde a todos los valores posibles para evaluar en la función.

La ecuación 2.1 da a entender que existen varias resonancias y ocurren en múltiplos impares de la frecuencia de resonancia fundamental, $C_o/4L$. En el caso de que el tracto vocal fuese un tubo acústico simple, las frecuencias de resonancia aparecerían, en aproximadamente, 500 Hz, 1500 Hz, 2500 Hz, etc. ya que su longitud típica es de unos 17,5 cm. (valor típico de un hombre adulto). En la figura 2.9 se muestra las tres primeras resonancias de un tubo acústico simple de 17 cm [33].

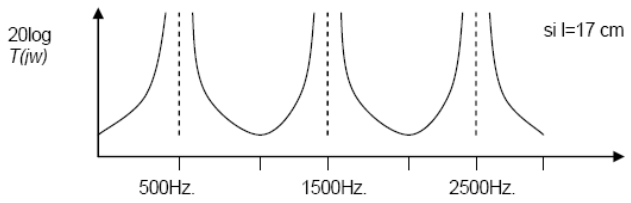


Figura 2.10: Resonancias del modelo de tubo acústico simple [11].

En el caso del tracto vocal real es necesario considerar su sección no uniforme como una concatenación de tubos simples de diferentes secciones. En la figura

2.11 se muestra al tracto vocal como una concatenación de tubos uniforme, exceptuando la cavidad nasal.

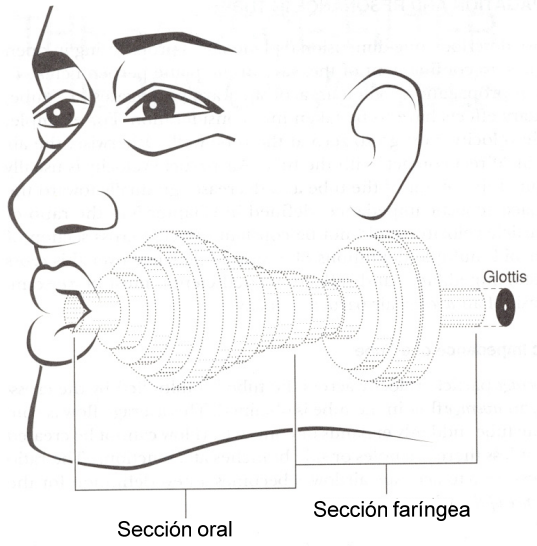


Figura 2.11: modelo del tracto vocal, exceptuando la cavidad nasal.

En el análisis de modelos de varios tubos de secciones distintas concatenados, aparece un nuevo concepto llamado coeficientes de reflexión, El cual define que en la conexión de los tubos hay una porción de la onda que se transmite y otra que se refleja, de tal manera que no se puede aplicar la ecuación 2.1 para hallar de manera directa sus resonancias, sino que hay que realizar análisis en las fronteras entre dos tubos. Según observaciones experimentales sobre las características espectrales de la señal de voz, muestran la presencia de múltiples resonancias relacionadas con la configuración del tracto vocal. En la

figura 2.12 se muestran las configuraciones del tracto vocal y su respuesta de frecuencia para las vocales /o/, /i/ y /u/ respectivamente.

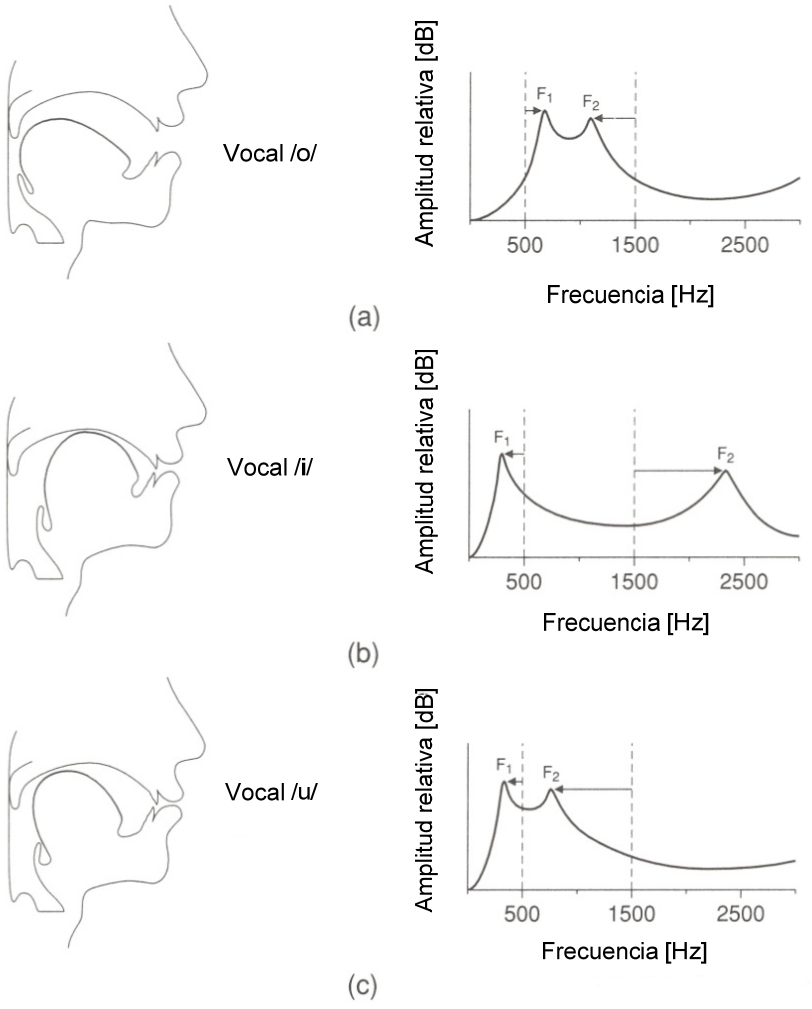


Figura 2.12: Configuraciones del tracto vocal en la pronunciación de vocales /o/, /i/, /u/ y resonancias resultantes.

A los máximos espectrales generados por las resonancias del tracto vocal se les llama **formantes**, y están en posiciones determinadas para cada uno de los sonidos que podemos producir, aunque hay una variabilidad inherente en la producción de los mismos por parte de cada locutor.

Radiación

La salida del sonido tiene asociada una impedancia de radiación que influye sobre el sonido variando la composición espectral del sonido y los niveles de presión sonora con respecto al ángulo de salida.

2.2 Modelo digital del habla

Un modelo del habla que se considere completo debe incluir: los cambios en la señal de excitación, la respuesta del tracto vocal y los efectos de los labios en la radiación.

Tal modelo es **fuente-filtro** que ha sido usado por casi todos los sistemas de procesamiento de voz. En la figura, 2.13 se muestra al modelo fuente filtro

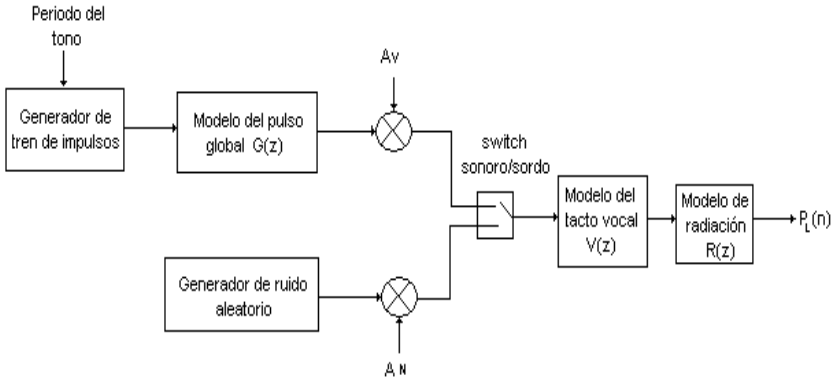


Figura 2.13 Modelo de producción de voz fuente – filtro

Según la figura 2.13 el modelo puede describirse en términos de la transformada Z, mediante las siguientes ecuaciones 2.2 para el caso de excitación sonora y 2.3 excitación sorda.

$$P_L(z) = R(z)\,V(z)\,G(z)\,X(z) \qquad (2.2)$$

$$P_L(z) = R(z)\,V(z)\,N(z) \qquad (2.3)$$

Donde X (z) y N (z) son las transformadas Z de las secuencias discretas x(n) y n(n), resultantes de muestrear x (t) representado por tren de impulsos y n (t) por ruido aleatorio. y G (z), V(z) y R(z) son las funciones de transferencia de los sistemas discretos que modelan los efectos de la glotis, el tracto vocal y los labios, respectivamente.

A continuación se describen los bloques que forman parte del modelo digital del habla:

- Excitación glotal
- Modelo del tracto vocal
- Modelo de radiación

2.2.1 Modelo de excitación

La excitación glotal consiste en la producción de ondas glotales a partir de la expulsión del aire a través de la abertura y cierre de las cuerdas vocales provocando como resultado dos tipos de excitación señales:

- Periódicas (excitación sonora)
- Ruido turbulento (excitación sorda)

Excitación sonora

El espectro glotal esta modelado como un tren de impulsos espaciados a frecuencias iguales a la frecuencia del tono fundamental que entran al modelo del pulso glotal que es un filtro de aproximadamente, una caída de 12 dB/octava alrededor de 0.8 a 1 KHz. En la figura 2.14 se muestra las representaciones en tiempo y frecuencia de: (a) tren de impulsos espaciados con un periodo de

0.008 ms (f = 125 Hz), (b) Modelo de pulsos glotal (Modelo de Rosenberg) y (c) excitación sonora (salida del modelo de pulso glotal).

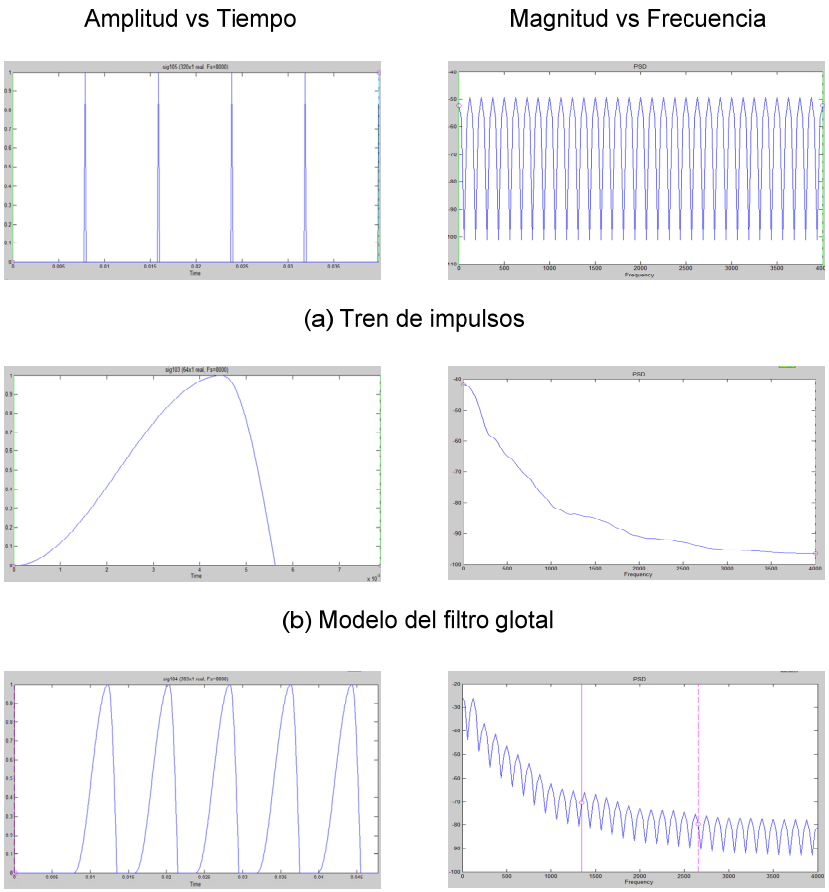


Figura 2.14: Representación temporal y espectral de: (a) tren de impulsos, (b) Modelo del filtro glotal y (c) salida del filtro.

En la figura 2.14 se puede notar que la salida del sistema de excitación sonora consiste de frecuencias armónicas espaciadas por el inverso del periodo de los

impulsos. También se nota la caída del espectro a medida que aumenta la frecuencia.

Excitación sorda

La excitación sorda modelada como una señal aleatoria con distribución gamma o laplaciana y espectro plano. En la figura 2.15 se muestra las representaciones de tiempo y frecuencia de una excitación sorda

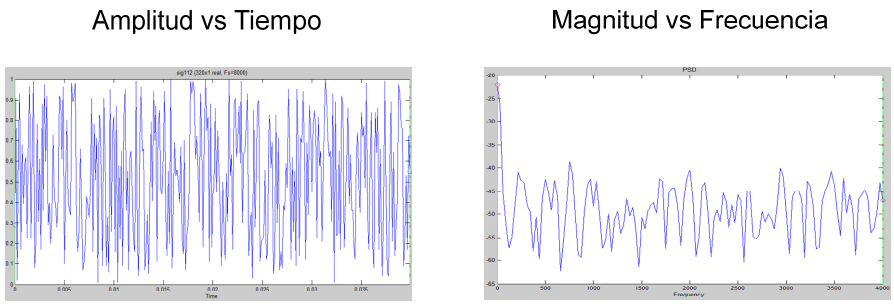


Figura 2.15: Representación en tiempo y frecuencia de una excitación sorda

2.2.2 Modelo del tracto vocal

El tracto vocal puede ser modelado por la función de transferencia:

$$H(z) = \frac{G}{1-\sum_{i=1}^{N} \propto_i z^{-i}} \tag{2.4}$$

Donde G representa el factor de ganancia total y α_i las ubicaciones de los polos. Los polos de H (z) corresponden a las resonancias o formantes de la voz. En la figura 2.16 se muestra un diagrama de polos y ceros donde se disponen los polos y ceros de un modelo de tracto vocal.

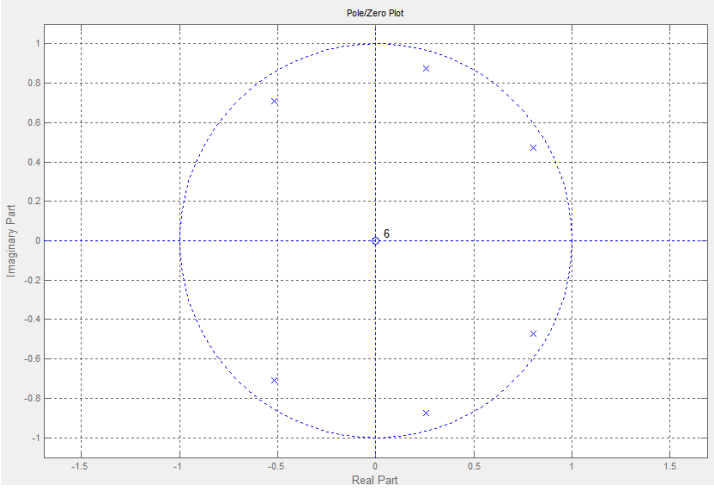


Figura 2.16: Polos y ceros un modelo del tracto vocal.

En la práctica, en la mayoría de las aplicaciones se modela como un sistema todo-polos formado por una cascada de un pequeño número de resonadores de dos polos. Cada resonancia se define como un formante con su frecuencia central y su ancho de banda correspondientes. La razón fundamental por la que se utiliza un modelado todo-polos es que un modelo todo-polos permite aproximar cualquier modelo racional utilizando un número suficientemente elevado de polos. En la figura 2.17 se muestra la representación en tiempo y

frecuencia de: (a) pulsos glotales, (b) modelo del tracto vocal y (c) salida del tracto vocal.

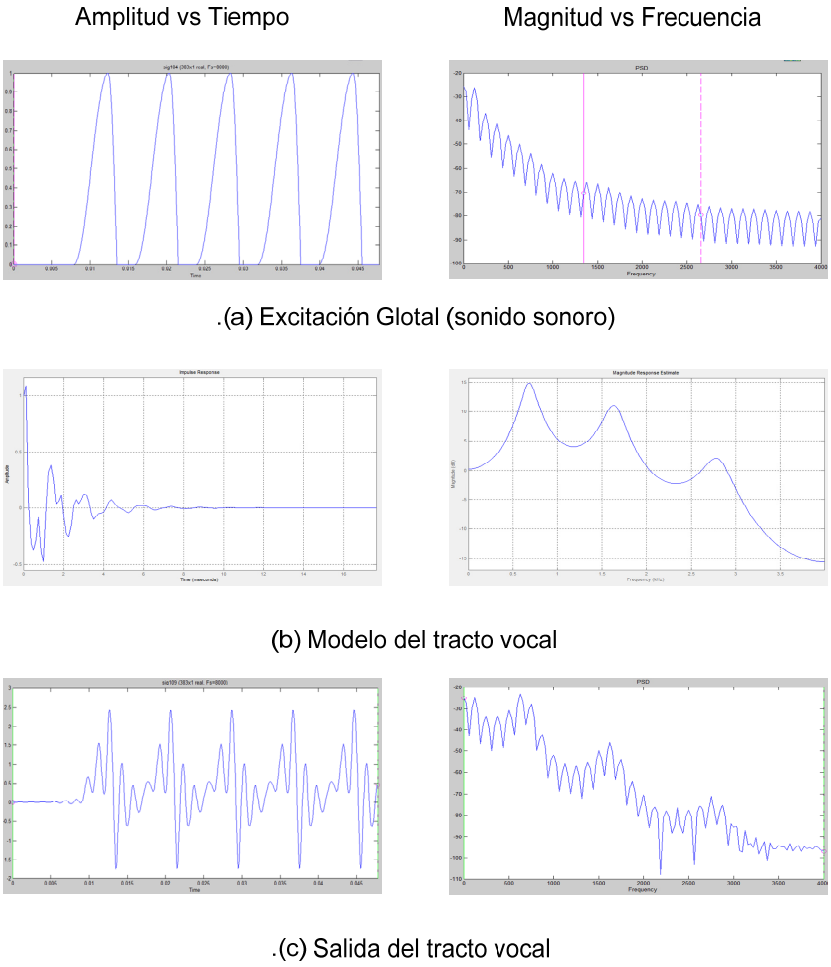
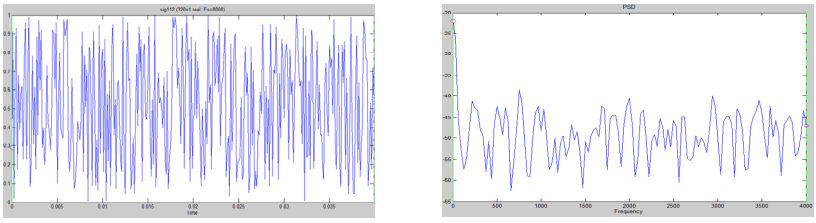


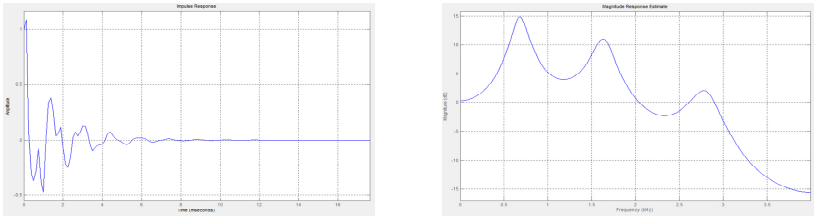
Figura 2.17: Representación en tiempo y frecuencia de (a) pulsos glotales, (b) modelo del tracto vocal y (c) salida del tracto vocal.

En la figura 2.17 se muestra el proceso de resonancia realizado por el tracto vocal sobre una excitación sonora. El modelo del filtro se basa en un filtro construido por medio de los polos y ceros de la figura 2.16, donde existen 3 pares de polos. El mismo filtro se aplica a una excitación sorda en la figura 2.18.

