INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

ESCOM

Trabajo Terminal

Sistema De Reconocimiento De Palabras Clave En Conversaciones De Voz.

2014-A004

Presentan

Juan Becerra Becerra

Diego Alberto Farías Pineda

Víctor Martínez Sánchez

Directores

M.C. Rubén Hernández Tovar

M.C. Víctor Hugo García Ortega

julio de 2015

INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

No. de Registro: 2014-A004 Serie: Amarilla Julio 2015

Documento Técnico

Sistema De Reconocimiento De Palabras Clave En Conversaciones De Voz.

Autores:

Juan Becerra Becerra[1]

Diego Alberto Farías Pineda[2]

Víctor Martínez Sánchez[3]

Directores:

M.C. Víctor Hugo García Ortega **M.C. Rubén Hernández Tovar**

Resumen: En este trabajo se propone el desarrollo de un prototipo que sea capaz de reconocer palabras clave dentro de conversaciones de voz, esto con el fin de recabar información que se considere importante y que pueda ser utilizada para el desarrollo de otras aplicaciones. Para lograrlo se planea hacer uso de técnicas de procesamiento digital de señales en el preprocesamiento de la señal digital de voz, redes neuronales como técnica de clasificación de los vectores de características de la señal de voz y reconocimiento de patrones con el fin lograr la correcta identificación de las palabras emitidas dentro de la conversación de voz.

Palabras clave: Procesamiento Digital De Señales, Procesamiento Digital De Voz, Reconocimiento De Patrones, Reconocimiento de Voz, Redes Neuronales.

[1] juanbecerrab@gmail.com
[2] diego@ideaslibres.net
[3] mvictor619@gmail.com

ESCUELA SUPERIOR DE CÓMPUTO
SUBDIRECCIÓN ACADÉMICA
DEPARTAMENTO DE FORMACIÓN INTEGRAL E INSTITUCIONAL
COMISIÓN ACADÉMICA DE TRABAJO TERMINAL

México, D.F. a 1 de julio de 2015

Dr. FLAVIO ARTURO SÁNCHEZ GARFIAS
PRESIDENTE DE LA COMISIÓN ACADÉMICA
DE TRABAJO TERMINAL
PRESENTE

Por medio del presente se informa que los alumnos que integran el **TRABAJO TERMINAL**: 2014-A004, titulado "Sistema De Reconocimiento De Palabras Clave En Conversaciones De Voz" concluyeron satisfactoriamente su trabajo.

Los discos (DVDs) fueron revisados ampliamente sus servidores y corregidos, cubriendo el alcance y objetivo planteados en el protocolo original y de acuerdo a los requisitos establecidos por la Comisión que Usted preside.

ATENTAMENTE

M.C. Victor Hugo García Ortega

M.C. Rubén Hernández Tovar

4

Advertencia

"Este documento contiene información desarrollada por la Escuela Superior de Cómputo del Instituto Politécnico Nacional, a partir de datos y documentos con derecho de propiedad y por lo tanto, su uso quedará restringido a las aplicaciones que explícitamente se convengan."

La aplicación no convenida exime a la escuela su