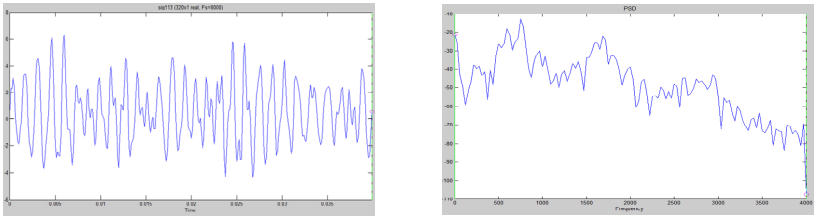
Amplitud vs Tiempo | Magnitud vs Frecuencia



.(a) Excitación de susurro (sonido sordo)



(b) Modelo del tracto vocal



.(c) Salida del tracto vocal

Figura 2.18: Representación en tiempo y frecuencia de (a) excitación sorda, (b) modelo del tracto vocal y (c) salida del tracto vocal.

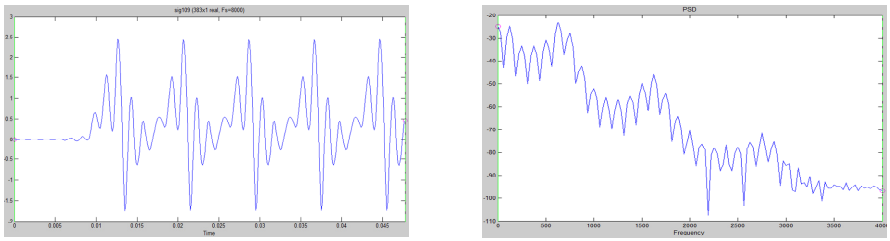
2.2.3 Modelo de radiación

El modelo de radiación describe la impedancia de radiación vista por la presión de aire cuando abandona los labios. Este efecto tiene la propiedad que a bajas frecuencias la presión del sonido es proporcional a la derivada de la velocidad volumétrica. Esto introduce un levantamiento de 6 dB/octava en el espectro que puede ser modelado por la ecuación 2.5

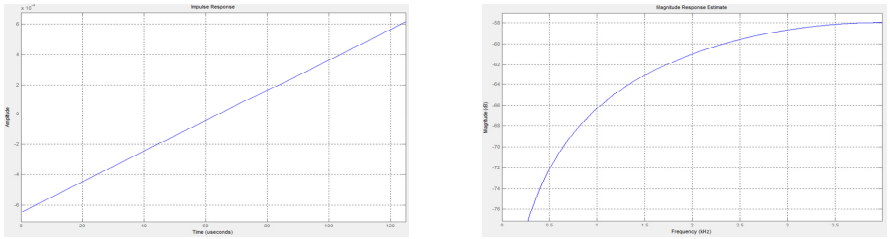
$$R(z) = 1 - z^{-1} \tag{2.5}$$

En la figura 2.19 se muestra la representación en tiempo y frecuencia de: (a) salida del tracto vocal (sonido sonoro), (b) modelo de radiación y (c) del modelo de radiación.

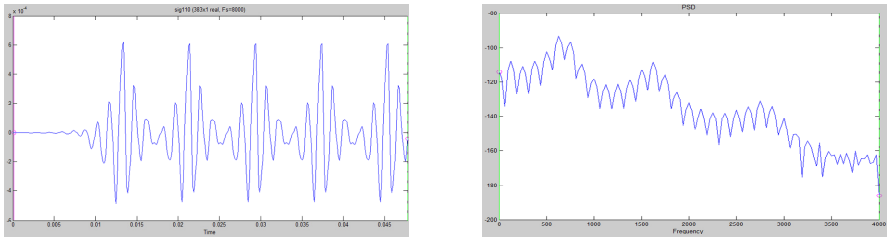
Amplitud vs Tiempo | Magnitud vs Frecuencia



(a) Salida del tracto vocal (sonido sonoro)



(a) Modelo de radiación

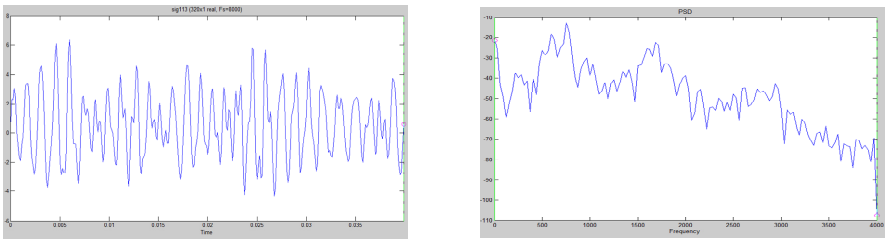


.(c) Señal de habla (sonido sonoro)

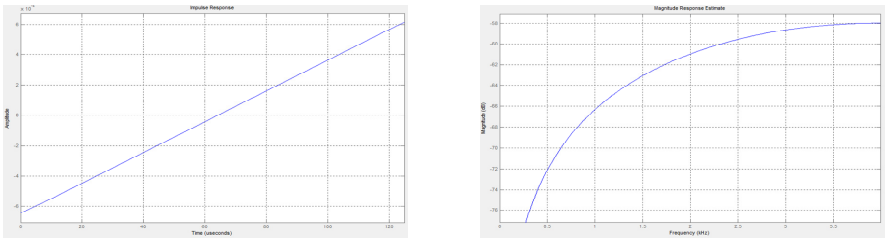
Figura 2.19: Representación en tiempo y frecuencia de (a) salida del tracto vocal (sonido sonoro), (b) modelo de radiación y (c) salida del modelo de radiación.

En la figura 2.20 se muestra a la salida del modelo de radiacion como una atenuación de bajas frecuencias. En la figura 2.20 se muestra la aplicación del mismo modelo de radiación para un sonido sordo que es salida del tracto vocal.

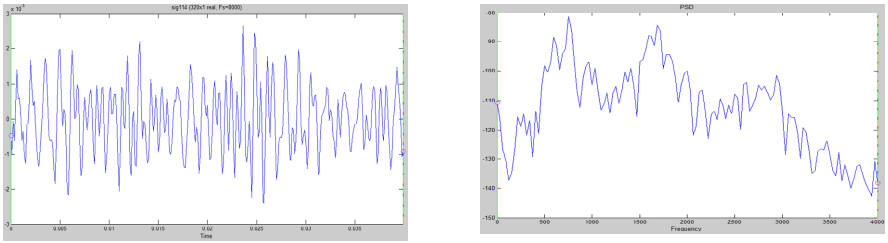
Amplitud vs Tiempo | Magnitud vs Frecuencia



(a) Salida del tracto vocal (sonido sordo)



(b) Modelo de radiación



(b) Señal de habla (sonido sordo)

Figura 2.20: Representación en tiempo y frecuencia de (a) salida del tracto vocal (sonido sordo), (b) modelo de radiación y (c) salida del modelo de radiación.

CAPÍTULO 3

3. ANÁLISIS DE LA SEÑAL DE HABLA

En el capítulo dos se presenta un modelo digital de producción de habla, este modelo establece que todos los sonidos del habla son generados a partir de la interacción de la una excitación y el tracto vocal. Este capítulo pretende analizar los sonidos producidos por el aparato fonador, mediante el análisis articulatorio y acústico de los mismos, con la finalidad de observar información importante que pueda ser extraída.

3.1 Principios básicos sobre fonética

Fonética es la rama de la lingüística que estudia la producción de los sonidos de la lengua, en sus manifestaciones físicas, sin importar el significado o el uso de estos en la lengua.

Los sonidos de la lengua pueden ser representados por un conjunto de unidades simples llamadas fonemas. Los fonemas son también definidos como unidades mínimas de distinción ya que un fonema posee información que lo hace único con respecto a los demás fonemas, esta información puede ser extraída mediante la ayuda de las siguientes ramas de la fonética:

- Fonética articulatoria.
- Fonética acústica.

3.2 Análisis Articulatorio de la señal habla

La fonética Articulatoria estudia los sonidos de una lengua desde el punto de vista fisiológico, es decir, describe que órganos vocales intervienen en su producción, en qué posición se encuentran y como esas posiciones varían los distintos caminos que puede seguir el aire cuando sale por la boca, nariz, o garganta, para que se produzcan sonidos diferentes.

Los símbolos fonéticos se usan para representar de manera escrita cualquier fonema, los símbolos que se usan más frecuentemente son adoptados por la asociación fonética internacional (AFI) en el alfabeto internacional donde escriben a cada fonema con un símbolo asociado. En la tabla II se observan los diferentes fonemas que se presentan en el idioma español representados por los símbolos de la AFI.

Tabla II: Fonemas del alfabeto español

Fonema	**Letras**	**Ejemplos**	**Fonema**	**Letras**	**Ejemplos**
/a/	A	Cuatro	/n/	N	Nada
/e/	E	Seis	/l/	L	Lado
/i/	I	Cinco	/λ/	Ll	Llama
/o/	O	Dos	/m/	M	Mano
/u/	U	Uno	/ɳ/	Ñ	Niño
/b/	b	Bola	/p/	P	Ópera
	V	Vaso	/r/	R	Trozo
/c/	Ch	Muchacho	/ r̄ /	R	Radio
/d/	D	Donde		Rr	Carro
/f/	F	Café	/s/	x, z, s, c	Mesa
/g/	G	Guerra	/t/	T	Tela
	W	Huevo	/j/	X	Xerox
/x/	j, x	Caja	/y/	Y	Mayo
/k/	c, x	Casa	/θ/	Z	Caza
	K	Kilo			
	Q	Queso			

3.2.1 Clasificación de los fonemas según obstrucción

Dado que algunos fonemas tienen características en común, podemos agruparlos a estos fonemas en clases. Una división fundamental de los fonemas es la clasificación entre vocales y consonantes (vocálicas y consonánticas).

Los fonemas vocálicos son aquellos que se producen con ninguna constricción del tracto vocal. Por el contrario los fonemas consonánticos son aquellos que se producen con alguna constricción. En la figura 3.1 se muestra la clasificación de los fonemas desde el punto de vista de obstrucción del tracto vocal.

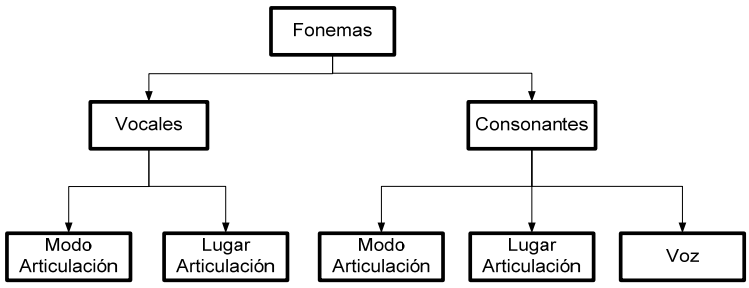


Figura 3.1: Clasificación de los fonemas del idioma español según obstrucción.

3.2.2 Las Vocales

Los fonemas vocálicos corresponden a las cinco vocales del alfabeto. Todos son articulaciones abiertas de corta duración, completamente sonoras, ya sea que

estén acentuadas o no. En la Figura 3.2 se muestra la configuración del tracto vocal en la producción de las vocales.

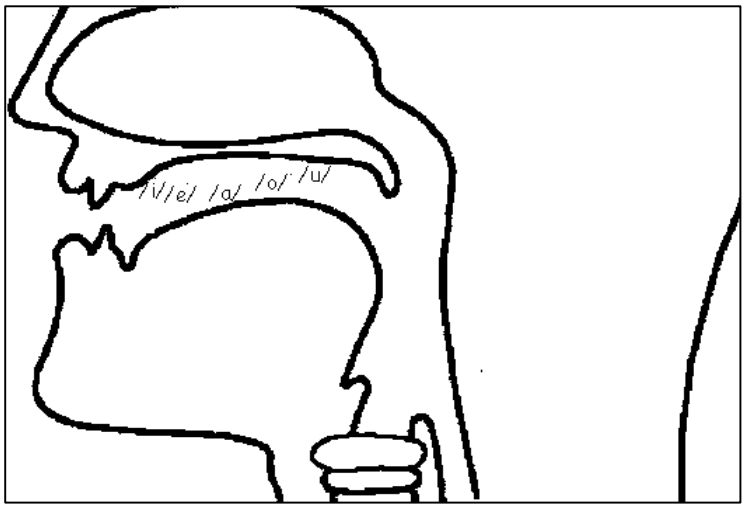


Figura 3.2 Configuraciones articulatorias de los fonemas vocálicos

Las vocales pueden ser clasificadas según:

- Modo de articulación
- Lugar de articulación

Modo de articulación

Las vocales según el modo de articulación se clasifican en relación a la abertura entre la lengua y el paladar y se denominan como:

- **Altas:** Se caracterizan por un movimiento de elevación de la lengua hacia el techo de la boca dejando una abertura estrecha por donde fluye el aire.
- **Medias:** Se caracterizan por un movimiento de elevación de la lengua hacia el techo de la boca dejando una abertura más amplia que la de las vocales altas.
- **Bajas:** El movimiento de elevación de la lengua es muy ligero de tal forma qu esta queda ocupando el hueco de la mandíbula inferior.

Lugar de articulación

Las vocales según el lugar de articulación se clasifican en relación a la posición de la lengua y se denominan como:

- **Anterior:** la lengua ocupa la región delantera o zona del paladar duro, como para articulación de las vocales.
- **Central:** la lengua ocupa la zona intermedia cubierta por el paladar medio, como para la articulación de la vocal a.
- **Posterior:** la lengua ocupa la región posterior o zona del paladar blando.

En la tabla III se muestra la clasificación de los fonemas vocálicos según su modo de articulación y lugar de articulación.

Tabla III: Clasificación de los fonemas vocálicos

	anterior	central	posterior
cerrada	i		u
media	e		o
abierta		a	

3.2.3 Consonantes

En la articulación de los fonemas consonánticos siempre está presente una obstrucción parcial o total al paso de la columna de aire, por su forma de ser producidas, las consonantes atienden a tres criterios

- Modo de articulación
- Tipo de excitación
- Lugar de articulación

Modo de articulación

Las consonantes, dependiendo del modo de articulación se refieren a los grados de constricción del punto de articulación y la manera en que se exhala para el siguiente sonido. Sin embargo, es importante describir todas las categorías para las consonantes.

- **Africadas:** Existe un cierre inicial del tracto vocal seguido de una expiración gradual que produce turbulencia
- **Aspiradas:** El tracto vocal está cerrado inicialmente en el punto de articulación y se exhala aire antes del siguiente sonido.
- **Fricativas:** El tracto vocal está abierto parcialmente en el punto de articulación y el velo está cerrado. Se genera ruido en el punto de articulación.
- **Laterales:** El tracto vocal está cerrado en el punto de articulación
- **Nasales:** El tracto vocal está cerrado en el punto de articulación y el velo está abierto.
- **Plosivas:** El tracto vocal está cerrado en el punto de articulación, el pasaje nasal está cerrado, y existe una exhalación limpia y cortante.
- **Semivocales:** El tracto vocal está parcialmente abierto en el punto de articulación sin turbulencia.
- **Vibrato:** Existe una abertura y cierre oscilatorios en el punto de articulación, seguidos de una exhalación gradual que produce turbulencia.

Tipo de excitación

Todos los fonemas consonánticos que corresponden a semivocales y nasales se conocen colectivamente como **sonoros**. Los fonemas **sonoros** implican una excitación del tracto vocal solamente con pulsos cuasi-periódicos originados por

la vibración de las cuerdas vocales. En contraste, las restantes clases son excitadas fundamentalmente por excitación ruidosa en el tracto vocal y se denominan **sordas**.

Lugar de articulación

El lugar de articulación identifica diferencias en el tracto vocal de acuerdo al punto máximo de constricción en el tracto vocal y permiten diferenciar fonemas que tienen la misma forma de articulación:

- **Alveolar:** La punta de la lengua se acerca o toca la punta alveolar en el techo de la boca.
- **Dental:** La punta de la lengua hace contacto con la parte posterior del diente incisivo superior.
- **Glotal:** Los dobleces de las cuerdas se cierran o constriñen.
- **Labial:** Los labios se constriñen. Bilabial denota constricción en ambos labios, mientras que labiodental denota contacto del labio inferior con los dientes superiores.
- **Palatal:** El dorso de la lengua se constriñe en el paladar duro.
- **Velar:** El dorso de la lengua se aproxima al paladar suave.

En la tabla IV se muestra la clasificación de los fonemas consonánticos según su modo de articulación y lugar de articulación y voz o tipo de excitación (12).

Tabla IV: Fonemas consonánticos

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

3.3 Análisis acústico de la señal de habla

La fonética acústica es rama de la fonética que estudia a la onda sonora como la salida de un resonador cualquiera; esto significa que, equipara el sistema de fonación con cualquier otro sistema de emisión y reproducción de sonidos.

Objetivos de la fonética acústica

- Estudio de las propiedades físicas de los sonidos del habla, considerándolos como ondas sonoras

- Visualización de las propiedades derivadas de los movimientos que tienen lugar en el tracto vocal
- Explicación de la relación entre las ondas sonoras y el mensaje del que son portadoras
- Determinación de los elementos invariantes que permite percibir como equivalente sonidos que desde el punto de vista acústico son distintos

Para poder examinar de manera profunda a las señales emitidas por el aparato de producción de habla, es necesario tener herramientas que nos ayuden a observar las propiedades que posee un fonema determinado. Estas herramientas serán llamadas representaciones y expresan a la señal de habla sobre el dominio del tiempo y de frecuencia.

3.3.1 Análisis en el dominio del tiempo

Los análisis que se desarrollan en el dominio del tiempo son:

- Análisis oscilográfico
- Análisis de energía
- Análisis de tasa de cruces por cero

Análisis oscilográfico

El análisis oscilográfico se basa en el oscilograma, que consiste en la representación de la señal de habla, obtenida a partir de un transductor (micrófono). En el eje horizontal del oscilograma se describe al tiempo mientras que en el eje vertical se describe a la presión de aire. En la figura 3.3 se muestra un oscilograma de la señal de habla “hola”.

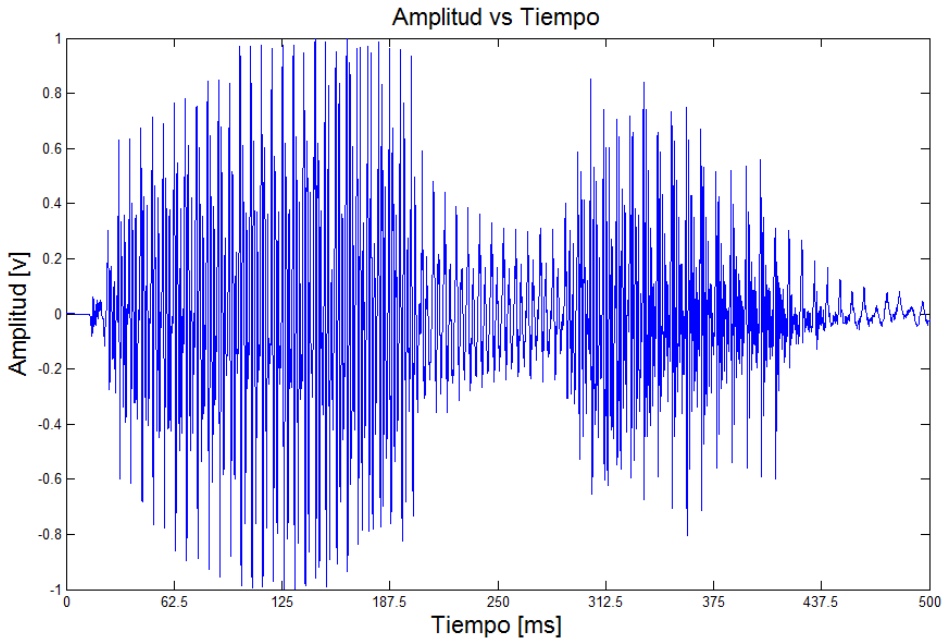


Figura 3.3: Oscilograma de la señal de habla “hola”.

Cabe destacar que la señal de habla se la considera cuasi-periódica en pequeños intervalos de tiempo ya que la mayor parte del tiempo contiene

sonidos sonoros en la formación de silabas, palabras, etc. El tamaño de intervalo para poder cumplir esta consideración debe de ser alrededor de 20 ms. En la figura 3.4 se muestra el oscilograma del fonema /a/ durante 93.75 ms, en que se puede destacar la forma cuasi-periódica en intervalos pequeños.

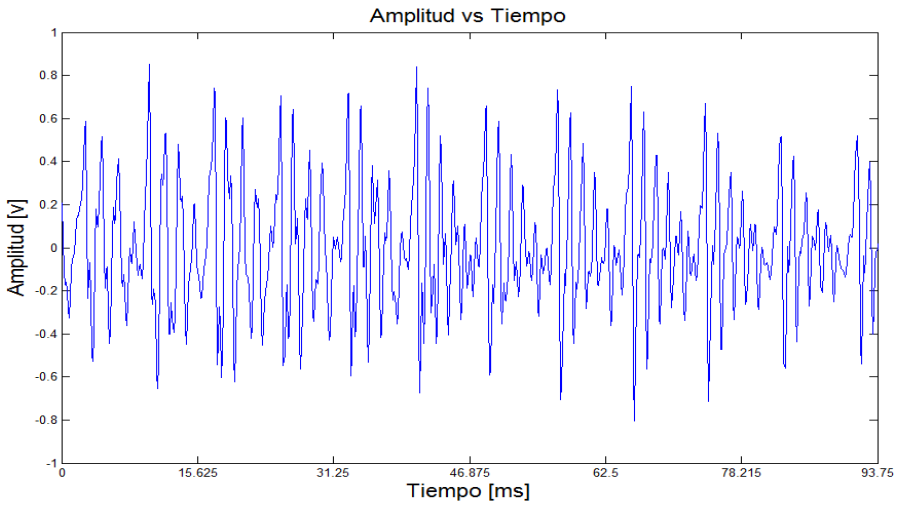


Figura 3.4 Oscilograma del fonema /a/

El objetivo del análisis oscilográfico en cortos periodos de tiempo, es la observación de las características de la señal de excitación sonora o sorda (periodicidad o ruido respectivamente). En la figura 3.5 se muestra la representación acústica del fonema /p/ y /a/ respectivamente.

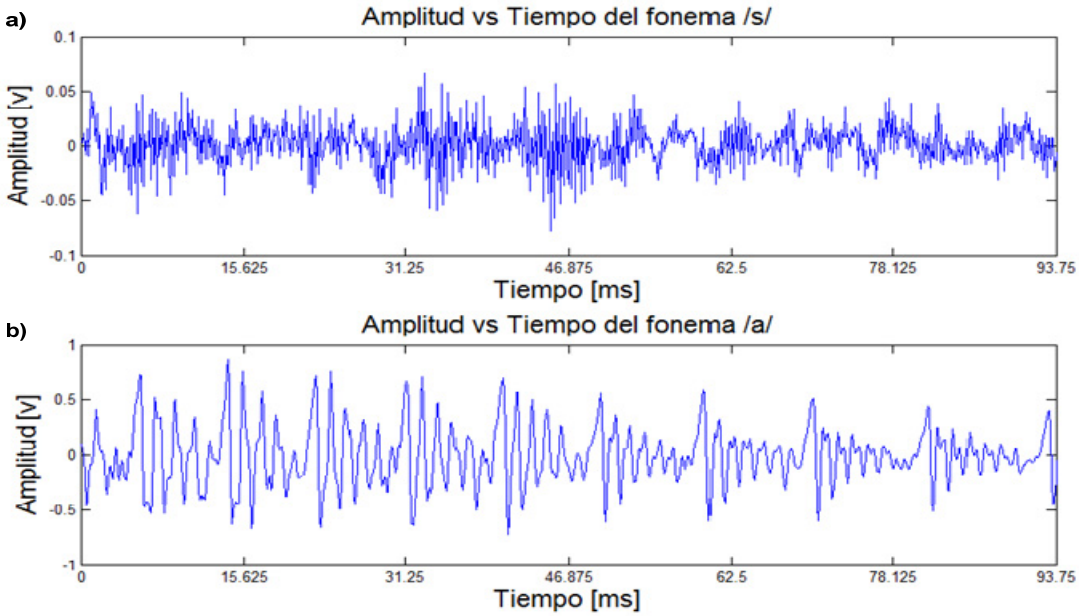


Figura 3.5 oscilograma de: (a) señal sorda, (b) señal sonora

Análisis de Energía

Hemos observado con el oscilograma, que la amplitud de una señal de habla varía apreciablemente con el tiempo. En el caso de los segmentos sordos poseen una amplitud relativamente baja con respecto a los sonidos sonoros. En la figura 3.6 se muestra una señal de habla “sofá” en la que se puede apreciar la diferencia de amplitudes entre segmentos de voz sonoros y sordos.

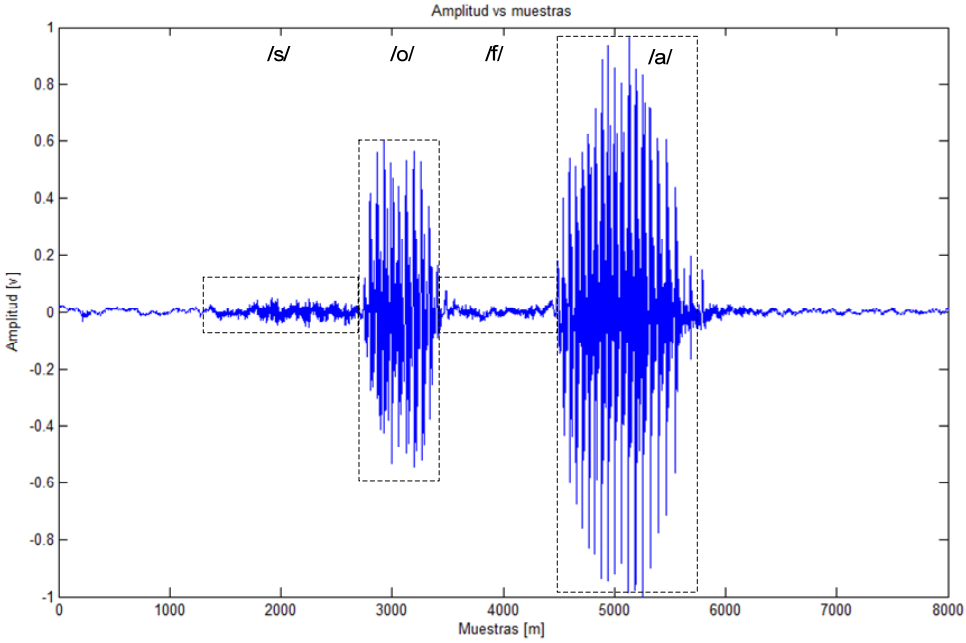


Figura 3.6: Segmentación de la señal de habla “sofá”.

El análisis de energía se basa en la representación de la energía promedio en tiempo corto, la cual refleja las variaciones en amplitud de manera más explícita posibilitando la diferenciación entre segmentos sordos y sonoros, además puede ser usada para distinguir la voz del silencio.

En general, definimos la energía en tiempo corto como:

$$E_n = \sum_{m=-\infty}^{+\infty} x(m)^2\, h(n-m) \qquad -\infty < n < \infty \qquad (3.1)$$

Donde h (n-m) es una ventana cuya función principal para el caso particular de análisis en el tiempo es la selección una porción de señal, x (m) es la señal de habla y E_n es el vector que almacenará los valores de energía de cada segmento de la señal. En la figura 3.7 se muestra a la ecuación 3.1 a manera de diagrama de bloques.

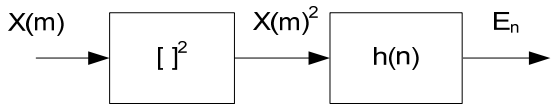


Figura 3.7: Diagrama de bloques de la energía promedio.

El efecto de la ventana en la representación de la energía dependiente del tiempo es discutido en el apéndice A, para la representación de energía en tiempo corto utilizaremos la ventana rectangular descrita en la ecuación 3.4.

$$h(n) = \begin{cases} \dfrac{1}{N} & 0 \le n \le N-1 \\ 0 & \text{c.c.} \end{cases} \tag{3.2}$$

En la figura 3.8 se muestra la aplicación de energía promedio sobre la señal de habla “sofá”, en la que se aprecia los segmentos sonoros y sordos por medio de la amplitud de la energía.

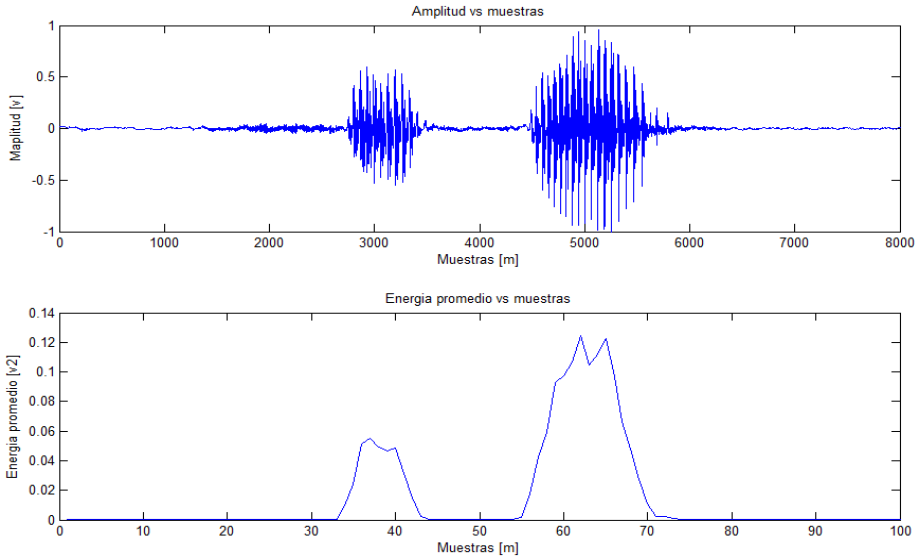


Figura 3.8: oscilograma y energía promedio de la palabra “sofá”

Análisis de tasa de cruces por cero

En el contexto de las señales en tiempo discreto, se dice que un cruce por cero ocurre si muestras sucesivas tienen signos algebraicos diferentes.

Las señales de voz son señales de banda ancha y la interpretación de la frecuencia promedio de cruces por cero es por tanto menos precisa. Sin embargo se pueden obtener estimados de las propiedades espectrales usando una representación basada en la tasa promedio de cruces por cero en tiempo corto. Antes de discutir la interpretación de la tasa de cruces por cero para la

voz, definiremos y discutiremos los cálculos requeridos. Una definición apropiada es:

$$Z_n = \sum_{m=-\infty}^{\infty} \left| \operatorname{sgn}\left[x(m)\right] - \operatorname{sgn}\left[x(m-1)\right] \right| \, w(n-m) \tag{3.3}$$

Donde

$$\operatorname{sgn}\left[x(n)\right] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases} \tag{3.4}$$

Y

$$w(n) = \begin{cases} \dfrac{1}{2N} & 0 \leq n \leq N-1 \\ 0 & \text{c.c.} \end{cases} \tag{3.5}$$

Tal como la representación de energía promedio, se utiliza una ventana para segmentar la señal de habla, la ventana utilizada en este caso está definida por la función w (n).

Las operaciones involucradas en la ecuación (3.3) se encuentran representadas en forma de diagrama de bloques en la figura 3.9. Esta representación muestra que la tasa promedio de cruces por cero en tiempo corto tiene las mismas propiedades generales que la energía en tiempo corto.

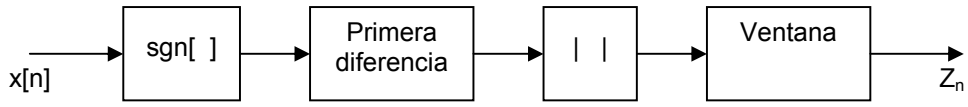


Figura 3.9: Representación en diagrama de bloques de los cruces por cero promedio en tiempo corto.

El modelo para la producción de voz sugiere que la energía de la voz sonora se encuentra concentrada por debajo de los 3 kHz debido a la caída del espectro introducida por la onda glotal, mientras que para la voz sorda, la mayor parte de la energía se encuentra a altas frecuencias. Debido a que las altas frecuencias implican altas tasas de cruces por cero, y bajas frecuencias implican bajas tasas de cruces por cero, existe una fuerte correlación entre la tasa de cruces por cero y la distribución de energía con la frecuencia. Una generalización razonable es que si la tasa de cruces por cero es alta, la señal de voz es sorda, mientras que si la tasa de cruces por cero es baja, la señal de voz es sonora. Esto, sin embargo, es una afirmación muy imprecisa ya que no hemos establecido qué es alto y qué es bajo, y por supuesto, no es posible ser preciso. En la figura 3.10 se muestra la representación de tasa de cruces por cero de la señal de habla “sofá” y su oscilograma (13).

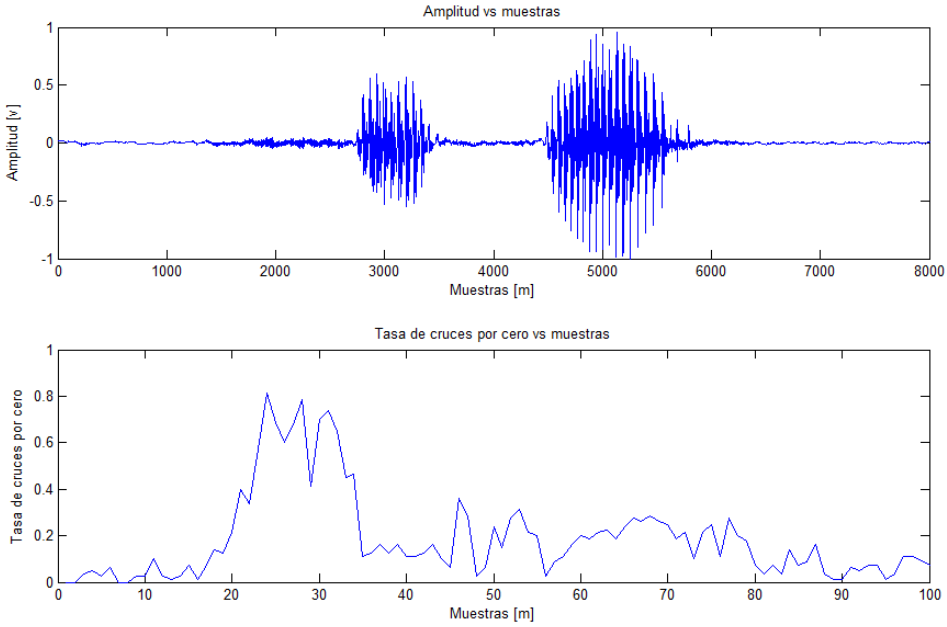


Figura 3.10: Representación oscilográfica y tasa de cruces por cero de la señal de "sofá"

3.3.2 Análisis en el dominio de frecuencia

Matemáticamente el análisis espectral está relacionado con una herramienta llamada transformada de Fourier. Ese análisis debe llevarse a cabo en intervalos pequeños de la señal de voz . La transformada no solamente contiene información sobre la intensidad de determinada frecuencia, sino también sobre su fase. Esta información se puede representar como un vector bidimensional o como un número complejo.

Con el fin de estudiar las propiedades espectrales de las señales de voz, debemos encontrar conveniente el introducir formalmente el concepto de una representación de Fourier dependiente en el tiempo de una señal.

Análisis de espectro de potencia

La representación espectral de potencia que es una función de la frecuencia (parte real) para un segmento de voz en un instante de tiempo específico. Una escala logarítmica es usada para la potencia en el eje vertical, y una escala lineal para la frecuencia en el eje horizontal.

La representación de frecuencia de un corto periodo de la señal puede ser usada para analizar las propiedades de un segmento en particular que contenga un fonema y es muy útil porque la estructura de frecuencia de un fonema es generalmente única. En la figura 3.11 se muestra la representación de potencia espectral de un segmento sonoro.

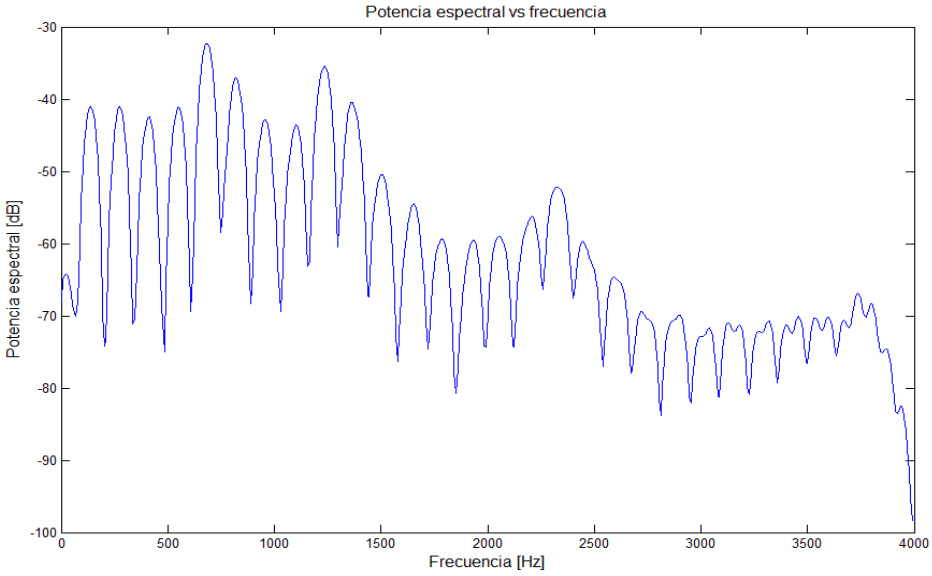


Figura 3.11: Representación de potencia espectral del fonema /a/

Descomposición del espectro de potencia

La señal de voz es a menudo caracterizada en los términos de las propiedades de la potencia espectral. Los dos atributos importantes de la potencia espectral son:

- La envolvente espectral.
- La estructura fina espectral.

En la figura 3.12 se muestra la descomposición de la potencia espectral del fonema /a/.

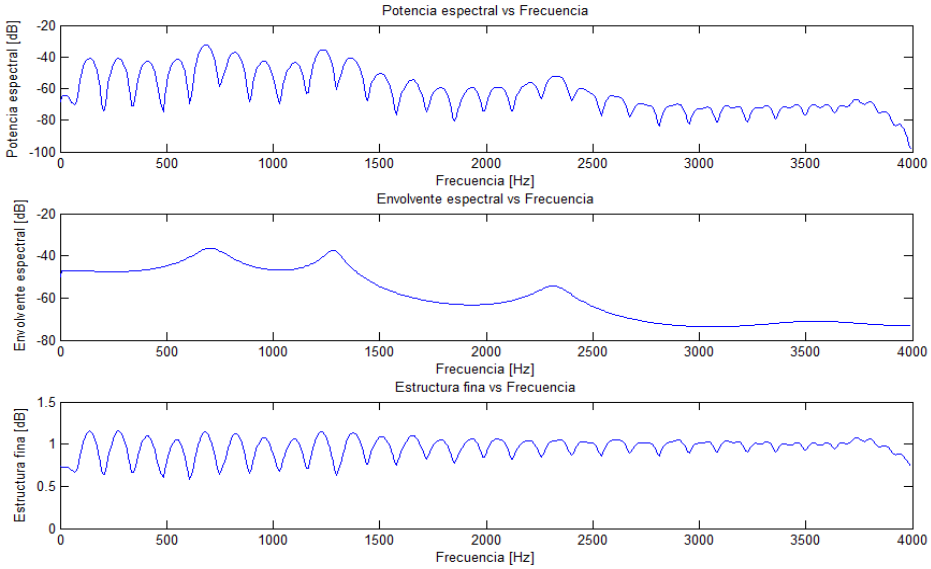


Figura 3.12: Descomposición de la potencia espectral del fonema /a/

Envolvente espectral

La envolvente espectral caracteriza a el tracto vocal con la forma de posee una estructura única compuesta por formantes capaz de diferenciar este fonema de otros.

Los formantes son determinadas frecuencias de resonancia introducidas por las cavidades en la boca y la nariz (configuración específica de los órganos del tracto vocal al pronunciar un determinado sonido), desempeñan un papel importante en la diferenciación de sonidos ya que cada sonido conllevara

frecuencias formantes características y diferenciadores respecto al resto de los sonidos.

La detección de estas frecuencias de resonancia se realiza en la envolvente espectral, constituyendo los formantes los máximos relativos de dicha envolvente; es decir las frecuencias formantes se definen como lo máximos parciales de la envolvente espectral.

Para los sonidos sonoros las frecuencias armónicas cerca de las resonancias son enfatizadas. Las resonancias de las cavidades que son típicos de las diferentes configuraciones de articulación son llamadas formantes. En la Figura 3.13, se muestran 3 formantes del fonema /a/ (máximos parciales de la envolvente espectral). Los sonidos sordos muestran solo formantes menos pronunciados debido a la excitación ruidosa. En la figura 3.14 se muestran los formantes del fonema /s/.

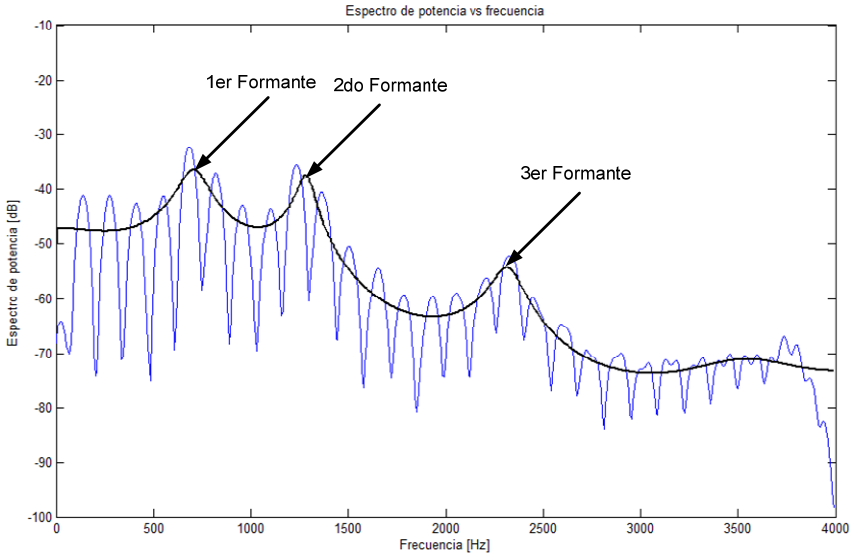


Figura 3.13: Envolvente espectral del fonema /a/.

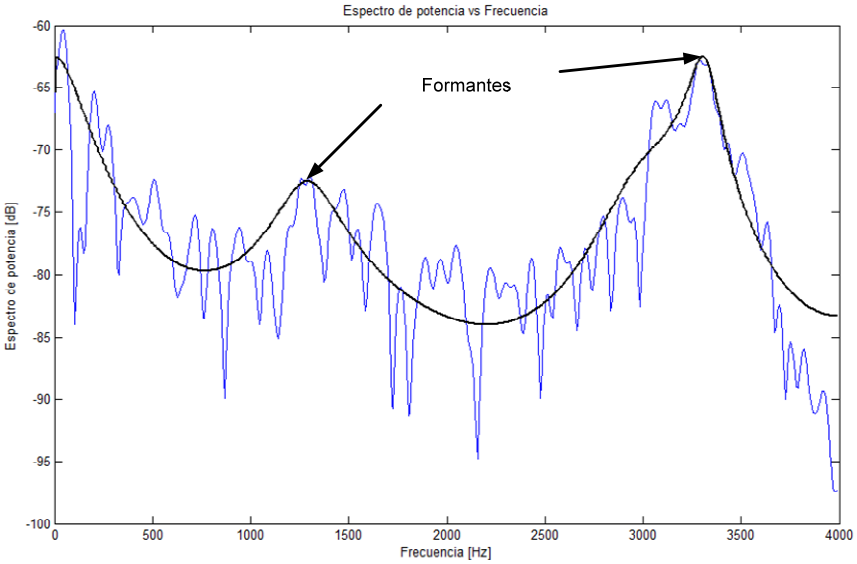
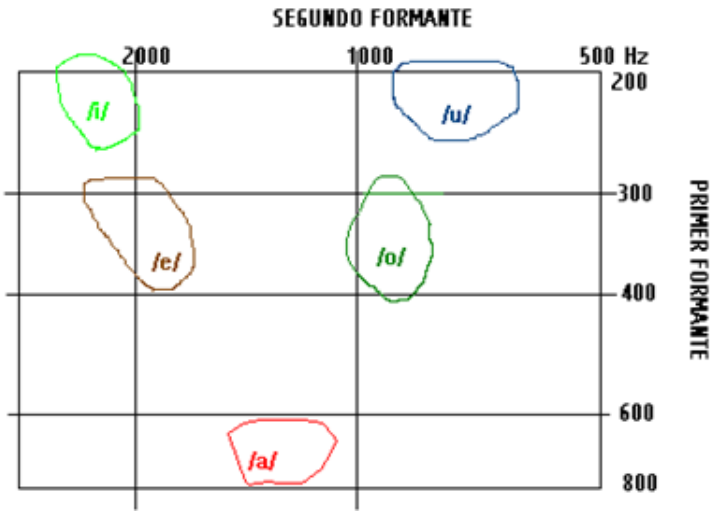


Figura 3.14: Envolvente espectral el fonema /s/.

En el caso de las vocales son necesarias tan solo las dos primeras frecuencias formantes. Por lo tanto los formantes son una característica importante de la percepción de la voz. Los sistemas de reconocimiento automático basan su eficiencia en la capacidad de extraer adecuadamente la envolvente espectral. En la figura 3.15 se muestra el intervalo de los formantes primero y segundo de las vocales.



3.15: Primer y segundo formante de los fonemas vocálicos.

Estructura fina

Estructura fina caracteriza a las cuerdas vocales y es usada para distinguir entre sonidos sonoros y sordos. Para sonidos sonoros hay una estructura armónica con picos en múltiples frecuencias fundamentales. Si una estructura armónica

en la estructura fina está perdida entonces los sonidos pueden ser considerados como sordos. La figura 3.16 muestra la descomposición de la potencia espectral extrayendo la estructura fina del fonema /a/ y en la figura 3.17 la estructura fina del fonema /s/.

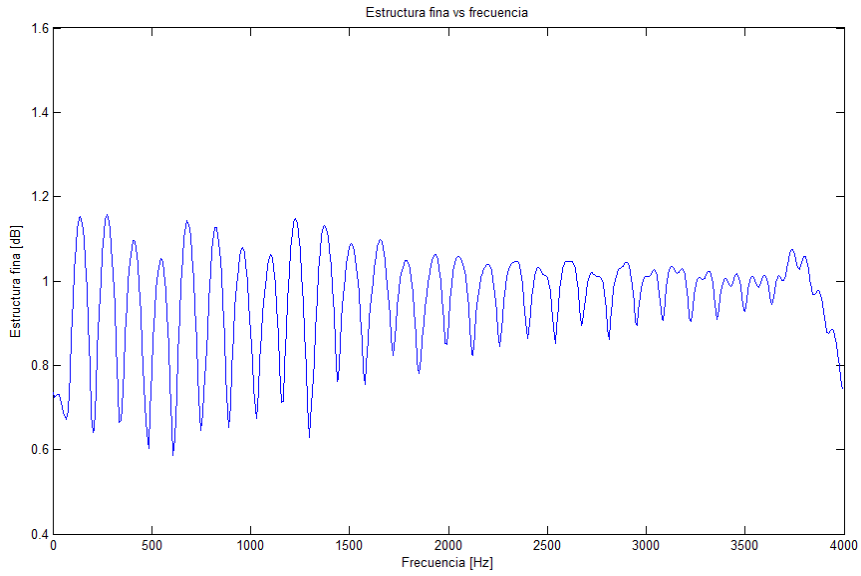


Figura 3.16: Estructura fina del fonema /a/.

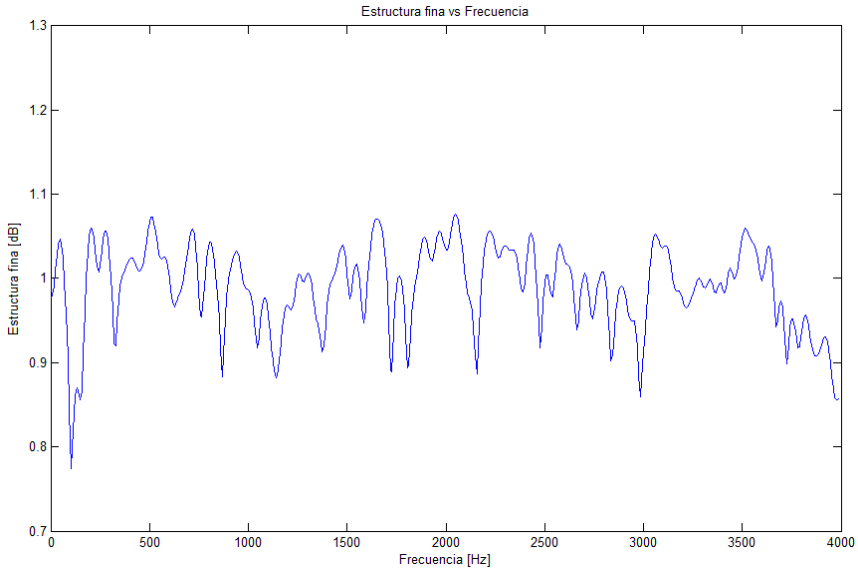


Figura 3.17: Estructura fina del fonema /s/.

Espectrograma

La señal de voz representada sobre un espectrograma es otra representación muy usada. En el eje horizontal representa al tiempo. Y en eje vertical representa a la frecuencia (potencia espectral) y la intensidad en una frecuencia dada en el tiempo está representada como tercera dimensión por los colores que van desde el azul hasta el rojo, menor a mayor respectivamente. Finalmente lo más importante de esta representación son las frecuencias formantes (frecuencias de mayor intensidad), las cuales hacen distinción de los fonemas.

En la figura 3.18 se muestra el procedimiento para llegar a construir un espectrograma.

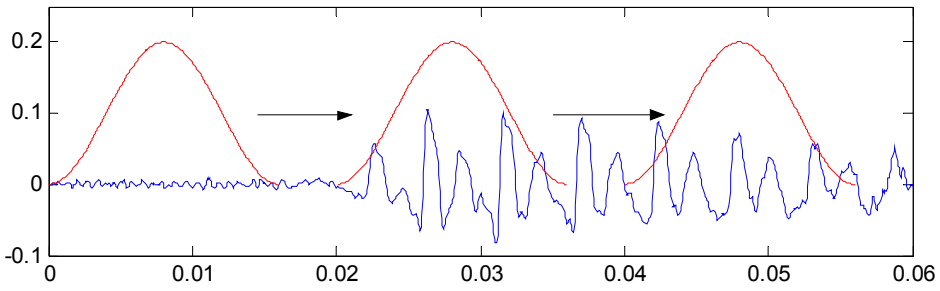


Figura 3.18a representación acústica de la voz en un intervalo de 60 ms.

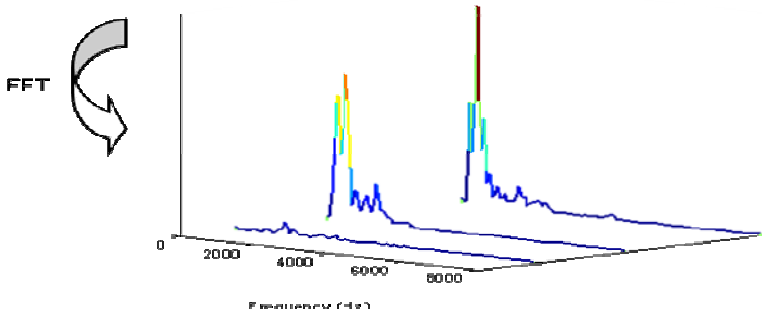


Figura 3.18b representación de frecuencia de tres tramas

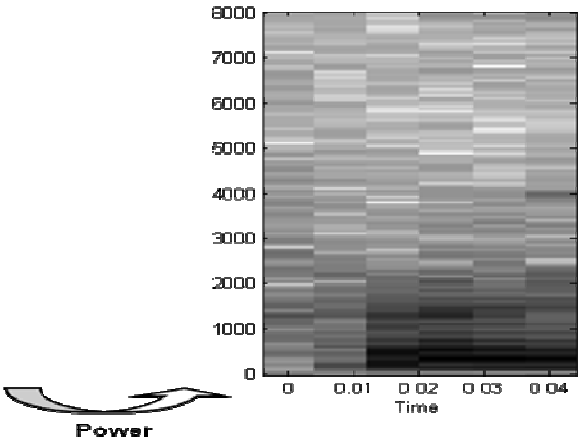


Figura 3.18c Espectrograma en un intervalo de 40 ms.

El procedimiento de construcción de un espectrograma es fácilmente apreciable: la figura 3.18a muestra que se parte desde la representación acústica, dividiéndola en tres pequeños intervalos de alrededor de 10 ms. Los cuales en la figura 3.18b se observa su transformación a frecuencia mediante la transformada de Fourier, y también en colores la magnitud. Finalmente mediante la obtención de la intensidad o potencia de las tramas se puede llegar al espectrograma apreciado en la figura 3.18c.

En la figura 3.19 se muestra el oscilograma y el espectrograma de la señal de habla “dos”.

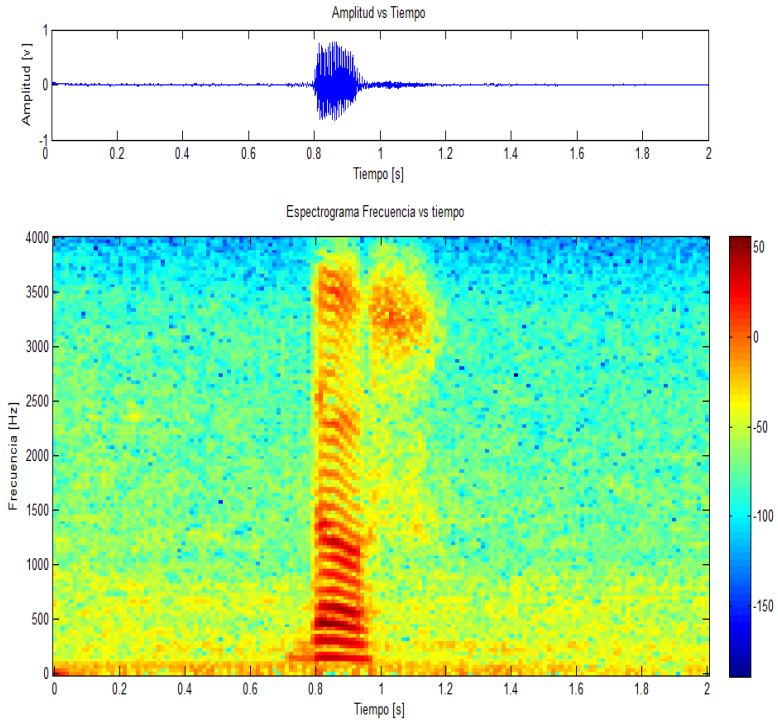


Figura 3.19: Representación acústica y espectrograma de la señal de habla “dos”.

La particularidad del espectrograma es que nos permite observar las propiedades espectrales de manera más explicitas. En la figura 3.20 se muestran los espectrogramas de los fonemas vocálicos, ya que estos se definen como zonas de color rojizo (14).

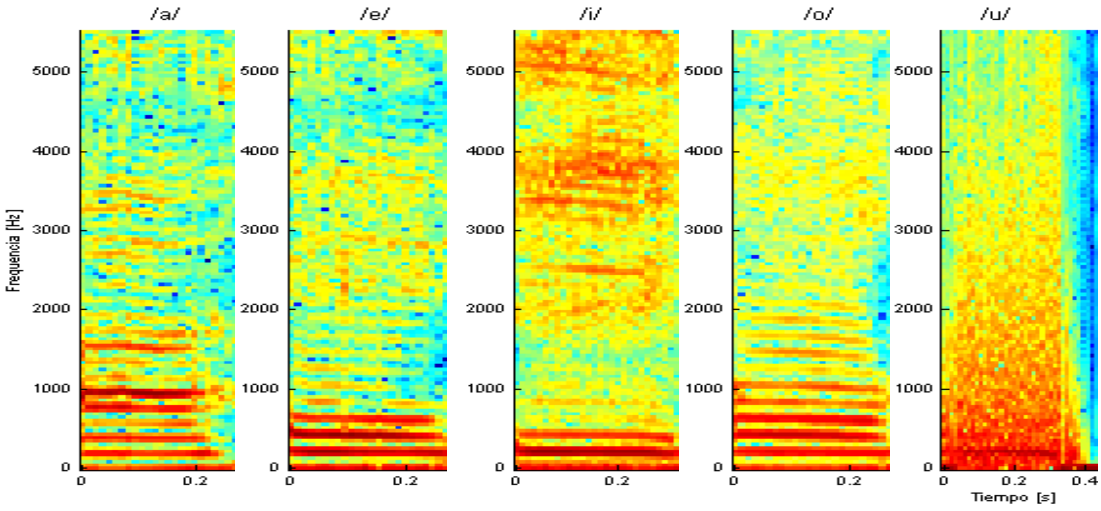


Figura 3.20: Espectrogramas de los fonemas vocálicos

CAPÍTULO 4

4. DISEÑO DEL SISTEMA

Este capítulo tiene como objetivo la explicación del sistema ASR (reconocimiento de voz automático por sus siglas en inglés) a implementar en matlab por lo se define una estructura o diagrama de bloques que contiene los puntos importantes del sistema.

4.1 Estructura del sistema a implementar

La estructura o diagrama de bloques del sistema ASR (reconocimiento de voz automático por sus siglas en inglés) a desarrollar, consiste de los bloques básicos que debe poseer todo sistema de reconocimiento: adquisición de la

señal, extracción de parámetros o información comparable de la señal, y bloque de comparación; que se define como bloque de entrenamiento, reconocimiento y plantillas. En la figura 4.1 se muestra la estructura expuesta.

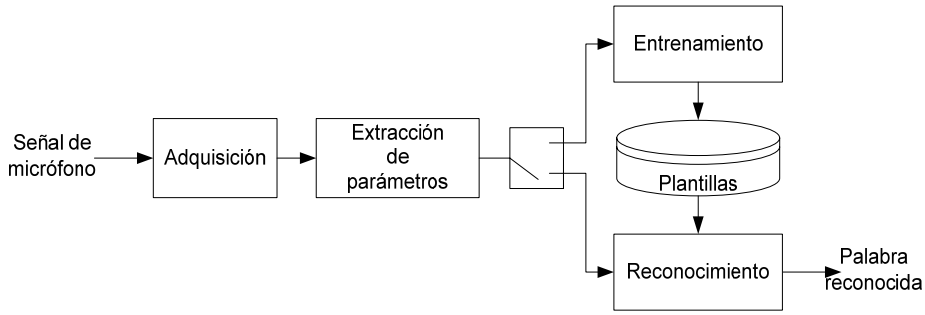


Figura 4.1 Diagrama de bloques del sistema de reconocimiento del habla.

4.2 Adquisición

El bloque de adquisición recibe a la entrada una onda acústica (voz) y genera a la salida un arreglo con los valores de magnitud de dicha onda. Internamente esta compuesto por tres bloques muestreo, cuantización y el bloque de detección de actividad de voz, En la figura 4.2 se representa los bloques que componen a la función de adquisición.

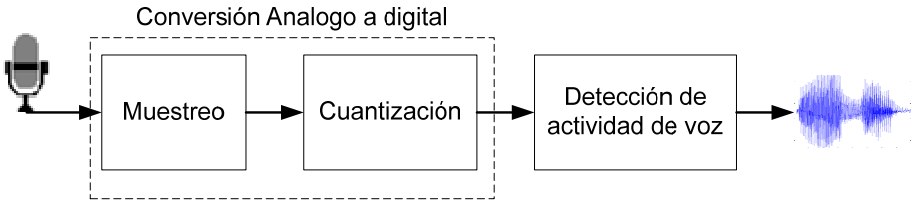


Figura 4.2: Diagrama de bloques de la adquisición.

4.2.1 Muestreo y cuantificación

Muestreo

La etapa de muestreo consiste en convertir la señal analógica, continua en el tiempo, en una señal discreta en el tiempo. El principio fundamental del muestreo es el denominado teorema de Nyquist. El cual enuncia que si la frecuencia de muestreo es mayor o igual al doble del ancho de banda de la señal a muestrear, se podrá recuperar la señal en su totalidad mediante una interpolación basada en funciones seno. Este hecho es importante, ya que nos indica que si el muestreo se realiza de forma correcta, no se pierde información.

Los estudios sobre las características de las señales de voz han demostrado que la mayor parte de la información necesaria para la inteligibilidad del habla se encuentra por debajo de los 4 KHz. De hecho el ancho de banda disponible tradicionalmente en las líneas telefónicas es algo menor de 4 KHz. Aunque hay que destacar que algunos sonidos emitidos por el aparato fonador poseen frecuencias mucho más elevadas, por ejemplo los sonidos fricativos (sonidos que se producen cuando un articulador se acerca a una zona de articulación de modo que el paso del aire se obstruye parcialmente, produciendo una fricción) los cuales pueden alcanzar los 10 KHz., pero la pérdida de esta información no

supone un déficit sustancial en la información del habla. Por tanto teniendo en cuenta el teorema de Nyquist, con una frecuencia de muestreo de 8000 Hz sería suficiente para los propósitos de reconocimiento del habla.

Las posibles mejoras que se obtendrían utilizando una frecuencia de muestreo mayor, no compensarían la carga computacional adicional debida a la mayor cantidad de datos a procesar.

Cuantificación

Una vez se tiene la señal discreta en el tiempo se debe discretizar en magnitud para tener una señal digital. Para cuantificar una señal con K bits, el número de niveles que garantiza la mayor eficiencia en el uso de las palabra en código binario es 2^K. La amplitud y el número de niveles se eligen a la vez para cubrir adecuadamente el rango de la señal. Si X_{MAX} es el valor máximo de la amplitud de la señal, los parámetros de la cuantificación deben elegirse de acuerdo a la siguiente expresión:

$$2X_{max} = \Delta 2^K,$$

Donde K es el número de bits empleados en la cuantificación, y Δ es la amplitud de los intervalos.

4.2.2 Algoritmo de conversión analógico a digital

En matlab la conversión análoga a digital de la señal de habla se la hace a través de la función **wavread.**

Habla = wavread (duración, fs., tipo de dato);

En matlab se realiza un almacenamiento de la señal de voz, la duración esta en términos de muestras por lo que si se desea adquirir 2 segundos se tendrá que colocar fs * 2.

Tipo de dato es la cuantificación en el caso a implementar es doble.

4.2.3 Detección de actividad de voz

La detección de actividad de voz consiste en localizar de manera automática el inicio y final de una señal de voz, con el propósito de almacenarla sin información redundante (silencios al inicio o final). Dado que en la práctica la señal de voz está acompañada de ruido de fondo, el problema de la detección de voz se complica dependiendo de la relación señal a ruido.

La señal del habla del lenguaje castellano está compuesta fundamentalmente por vocales acompañadas por consonantes, las características antes descritas en el capítulo 3 concluyen que la representación de energía permite localizar a

los sonidos sonoros (mayormente representados por las vocales); mientras que la tasa de cruces por cero permite la observación de los sonidos sordos (mayormente representados por las consonantes a excepción de los fricativos por poseer excitación sonora).

En el capítulo 3 se introdujeron las representaciones de la señal de voz mediante energía y tasa de cruces por cero. El algoritmo a implementar en la detección de voz es el algoritmo de Rabiner y Saburm (15).

Algoritmo en matlab

El algoritmo de Rabiner y Saburn:

1. Calcular la medida y desviación estándar para E[n] y Z[n] a partir de los primeros 100 ms de señal.
2. Calcular umbral para Z[n] basado en la distribución de Z[n] en tramos sordos: IZCT (umbral de tasa de cruces por cero).
3. Calcular umbrales “upper” y “lower” para E[n]: ITU (umbral superior de energía), ITL (umbral superior de energía) respectivamente.
4. Encontrar un intervalo que sobrepase ITU.
5. Encontrar puntos iníciales y finales tentativos N1 y N2 buscando los cruces del umbral inferior ITL fuera del intervalo anterior.

6. Mover N1 hacia atrás hasta el primer punto en el que Z[n] excede el umbral IZCT.
7. Mover N2 hacia adelante hasta el primer punto en el que Z[n] excede el umbral IZCT.

El algoritmo puede ser descrito con referencia en la figura 4.3. Las representaciones básicas usadas son la cantidad de cruces por cero en una trama de 10ms y la magnitud promedio calculados con una ventana de 10ms. Ambas funciones son calculadas para todo el intervalo de grabación a una tasa de 100 veces/s. Se asume que los primeros 100ms del intervalo no contienen voz. La media y desviación estándar de la magnitud promedio y la tasa de cruces por cero son calculados para este intervalo para dar una caracterización estadística del ruido de fondo. Usando esta caracterización estadística y la magnitud promedio máxima del intervalo, se calculan los umbrales de energía y de cruces por cero. Se busca la magnitud promedio para encontrar el intervalo en el que siempre se excede un cierto umbral conservador (ITU en la figura 4.3). Se asume que los puntos de inicio y final se encuentran fuera de este intervalo. Entonces trabajando hacia atrás desde el punto en el que M_n excede por primera vez el umbral ITU, el punto (marcado como N_1 en la figura 4.3) donde M_n cae por primera vez debajo de un umbral inferior ITL es tentativamente

seleccionado como el punto de inicio. Se sigue un procedimiento similar para encontrar el punto tentativo final N_2. Este procedimiento de doble umbral asegura que las caídas en la función magnitud promedio no proporcionen un punto final falso. En este punto es razonablemente seguro asumir que los puntos de inicio y final no se encuentran en el intervalo de N_1 a N_2. El siguiente paso es moverse hacia atrás desde N_1 (hacia delante de N_2) comparando la tasa de cruces por cero con un umbral (IZCT en la figura 4.3) determinado a partir de las estadísticas de la tasa de cruces por cero para el ruido de fondo. Este se encuentra limitado a 25 tramas antes de N_1 (después de N_2). Si la tasa de cruces por cero excede el umbral 3 ó más veces, el punto de inicio N_1 se traslada hacia atrás al primer punto en que se excede el umbral de cruces por cero. Si lo anterior no sucede, N_1 se define como el inicio. Se sigue un procedimiento similar para el final.

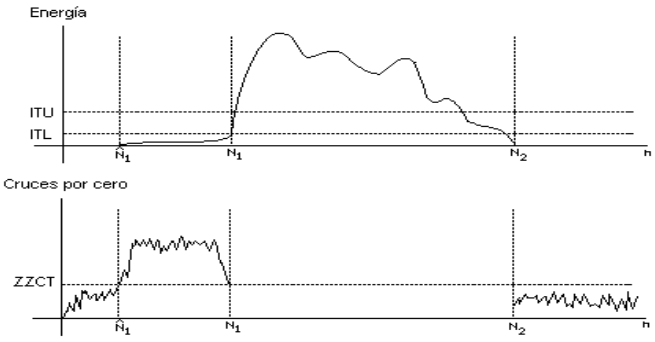


Figura 4.3. Ejemplo típico de las mediciones de la magnitud promedio y cruces por cero para una palabra que comienza con una fricativa fuerte.

A pesar de las dificultades creadas por las situaciones anteriores, las representaciones de energía y tasa de cruces por cero pueden ser combinadas para ser usadas como base de un algoritmo que localice el inicio y fin de una señal de voz. Uno de dichos algoritmos fue estudiado por Rabiner y Cambur en el contexto de un sistema de reconocimiento de voz de palabras aisladas. En este sistema un emisor pronuncia una palabra durante un intervalo de grabación prescrito, y el intervalo completo es muestreado y guardado para el procesamiento. El propósito del algoritmo es encontrar el inicio y final de la palabra con el fin de que el subsecuente procesamiento y comparación de patrones puedan ignorar el ruido de fondo.

4.3 Extracción de parámetros

La extracción de parámetros tiene como objetivo obtener la información más relevante de la señal de habla, esta información fue definida en el capítulo 2 con el nombre de envolvente espectral. La cual describe de manera explícita las frecuencias formantes únicas de cada fonema.

Atendiendo al modelo usual de producción del habla, La voz se produce mediante la convolución de la excitación glotal (vibración de las membranas de

las cuerdas vocales con el paso del aire) y la respuesta al impulso del tracto vocal, por lo que es necesario procesar esta señal de habla para poder extraer la señal de tracto (envolvente espectral), ya que esta posee la estructura que diferencia a un fonema de otro. Esta separación no se puede conseguir mediante filtrado ya que ambas componentes no están combinadas linealmente. Para lo cual es necesario aplicar un sistema homomórfico.

4.3.1 Análisis homomórfico

Este término significa literalmente: la misma estructura. La adición no es la única manera para que el ruido y la interferencia se puedan combinar con una señal de interés; la multiplicación y la convolución son también medios comunes de mezcla de señales próximas, si las señales se combinan de una manera no lineal, no pueden ser separadas por filtración lineal. Las técnicas homomórficas procuran separar señales combinadas de una manera no lineal haciendo que el problema se convierta en lineal. Es decir, el problema se convierte en la estructura de un sistema lineal.

Desde la introducción en los primeros años de la década de los 70, de las técnicas homomórficas de procesado de señal, su importancia dentro del campo del reconocimiento de voz ha sido muy grande. Los sistemas homomórficos son

una clase de sistemas no lineales que obedecen a un principio de superposición.

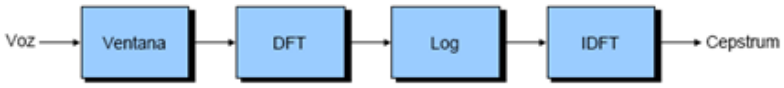
De estos, los sistemas lineales constituyen un caso especial.

4.3.2 Análisis cepstral

Atendiendo al modelo usual, la voz se produce mediante la convolución de una excitación y la respuesta impulsiva del modelo del tracto vocal. En algunas ocasiones es importante aislar una de las componentes (excitación o modelo del tracto vocal) para su uso por algoritmos de tratamiento digital de la voz.

Una de las técnicas más sencillas para separar componentes de una señal es el filtrado, es útil cuando las componentes aparecen combinadas linealmente, puesto que la operación de filtrado es lineal.

En la producción de la señal de voz, la excitación y la respuesta impulsiva del filtro que modela el tracto vocal no se combinan linealmente, sino mediante una convolución (operación lineal, pero no combinación lineal).



$$c(n)=\frac{1}{N_s}\sum_{k=0}^{N_s-1}\log_{10}\left|S_{med}(k)\right|e^{j\frac{2\pi}{N_s}kn} \qquad 0\le n\le N_s-1$$

Figura 4.4 Análisis Cepstral partiendo de la transformada discreta de Fourier

El cepstrum $c(n)$ de la señal de voz se define como la transformada inversa de Fourier del logaritmo de su espectro $S(f)$, es decir,

$$c(n)=F^{-1}\log(S(f)) \qquad (4.1)$$

El término cepstrum (obsérvese la inversión intencionada del orden de las primeras letras con respecto a spectrum) es indicativo de haber realizado una transformación inversa del espectro. La variable independiente del cepstrum se denomina quefrency (proveniente de la variable inglesa frequency, también invertida) y tiene carácter temporal.

El análisis cepstral es un caso particular de procesado homomórfico que permite obtener una representación de la señal de voz en un dominio donde la

excitación y el modelo del tracto vocal se combinan atendiendo a las siguientes propiedades:

- Las componentes representativas aparecen separadas en ese nuevo dominio
- Las componentes representativas aparecen combinadas linealmente

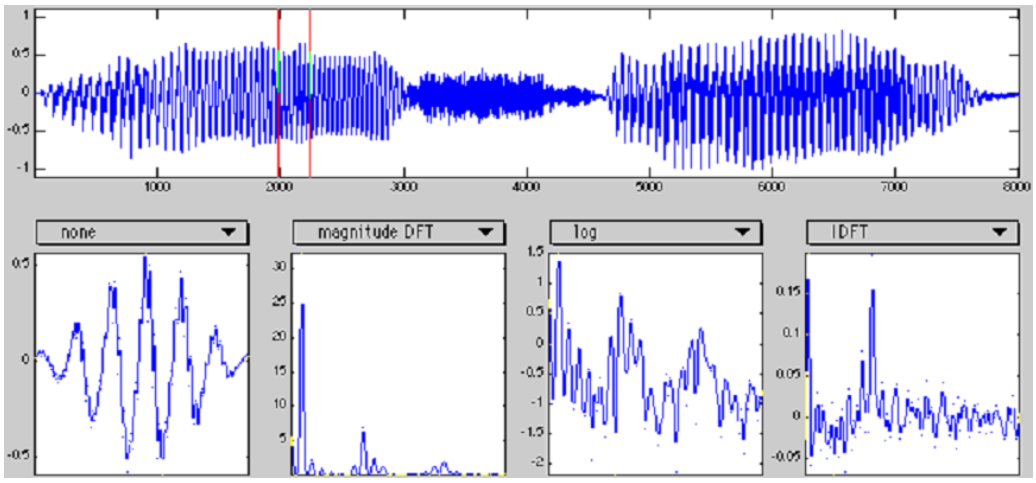


Figura 4.5 a) Forma acústica, b) magnitud de DFT, c) logaritmo,d) idft de la senal y dominio cesptral

4.3.3 Liftro

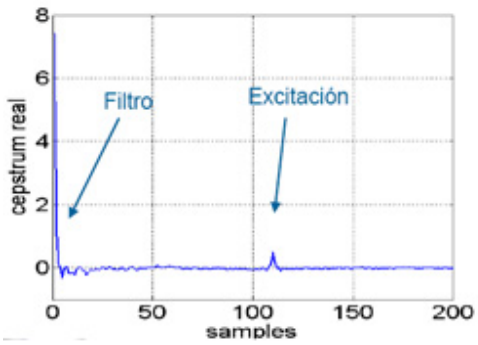


Figura 4.6 Separación excitación-filtro por filtrado homomórfico

El proceso de separar las componentes cepstrales se denomina liftrado (liftering en ingles, derivado de filtering, filtrado) . Las señales mezcladas por convolución en el dominio del tiempo aparecen mezcladas aditivamente en el dominio de la frecuencia. Por tanto la representación cepstral es ideal para separar las características correspondientes a la excitación y al filtro que modela el tracto vocal, los coeficientes cepstrales de orden bajo contienen información sobre la envolvente espectral y caracterizan el filtro que modela el tracto vocal, los coeficientes cepstrales de orden mayor representan el rizado espectral y es en estos donde se manifiesta el pitch. Aplicando ventanas adecuadas al conjunto de coeficientes cepstrales se pueden separar estas características (excitación y envolvente espectral).

Los coeficientes de orden bajo proveen entonces información sobre la envolvente. La palabra está originada por la alteración de las letras que forman spectrum ya que en realidad su dominio no es temporal ni frecuencial sino un nuevo dominio cepstral.

4.4 Comparación de parámetros

En el bloque de comparación de patrones, se comparan las características de una palabra de entrada desconocida con las características de las palabras almacenadas en memoria, por lo tanto la finalidad del bloque de comparación es seleccionar la palabra en memoria más parecida, devolviendo como resultado una cadena de caracteres con el nombre de la palabra.

Hay que recordar que la pronunciación de las palabras, se caracteriza por una gran variabilidad temporal de su forma; es decir una misma palabra se puede pronunciar a distintas velocidades, lo que produce importantes distorsiones temporales. Estas distorsiones no solo afectan a la duración total de la palabra, sino que también influyen en los sucesivos elementos fonéticos que la componen. La necesidad de comparar dos secuencias acústicas independientemente de las distorsiones temporales respectivas, exige una técnica de normalización temporal no lineal. Que para el caso particular de esta

tesis es DTW (dynamic time warping) en inglés, alineamiento temporal dinámico en español. En la figura 4.7 se muestra el diagrama de bloques de la comparación de patrones (16).

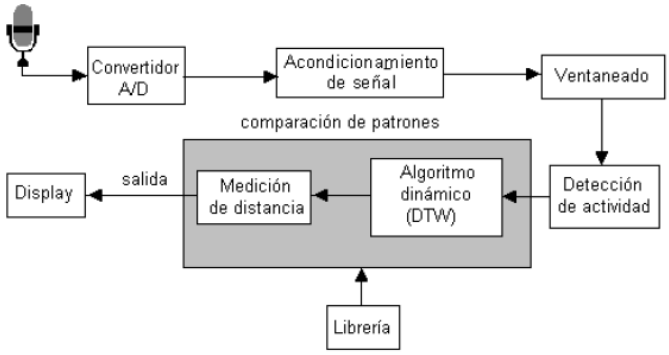


Figura 4.7 Diagrama de bloques de la comparación de patrones

4.4.1 Alineamiento temporal dinámico DTW

La técnica utilizada para la etapa de reconocimiento de señales de voz es el alineamiento temporal dinámico o DTW (Dynamic Time Warping), este modelo corresponde a uno de los primeros métodos desarrollados en reconocimiento y es uno de los más eficaces en el reconocimiento monolocutor. El problema principal recae en el tamaño del diccionario de palabras a reconocer y en el tiempo de procesamiento ocupado por los algoritmos. Sin embargo, como un primer acercamiento a los sistemas de reconocimiento de voz se constituye en

una de las herramientas básicas de reconocimiento, previo a la realización de sistemas más complejos.

En general DTW es un método que permite encontrar un emparejamiento entre dos secuencias dadas con ciertas restricciones. Las secuencias son "deformadas" de manera no lineal en la dimensión del tiempo para determinar una medida de su similitud independiente de las variaciones no lineales en la dimensión del tiempo.

Algoritmo:

Sea X=(x1, x2,…, xi) una trama de longitud N que contiene vectores cepstrales de la palabra desconocida, Y= (y1, y2,…, yj) una cadena de vectores cepstrales de longitud M. de las palabras almacenadas en una librería. La figura 4.8 muestra dos señales cualquiera x, y con sus respectivos valores.

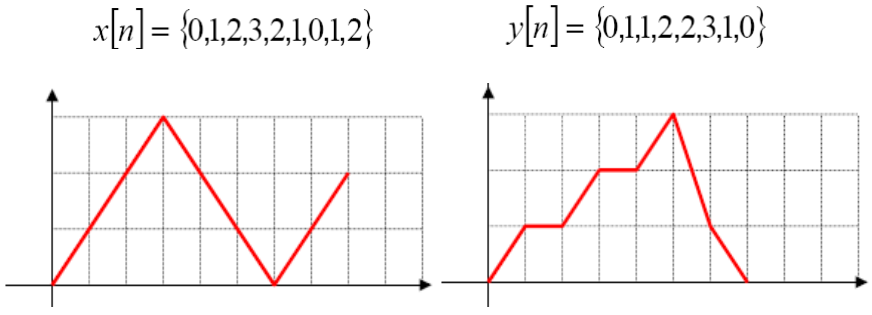


Figura 4.8: Señales x, y con sus valores

1. Construcción de una matriz de distancia local d(i, j) donde:

$$d(i,j) = |x[i] - y[j]| \; ; \quad 1 \leq i \leq N, \quad 1 \leq j \leq M \tag{4.2}$$

La ecuación 4.2 está representada en la figura 4.9

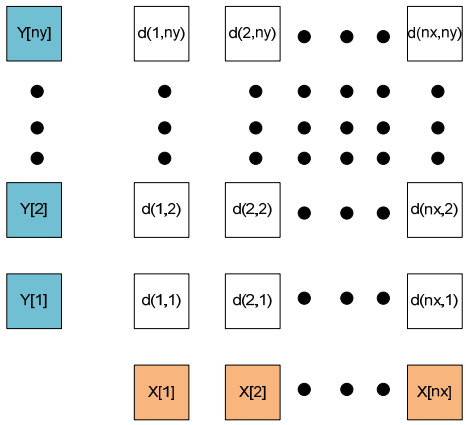


Figura 4.9: Matriz de distancia local

En el caso particular de las señales x, y de la figura 4.8 la matriz de distancia local está ilustrada en la figura 4.10.

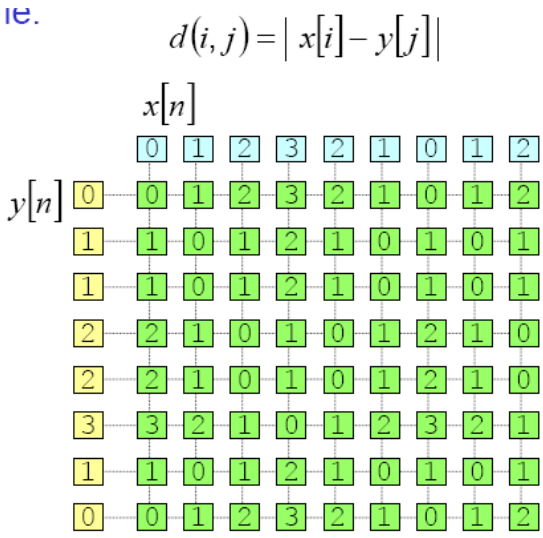


Figura 4.10: Matriz de distancia local para las señales

2. A partir de la matriz de distancias locales se aplica la función de recursión D(x, y) expuesta en la ecuación 4.3.

$$D(x,y) = d(x,y) + \min[D(x-1,y), D(x-1,y-1), D(x,y-1)] \quad (4.3)$$

$$0 \leq x \leq N, \qquad 0 \leq y \leq M$$

La misma función recursiva D(x, y) aplicada sobre la matriz de distancias locales de la figura 4.9 y mostrado de esta manera en la figura 4.11.

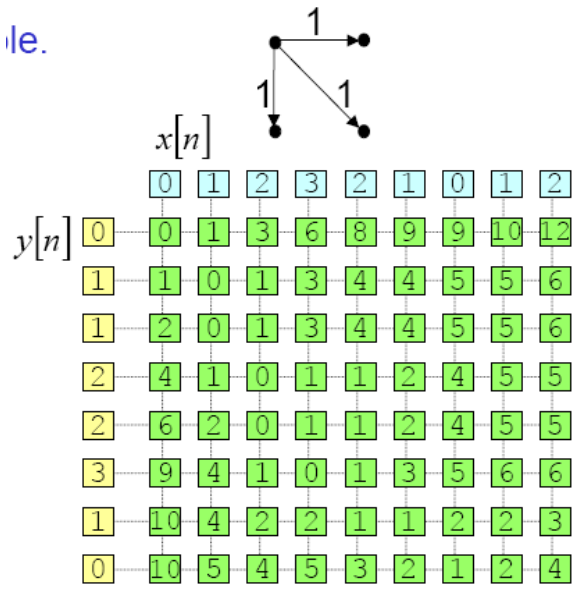


Figura 4.11: Matriz de distancia acumulativa A

3. Finalmente se obtiene que DTW(x, y) es:

$$DTW(x,y) = \frac{D(N,M)}{L} \qquad (4.4)$$

Donde D(N, M) es la función recursiva evaluada en (N, M) y L es la longitud del camino optimo desde (0, 0) y (N, M). En la figura 4.12 se muestran los valores necesarios para calcular el DTW(x, y) de la figura 4.8, y en la ecuación 4.9 se muestra el resultado.

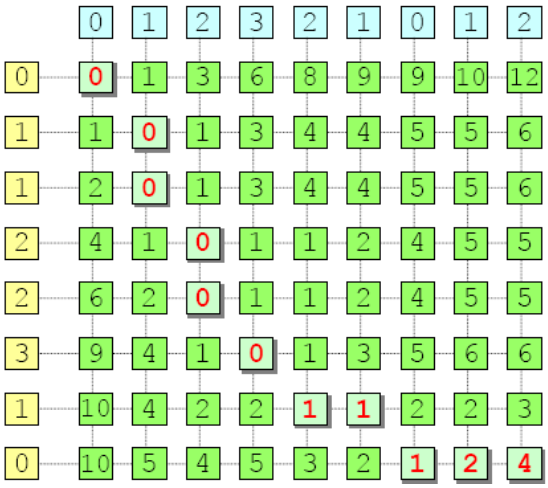


Figura 4.12: Matriz de distancia acumulativa B

$$DTW(x,y) = \frac{D(N,M)}{L} = \frac{4}{11} = 0.3636 \qquad (4.5)$$

4.4.2 Algoritmo de entrenamiento

El algoritmo de entrenamiento tanto como el algoritmo de reconocimiento aplica el algoritmo DTW. El propósito del algoritmo de entrenamiento es la creación de una base de datos de “palabras” donde cada “palabra” es una estructura que contiene 4 tipos de datos:

- Una cadena de caracteres que almacena el nombre de la palabra.
- Un arreglo de número reales que almacena una señal de habla

considerada señal patrón o referencia.

- Dos valores reales que almacenan los valores D_{max} y D_{1max}. Los cuales representan umbrales de aceptación mínimos para ser usados en la etapa de reconocimiento.

Una vez definida la estructura es necesario definir el procedimiento de entrenamiento que me de cómo resultado la obtención de todos los elementos de la estructura:

1. Almacenar 5 señales de habla de la misma palabra $\{t_1, t_2, t_3, t_4, t_5\}$, de entre estas señales posteriormente una será seleccionada como la más representativa llamada señal referencia.

2. Encontrar la señal más representativa aplicando la ecuación 4.6 y 4.7; Donde D_{final} es la suma de los resultados DTW entre tramas de las señales de habla $t_j(n), t_k(n)$.

$$S(j) = \sum_{k=1}^{5} D_{final}(t_j(n), t_k(n)) \qquad (4.6)$$

$$t_R(n) \text{ será tal que } S(R) < S(j), \forall j \mid j \neq R \text{ y } 1 \leq j \leq 5. \qquad (4.7)$$

Lo que expresa la ecuación 4.6 es que en cada señal será evaluada su disimilitud con respecto a las otras cuatro señales, así los valores $s(j)$ nos proporcionan un indicador importante para encontrar la señal referencia.

La ecuación 4.7 define a $t_R(n)$ como señal referencia, la cual debe poseer el menor valor de $s(j)$.

3. Encontrar los valores D_{max} y D_{1max} los cuales se definen como:

D_{max} Es la mayor distancia D_{final} entre el elemento más representativo o señal referencia y otro elemento de entre el conjunto de 5 señales.

D_{1max} Es la mayor distancia D_{final} entre dos elementos de entre todo el conjunto de señales.

Para tener una mejor idea acerca de estos valores se muestra gráficamente mediante la figura 4.13.

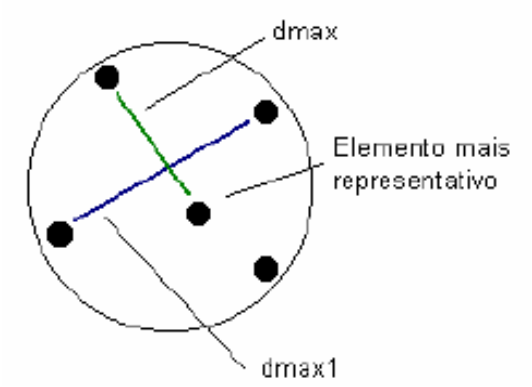


Figura 4.13: Representación de D_{max} y D_{1max} y la señal de referencia

4.4.3 Algoritmo de reconocimiento

El algoritmo de reconocimiento consiste en la ejecución de un procedimiento que realiza la comparación entre una señal de habla desconocida con la base de datos de “palabras”. Esta comparación está dada por la ecuación 4.8

$$D_{min} = \min\{D_{final}(g(n), p_k(n) | 1 \leq k \leq (\text{tamaño de la base de datos})\} \quad (4.8)$$

De la señal $p_k(n)$ que genere el menor valor de $D_{final}(g(n), p_k(n))$ se utilizaran sus valores D_{max} y D_{1max} almacenados previamente en la etapa de entrenamiento, los cuales serán utilizados como umbrales de reconocimiento tal y como se describe en las ecuaciones 4.9 y 4.10.

$$D_{min} \leq A * D_{max} \qquad (4.9)$$

$$D_{min} \leq A * D_{1max} \qquad (4.10)$$

Si D_{min} satisface las desigualdades de las ecuaciones 4.9 y 4.10 se dice que $g(n)$ se encuentra almacenada en la base de datos de palabras y es igual a $p_k(n)$, entonces el sistema arroja la cadena de caracteres relacionada a dicha señal. En el caso de que no sean satisfechas las desigualdades el sistema arroja la cadena “Palabra no almacenada en diccionario”.

El valor de A es una constante obtenida mediante experimentación. Siendo el valor optimo de A, el que genere un porcentaje mayor de aciertos (17).

4.4.5 Diseño de Pruebas

Todos los Pruebas fueron hechos sobre una aplicación desarrollada en Matlab y están basados sobre el reconocimiento de dígitos del cero al nueve. Cada digito será pronunciado 10 veces; Dando un total de 100 palabras a reconocer.

CAPÍTULO 5

5. Pruebas y Resultados

En el presente capitulo se muestran las pruebas realizadas sobre el sistema de reconocimiento y los resultados obtenidos. Existen varios parámetros que pueden ser escogidos para evaluar el sistema de reconocimiento, por ejemplo: El umbral de reconocimiento, Tamaño de la trama usada en la segmentación de la palabra y funcionamiento en ambientes con ruido.

5.1 Pruebas realizadas

Experimento 1: Variación del umbral de reconocimiento

El experimento de variación del umbral de reconocimiento consiste en variar A (factor de escala) relacionado con los umbrales D_{max} y D_{1max} para dar como resultado un desplazamiento del umbral de reconocimiento y en consecuencia un numero de aciertos y desaciertos capaces de ser medidos a manera de porcentajes. Los resultados obtenidos se muestran en la figura 5.1 y 5.2.

Los valores de A varían entre: {1, 1.1, 1.25, 1.3, 1.4, 1.5, 1.63, 1.69, 1.7, 1.76, 1.78, 1.8, 1.85, 1.9, 2}

El tamaño de la trama es 20 ms.

Experimento 2: Variación del tamaño de la trama

El experimento de variación del tamaño de trama consiste en calcular el número de aciertos y desaciertos producidos al variar el tamaño de la trama. Los resultados obtenidos se muestran en la figura 5.3 y 5.4.

Los valores de tamaño de trama son: {10 ms, 15 ms, 20 ms, 25 ms, 30 ms}

El valor de A es 1.3

Experimento 3: Reconocimiento en ambiente de ruido

El experimento de reconocimiento en ambiente de ruido consiste en calcular el número de aciertos y desaciertos tanto en ambiente controlado como en ambiente con ruido de fondo. Los resultados obtenidos se muestran en la figura 5.5.

El tamaño de la trama es: 20 ms

El valor de A es 1.3

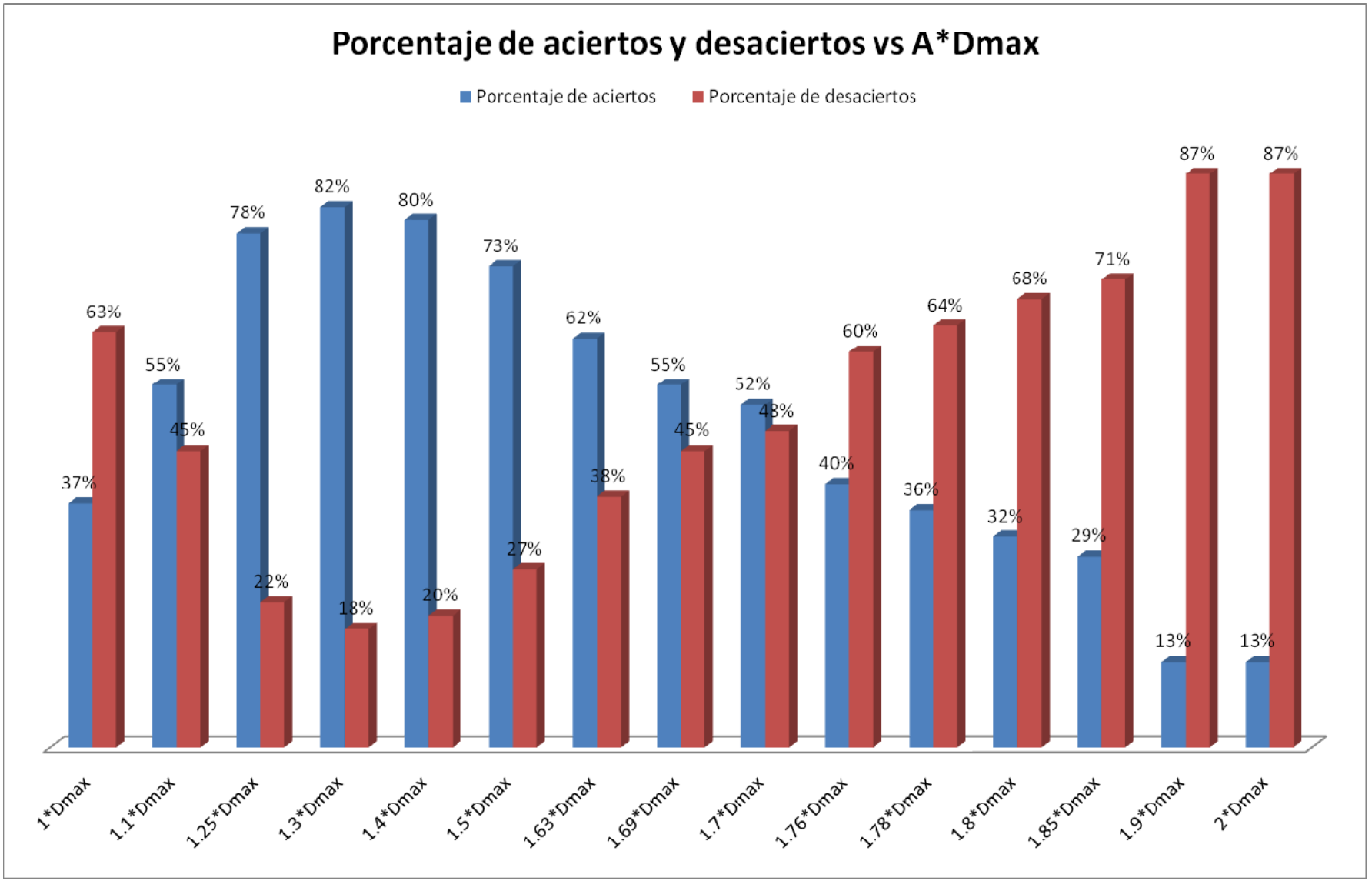


Figura 5.1: Gráfica de Porcentaje de aciertos y desaciertos versus A*Dmax

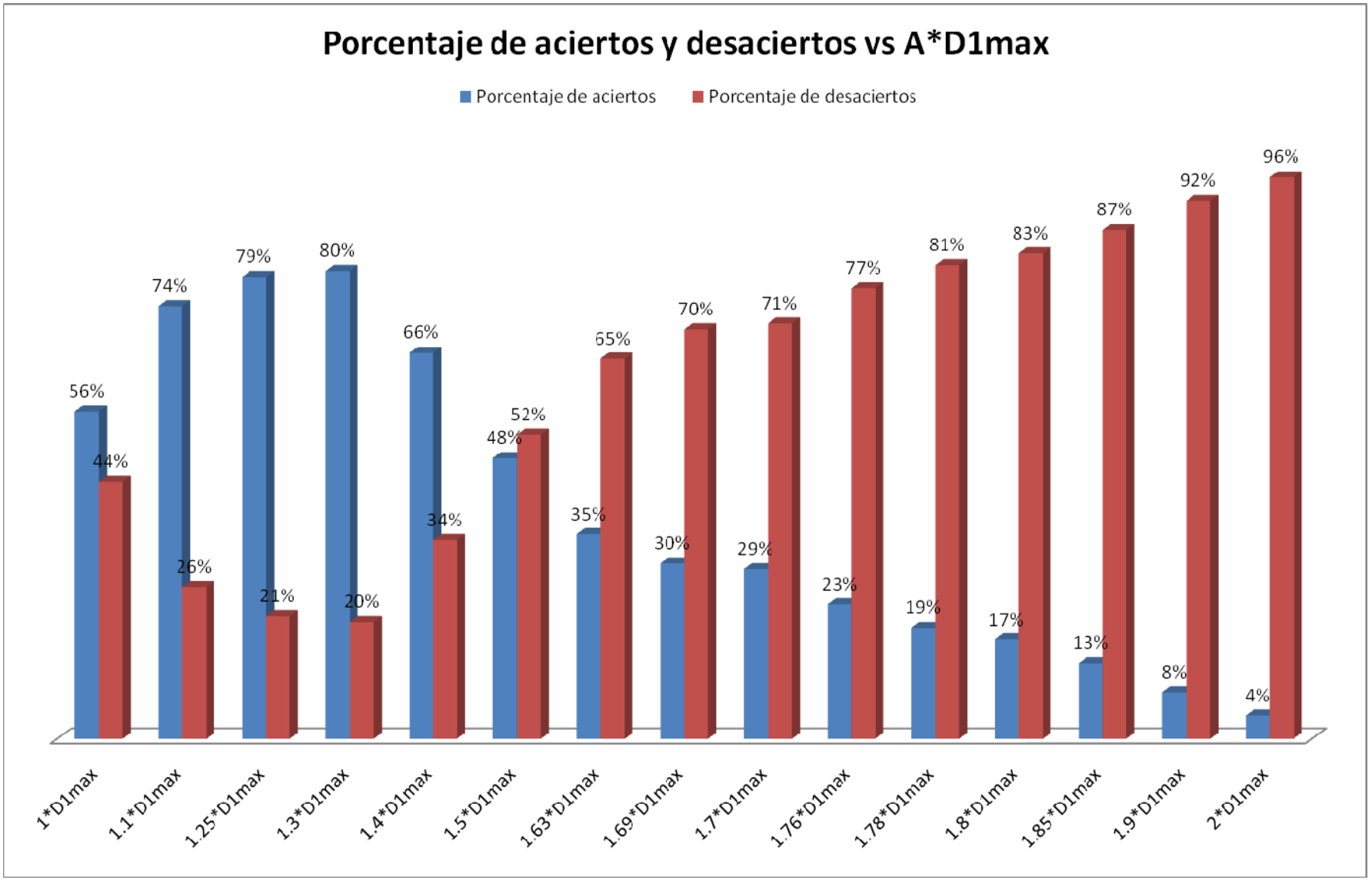


Figura 5.2: Gráfica de Porcentaje de aciertos y desaciertos versus A*D1max

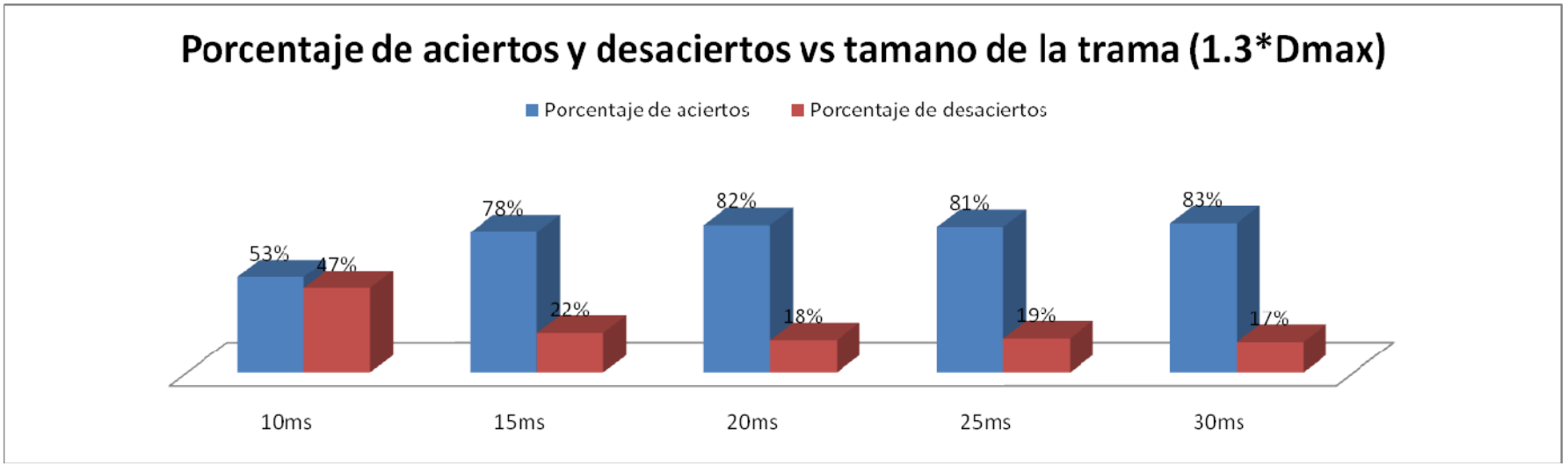


Figura 5.3: Gráfica de Porcentaje de aciertos y desaciertos versus 1.3*D1max

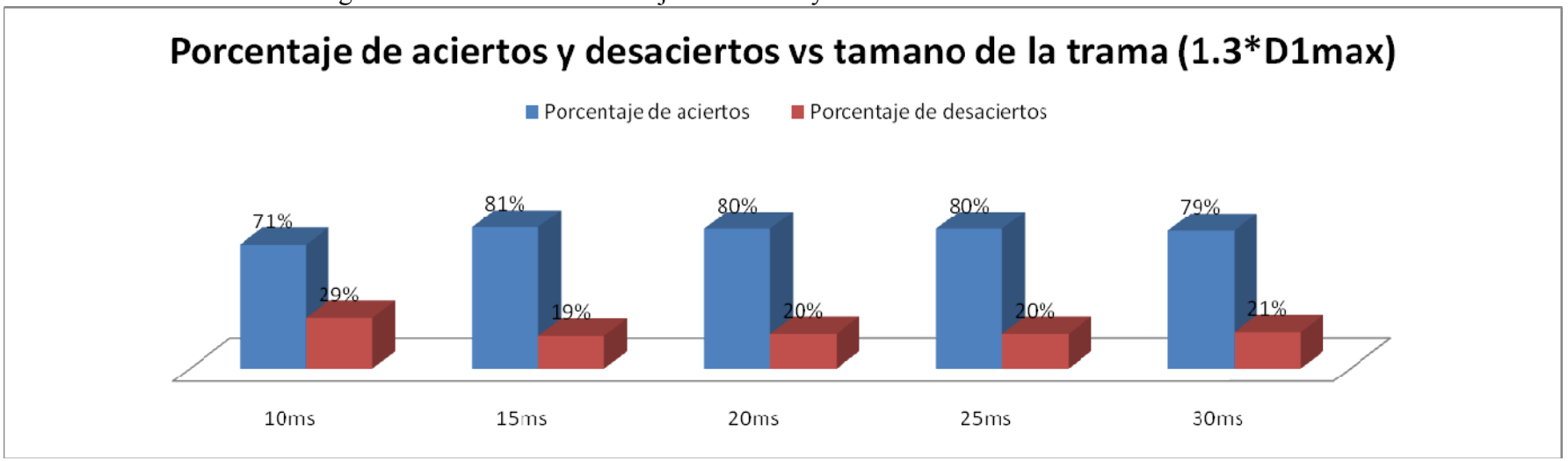


Figura 5.4: Gráfica de Porcentaje de aciertos y desaciertos versus 1.3*D1max

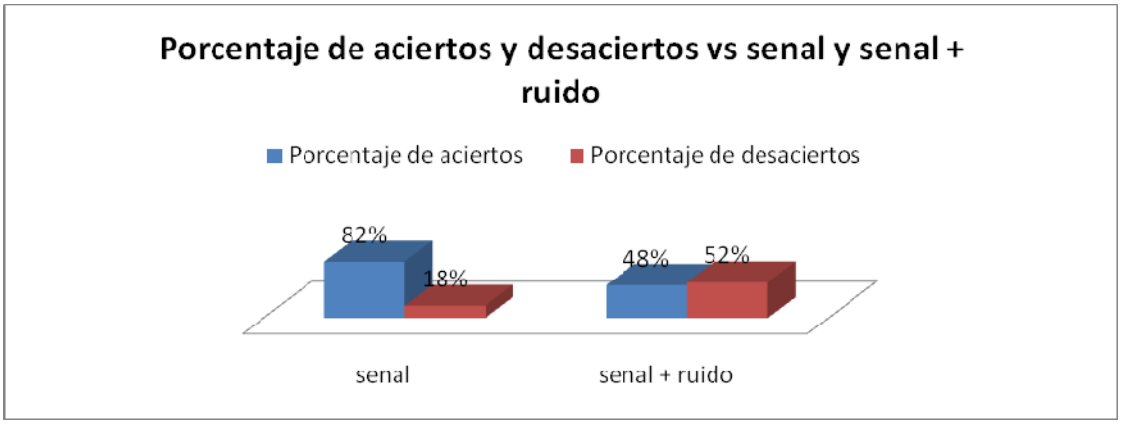


Figura 5.5: Gráfica de porcentaje de aciertos y desaciertos versus señal y señal + ruido

5.2 Análisis de Resultados

Experimento 1: Variación del umbral de reconocimiento

Según la figura 5.1 al utilizar en el sistema de reconocimiento del habla el umbral de reconocimiento $A * D_{max}$, variando el valor de A obtenemos que el valor óptimo para este experimento es de 1.3, Con un porcentaje de acierto de 82%.

Según la figura 5.2 al utilizar en el sistema de reconocimiento del habla el umbral de reconocimiento $A * D_{1max}$, variando el valor de A obtenemos que el valor óptimo para este experimento es de 1.3, Con un porcentaje de acierto de 80%.

Experimento 2: Variación del tamaño de la trama

Según la figura 5.3 al utilizar en el sistema de reconocimiento del habla un tamaño de trama variable, y un umbral igual $1.3 * D_{max}$ obtenemos que el valor óptimo del tamaño de trama para este experimento es de 30 ms. Con un porcentaje de acierto de 83%.

Según la figura 5.4 al utilizar en el sistema de reconocimiento del habla un tamaño de trama variable, y un umbral igual $1.3 * D_{1max}$ obtenemos que el valor óptimo del tamaño de trama para este experimento es de 15ms. Con un porcentaje de acierto de 81%.

Experimento 3: Reconocimiento en ambiente de ruido

Según la figura 5.5 al evaluar al sistema de reconocimiento del habla en ambiente con ruido fondo con datos de umbral de reconocimiento igual a $1.3 * D_{max}$ y 20 ms de tamaño de trama se obtiene que el sistema evaluado en ruido da un porcentaje de aciertos del 48% siendo esto muy por debajo del porcentaje de aciertos en condiciones contraladas que es 82%. Esto se debe a que el sistema no posee una etapa de reducción de ruido.

Conclusiones

1. Los experimentos realizados evaluaron al sistema de reconocimiento del habla en las los bloques de extracción de características y comparación de parámetros o reconocimiento. Dando como valores óptimos de A y tamaño de trama de 1.3 y 30 ms segundos respectivamente con un porcentaje de acierto del 83%.

2. El porcentaje de aciertos obtenido en sistema de reconocimiento es similar a otros proyectos realizados por otros departamentos de investigación. Como es el caso del proyecto “Reconocimiento de dígitos usando DTW” de la Universidad Federal de Rio de Janeiro [18]. Que posee un porcentaje de aciertos del 83%.

3. El sistema posee poca tolerancia frente al ruido con un porcentaje de aciertos del 48% esto se debe a que no posee ningún algoritmo de reducción de ruido de fondo. Este ruido influye principalmente en la detección del inicio y fin de señal de habla dando como resultado un mal almacenamiento de la señal a procesar y por ende resultados erróneos.

4. La herramienta Matlab fue muy útil en todo el proceso de desarrollo del sistema; dado que posee una interfaz grafica amigable en la presentación de señales. Por lo tanto se recomienda para cualquier diseño de sistemas de procesamiento de señales la utilización de este software.

Recomendaciones

1. Implementar un algoritmo para la eliminación de ruidos de fondo, para evitar resultados erróneos en el caso de ambientes ruidosos.

2. Pronunciamiento de las palabras pausado y claro para una mejor captación de la palabra por parte de la aplicación.

ANEXOS

ANEXO 1

Código utilizado

Algoritmo de extracción de características

```
function r = mfcc2(s, fs,tam)

x1 = 100/256;

tam2 = x1*tam;

n = floor((tam/1000)/(1/fs));

m = floor((tam2/1000)/(1/fs));
```

```
l = length(s);
nbFrame = floor((l - n) / m) + 1;
for i = 1:n
for j = 1:nbFrame
M(i, j) = s(((j - 1) * m) + i);
end
end
s2 = s'
h = hamming(n);
M2 = diag(h) * M;
for i = 1:nbFrame
frame(:,i) = fft(M2(:, i));
end
t = n / 2;
tmax = l / fs;
m = melfb(20, n, fs);
n2 = 1 + floor(n / 2);
z = m * abs(frame(1:n2, :)).^2;
r = dct(log(z));
```

Algoritmo de alineamiento temporal dinámico DTW

```
function dist2 = dtw2(c1,c2);
c1(1,:) = [];
c2(1,:) = [];
c1 = c1(1:9,:);
```

```
c2 = c2(1:9,:);
d = disteu(c1, c2);
D=zeros(size(d));
[M,N] = size(D)
D(1,1)=d(1,1);
for m=2:M
    D(m,1)=d(m,1)+D(m-1,1);
end
for n=2:N
    D(1,n)=d(1,n)+D(1,n-1);
end
for m=2:M
    for n=2:N
        D(m,n)=d(m,n)+min(D(m-1,n),min(D(m-1,n-1),D(m,n-1)));
    end
end
Dist=D(M,N);
n=N;
m=M;
k=1;
w=[M N];
while ((n+m)~=2)
    if (n-1)==0
        m=m-1;
    elseif (m-1)==0
```

```
        n=n-1;
      else
        [values,number]=min([D(m-1,n),D(m,n-1),D(m-1,n-1)]);
        switch number
        case 1
          m=m-1;
        case 2
          n=n-1;
        case 3
          m=m-1;
          n=n-1;
        end
    end
      k=k+1;
      w=[m n; w];
end
tam = size(w);
dist2 = Dist/tam(1);
```

Algoritmo de detección de actividad de voz

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%    Funcion 'Detector'                    %
%    Se encarga de realizar la deteccion    %
%    del principio y final de la palabra    %
%    de voz. Es muy importante que          %
%    inicialicemos el detector antes de     %
```

```
%    comenzar la deteccion de una palabra,  %
%    para ello utilizaremos la funcion      %
%    anterior.                              %
%    trama : numero de trama que es.        %
%    energia_trama : energia de la trama    %
%     correspondiente.                      %
%    linea : numero de la linea.            %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [ini,fin,E]=detector(senal, fs ,tramas_ruido ,long_win ,offset)
 [E,Eruido,total_tramas]= energia(senal, fs, tramas_ruido, long_win, offset);
%(total_tramas, energia, fs, offset, EnergiaInicial, DEBUG)
fprintf('Eruido : %d\n',Eruido);
EnergiaInicial = Eruido;
% Nivel low del detector
factor_low = 8; % dB (3)
% Nivel speech del detector
factor_speech = 16; % dB (15)
% Nivel high del detector
factor_high = 20; % dB (30)
% Tiempo mínimo por encima de SPEECH
ms80 = 60; % ms.
% Tiempo máximo por debajo de SPEECH
ms180 = 180; % ms.
TMaximoEncimaSpeech = ms180;
% Frecuencia de muestreo.
```

```
Frekhz = fs/1000; % khz. (8)
% Avance de trama.
SepTramas = offset; % tramas (80)
% Mínima duración de palabra.
minima_duracion_pal = 15; % tramas.
% Máxima duración click.
minimo_anchura_acept = 5; % muestras.
% Número de tramas iniciales en las que estimamos el ruido.
VentanaRuido = 10; % tramas.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%    Funcion Inicializa_Detector           %
%    Se encarga de inicializar todas las    %
%    variables necearias en la deteccion    %
%    de una palabra mediante la energia     %
%    de cada trama.                         %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
ini = 0;
fin = 0;
hay_palabra = 0;
fin_palabra = 0;
supera_speech = 0;
mantiene_speech = 0;
supera_high = 0;
energia_anterior = -50;%0.0;
energia_ruido = EnergiaInicial;
```

```
energia_low = energia_ruido + factor_low;
energia_speech = energia_ruido + factor_speech;
energia_high = energia_ruido + factor_high;
num_tramas_ruido = 0;
cuenta_debajo_speech = 0;
cuenta_arriba_speech = 0;
Inicio_Posible = 0;
inicio_detectado = 0;
fin_detectado = 0;
Posible_fin = 0;
anchura_click = 0;
maxima_duracion = ms180 * Frekhz / SepTramas;
minima_duracion = ms80 * Frekhz / SepTramas;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
for trama=1:total_tramas,
   % Hacemos una estimacion del nivel de ruido de la sala para
   % calcular a partir de ahi los umbrales adecuados para la deteccion.
  energia_ruido = energia_ruido + E(trama);
   num_tramas_ruido = num_tramas_ruido + 1;
   if num_tramas_ruido == VentanaRuido
      energia_ruido = energia_ruido / num_tramas_ruido;
      if (energia_ruido < energia_low) & (hay_palabra == 0)
         % Actualizamos los niveles: solo en terminadas
         % circunstancias que son cuando no hay una
         % variacion muy rapida.
```

```
        energia_low = energia_ruido + factor_low; % 3
        energia_speech = energia_ruido + factor_speech; % 10
        energia_high = energia_ruido + factor_high; % 25
      end
     % Cada vez que llegamos a VENTANARUIDO hacemos una
     % estimacion e inicializamos las variables
     % utilizadas.
     num_tramas_ruido = 0;
    energia_ruido = 0.0;
end
% Posible inicio: superamos el nivel LOW pero
% todavia no hemos llegado al HIGH (primera
% condicion de inicio).
if (E(trama) > energia_low) & (energia_anterior <= energia_low) & (supera_speech == 0)
    inicio_detectado = trama;
    Inicio_Posible = 1;
     fprintf(' inicio posible: %d \n',trama);
end

% Si bajamos de LOW sin haber llegado a HIGH
% entonces un posible inicio anterior, no era tal.
if (E(trama) <= energia_low) & (energia_anterior > energia_low) & (supera_speech == 0)
    inicio_detectado = 0;
     fprintf(' no era inicio: %d \n',trama);
end
```

```
% Hemos superado el nivel SPEECH por primera vez.
if (E(trama) > energia_speech) & (energia_anterior <= energia_speech)
   % Vamos a llevar la cuenta de las veces que le superamos.
   fprintf(' sobre el speech: %d \n',trama);
   cuenta_debajo_speech = 0;
   if supera_speech == 0
      cuenta_arriba_speech = 0;
      supera_speech = 1;
   end
end

% Hemos cumplido la segunda condicion para un
% inicio de palabra que es la de mantenernos por
% encima de SPEECH mas de minima_duracion.

if (cuenta_arriba_speech == minima_duracion)
  yw = 3;
end

if (E(trama) > energia_speech) & (cuenta_arriba_speech == minima_duracion)
    fprintf(' paso el minimo valor: %d \n',trama);
   mantiene_speech = 1;
end
```

```
cuenta_arriba_speech = cuenta_arriba_speech + 1;

% Si superamos el nivel HIGH hemos cumplido la
% tercera de las condiciones para un comienzo
% de palabra.
if E(trama) > energia_high
    supera_high = 1;
    % AÑADIDO
    fprintf('  supero high   %d\n',trama);
    hay_palabra = 1;
end

% Consideramos un posible final de trama al bajar
% del nivel LOW (primera condicion de final).
if (E(trama) <= energia_low) & (energia_anterior > energia_low) & (supera_speech == 1)

    Posible_fin = trama;
    if mantiene_speech == 0
        anchura_click = trama - inicio_detectado;
         fprintf(' eera click: %d    y supero high   %d\n',trama, supera_high);
        % calculo anchura de un posible click inicial

        % Verificamos si lo detectado es o no un click.
        if anchura_click < minimo_anchura_acept
        % no hay palabra, es click
```

```
            cuenta_debajo_speech = 0;
            cuenta_arriba_speech = 0;
            supera_speech = 0;
            mantiene_speech = 0;
            supera_high = 0;
            hay_palabra = 0;
        end
    end
end

% Bajamos de nivel SPEECH.
if (E(trama) <= energia_speech) & (energia_anterior > energia_speech) & (supera_speech ==
1)
    cuenta_arriba_speech = 0;
     fprintf(' bajo del speech y borro cuenta arriba   %d  %d\n',trama );
end

% Si seguimos debajo del SPEECH.
if (E(trama) <= energia_speech) & (supera_speech == 1)
    % Si seguimos debajo durante maxima_duracion se
    % cumplira la segunda condicion de posible final de
    % palabra.
    fprintf(' bajo del speech   %d  %d\n',trama, energia_speech );

    if cuenta_debajo_speech == maxima_duracion
```

```
        cuenta_debajo_speech = 0;
        cuenta_arriba_speech = 0;
        supera_speech = 0;
        mantiene_speech = 0;
        supera_high = 0;
        fin_detectado = Posible_fin;

        tramas_pal = fin_detectado - inicio_detectado;

        % Si la palabra es suficientemente grande entonces
        % realmente hemos detectado una palabra.
        if (tramas_pal > minima_duracion_pal) & (hay_palabra == 1)
          fin_palabra = 1;
            ini = inicio_detectado;
            fin = fin_detectado;
        end

        % Ya no hay palabra porque la hemos terminado.
        hay_palabra = 0;
    end
    cuenta_debajo_speech = cuenta_debajo_speech + 1;
end
% Ratificamos un posible inicio calculado
% anteriormente.
if (supera_high == 1) & (mantiene_speech == 1)
```

```
        hay_palabra = 1;
        if fin == 0 % No se ha detectado un final con anterioridad.
            ini = inicio_detectado;
        end
    end
    energia_anterior = E(trama);
end
    fprintf('INI : %d\n',ini);
    fprintf('FIN : %d\n',fin);
    [Z,ZCC_ruido,total_tramas]= tasacruces (senal, fs, tramas_ruido, long_win, offset);
    figure
subplot(5,1,2);
plot(Z);
f = 4
valor = 25
iniz = 0;
finz = 0;
y = var(Z(1:10));
x = max(Z(1:10));
umbralz = x + y*f
if fin ~= 0
    for i = 4 : valor
        if ( Z(ini - i) < umbralz)
            iniz = ini - i;
            break;
```

```
        end;
    end;
    if (iniz == 0)
      iniz = ini;
    else
        if (iniz == ini -4)
            iniz = ini;
        end;
    end;
    for i = 4 : valor
        if ( Z(fin + i) < umbralz)
            finz = fin + i;
            break;
        end;
    end;
    if (finz == 0)
        finz = fin;
    else
        if (finz == fin +4)
            finz = fin;
        end;
    end;
end;
ini = iniz;
fin = finz;
```

ANEXO 2

Manual de usuario

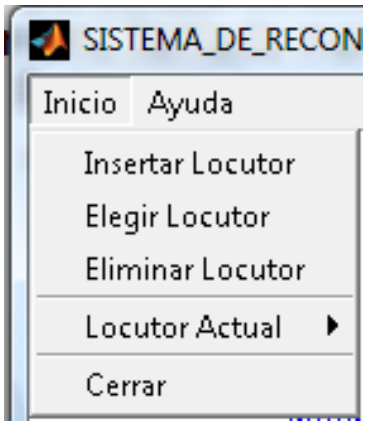
1. Pantalla principal:

El primer paso en el manual del usuario es la explicación de la pantalla principal. La pantalla principal consta de dos Menús en la parte superior las cuales son: Inicio y Ayuda. Y un menú llamado menú ASR (Automatic speech recognition) que contiene dos botones: Entrenar y reconocer. En la figura A2.1 se muestra el menú principal del sistema de reconocimiento del habla.

Figura A2.1: Menú principal del sistema de reconocimiento del habla.

Menú Inicio:

Al presionar el menú inicio se presentan en la pantalla cinco opciones o sub menús. Estas opciones están relacionadas a la creación de una base de datos de palabras. Puesto que el sistema de reconocimiento es monolocutor cada base de datos debe pertenecer a un mismo locutor. Las diferentes opciones a realizar son Insertar locutor, elegir locutor, eliminar locutor, locutor actual y salir. En la figura A2.2 se muestra el menú de opciones del menú inicio.



. Figura A2.2: Menú de opciones de la pestaña de inicio.

Insertar locutor:

Al presionar insertar locutor contiene se abre una ventana llamada "INSERTAR_LOCUTORES", la cual contiene una caja de texto, En la cual se debe colocar el nombre al cual se desea llamar la base de datos de palabras. Al realizar el correcto nombramiento de la base de datos es necesario presionar el botón aceptar para la creación de una base de datos originalmente vacía. Obsérvese la figura A2.3 para mayor información sobre la ventana de insertar locutor.

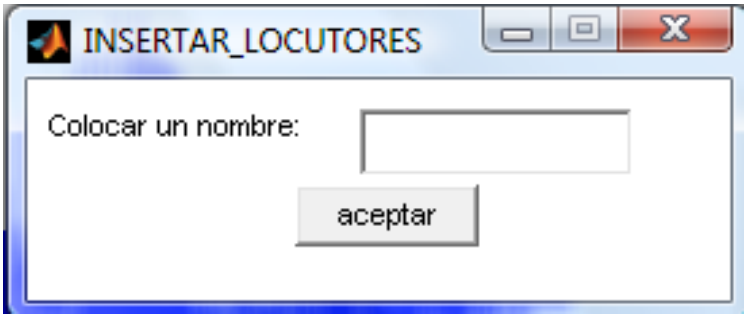


Figura A2.3: Ventana Insertar locutor.

Elegir locutor:

Al presionar elegir locutor se presenta una ventana llamada "ELEGIR_LOCUTORES", la cual contiene una lista que refleja todas las bases de datos almacenadas en el sistema de reconocimiento. En el caso de la figura A2.4 se observa que existe una única base de datos llamada Manuel.

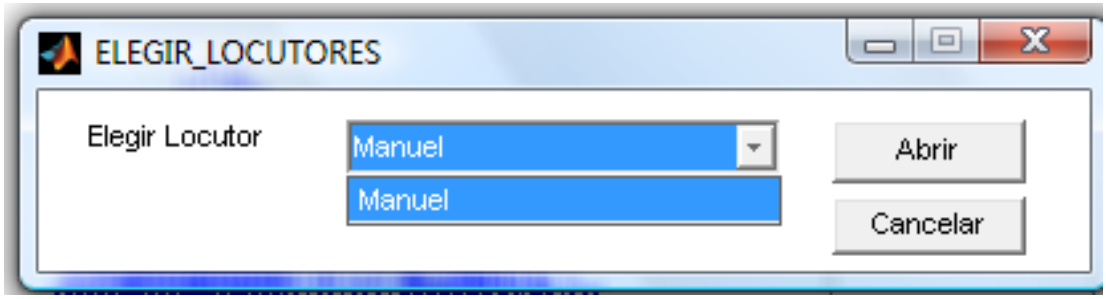


Figura A2.4: Ventana Elegir locutor.

Para seleccionar una base de datos es necesario la selección en la lista y presionar el botón abrir. Esto generará su selección y se visualizara en la pantalla principal bajo la sub pestaña Locutor actual. Véase la figura A2.5 para visualizar esta selección.

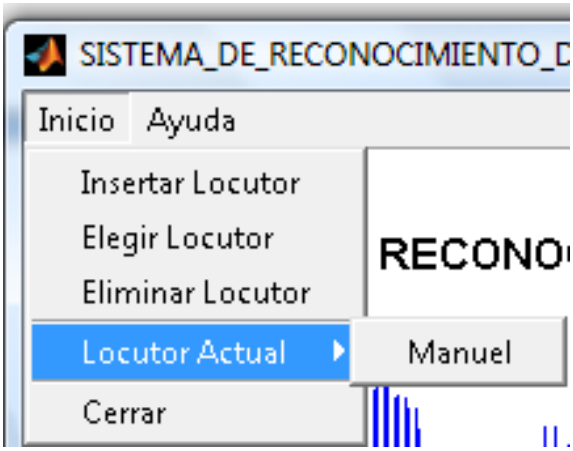


Figura A2.5: Visualización de locutor actual.

Eliminar locutor:

Al presionar eliminar locutor se presenta una ventana llamada "ELIMINAR_LOCUTORES", la cual contiene una lista que refleja todas las bases de datos almacenadas en el sistema de reconocimiento. El proceso de borrado es el mismo que selección del locutor, pero el resultado final es la eliminación de la base de datos de palabras seleccionada. En el caso de la figura A2.6 se observa que existe una única base de datos llamada Manuel.

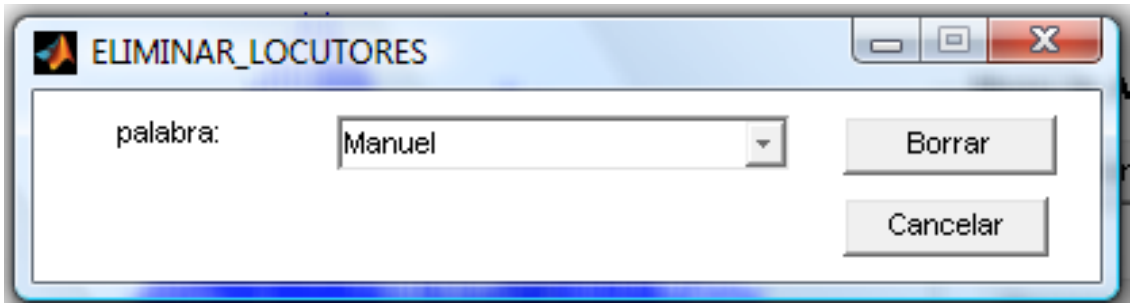


Figura A2.6: Ventana eliminar locutor.

Menú de ASR:

El menú de ASR contiene dos botones: Entrenar y reconocer. En la figura A2.7 se muestra el menú ASR.

Figura A2.7: Menú de ASR

Entrenar:

Al presionar el botón entrenar se abre una ventana llamada “TRAINING” esta ventana contiene dos menús en la parte superior y un menú con 3 botones en la parte lateral derecha. En la figura A2.8 se muestra la ventana “TRAINING”.

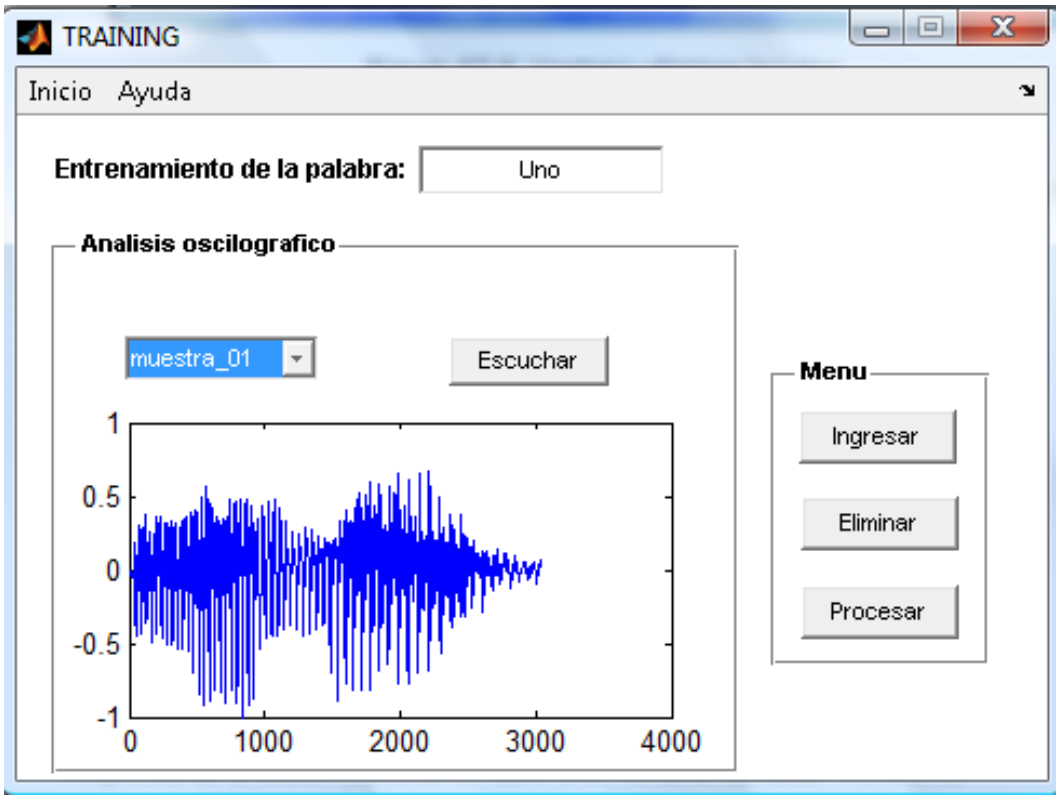


Figura A2.8: Ventana “TRAINING”

Procedimiento de entrenamiento:

Hay que recordar que en la ventana principal se ha seleccionado previamente una base de datos de palabras. Por lo tanto en la ventana de entrenamiento llamada “TRAINING” todas las palabras estarán almacenadas en la base de datos seleccionada.

1. Iniciar Entrenamiento: Para iniciar el entrenamiento hay que eliminar todas las palabras contenidas en una lista sobre la grafica. En dicha lista se encuentran todas las señales ingresadas bajo el nombre de muestras. Luego

hay que ir al menú de inicio de la ventana de entrenamiento y seleccionar la opción iniciar entrenamiento esto eliminará de manera automática todas las muestras contenidas en la lista. En la figura A2.9 se muestra la opción del menú de inicio “Iniciar entrenamiento”.

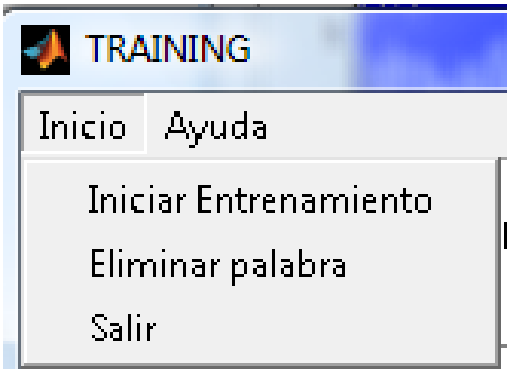


Figura A2.9: Menú de inicio en la ventana “TRAINING”.

2. Asignar un nombre a la palabra: Al colocar un nombre en el cuadro de texto todas las señales de hablas pronunciadas por el locutor deben ser la misma. Al no ser así deben de ser eliminadas o de lo contrario el entrenamiento de dicha palabra será erróneo. En la figura A2.10 se muestra en cuadro de texto antes mencionado.

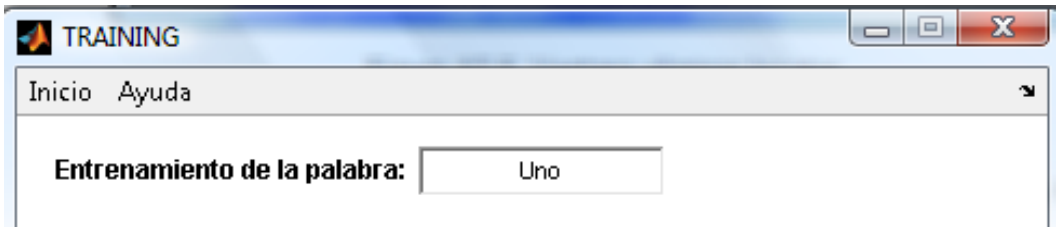


Figura A2.10: Etapa de asignación de nombre de la palabra a entrenar.

3. Ingresar señales de habla: La etapa de entrenamiento consiste en ingresar 5 pronunciaciones de una misma palabra y luego procesarlas para seleccionar de entre ellas la señal patrón y dos umbrales □□□ y □□□.

Cada palabra debe de ser ingresada por medio de botón “Ingresar”. Al presionar el botón ingresar el locutor escuchará un sonido generado por la computadora el cual indica que puede decir la palabra. Y 3 segundos aproximadamente escuchará otro sonido generado por la computadora que indica el fin de la palabra. Esto quiere decir que el usuario debería almacenar una palabra de duración máxima de 2.5 segundos aproximadamente para que no existan problemas de almacenamiento.

Una vez almacenada una señal esta será añadida a la lista muestras con su respectiva numeración. Para propósitos de control cada muestra presente en la

lista puede ser graficada mediante el análisis oscilográfico y escuchada presionando el botón “escuchar”. En la figura A2.11 se muestra estas opciones.

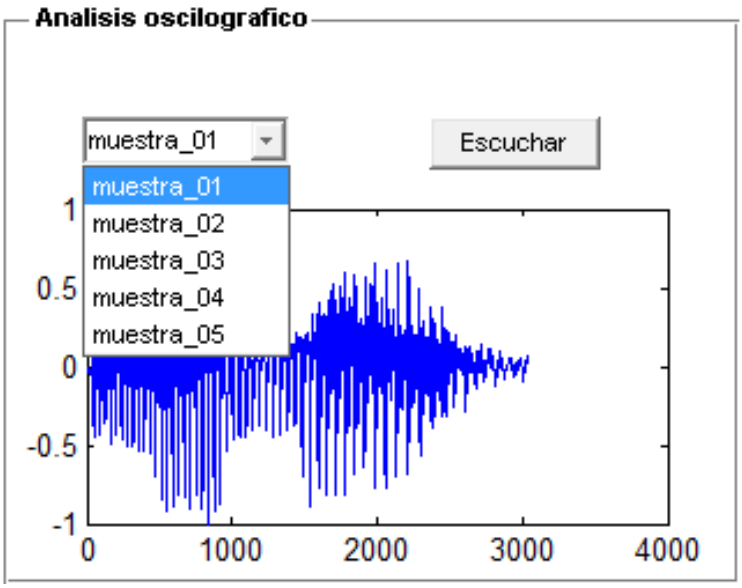


Figura A2.11: Análisis oscilográfico de las muestras.

4. Procesar: Una vez almacenadas las 5 muestras de la palabra a entrenar se debe de presionar el botón procesar presente en el menú lateral derecho. Una vez procesada la palabra y si se desea continuar con el entrenamiento de palabras se debe de iniciar el proceso de entrenamiento desde el comienzo.

Reconocer:

Al presionar el botón reconocer se generan los sonidos de inicio y fin de la palabra entre los cuales el locutor debe de pronunciar la palabra. Después de esto aparecerá un mensaje en la pantalla con el nombre de la palabra si se encuentra en la base de datos de lo contrario aparecerá el mensaje: “Palabra no almacenada en diccionario”.

BIBLIOGRAFÍA

1. J. Allen, Advances in Speech Signal Processing, editorial, Dekker, 1992, páginas 741–790..

2. F. Casacuberta and E. Vidal. Reconocimiento Automático del Habla , editorial, Marcombo, 1987.

3. Y. Ephraim, “A bayesian estimation approach for speech enhancement using hidden Markov models”,editorial,lugar, 1992, páginas 725–735.

4. S. Furui, Advances in Speech Signal Processing, editorial, Dekker, 1992, páginas 597–622.

5. Mark J.F. Gales, Model-based techniques for noise robust speech recognition. editorial, Gonville and Caius College, September 1995.

6. María Isabel Calonge Ramírez, http://www.gobiernodecanarias.org/ educacion/tamadaba/tama4/voz.htm

7. James Glass y Victor Zue, MITOPENCOURSEWARE Massachusetts Institute of Technology, http://ocw.mit.edu/NR/rdonlyres/Electrical-Engineering-and-Computer-Science/6-345Automatic-Speech-RecognitionSpring2003/8C754156-CFB0-438F-9006-A67996A7EBD9/0/lecture2.pdf

8.AbelHerrera,http://microsoft.fib.unam.mx/PaginasProfesores/AbelHerrera/Pro_dig_audio/frames.html

9. Vilela Aguirre, http://cta-2992.blogspot.com/

10. Paul Contreras, www.vi.cl/foro/index.php?showtopic=8205

11. Javier Macías Guarasa, Sistema de producción de habla, editorial, Madrid, , 2006

12. Bernal, J., Bobadilla, J. Y Gómez, P. ;Reconocimiento de Voz y Fonética acústica, editorial, México, 2000

13. González Rodríguez, Joaquín; Ortega García, Javier, Procesado de Voz: Reconocimiento de mensaje y locutor; editorial; Madrid; España; 1997.

14. Pardo, J. M.; Sistema de producción del habla, editorial, España, 2002

15. IFEACHOR, Emmanuel y jERVIS, Barrie. Digital signals processing: A practical approach.Cap.1.Addison-Wesley(Eds.),1993.

16. L. G. Johansen, P. Rubak, “Investigating Speech Quality by Homomorphic Deconvolution”, editorial, Praga, 1997

17. J. J. Kim, M. Bae, “On a Modified Cepstral Pitch Control Technique for the High Quality Text-to-Speech Type System”, editorial, Indiana, 1998.

18. Amaro Azevedo, Reconhecimento de dígitos usando DTW, editorial, Rio De Janeiro, 2007.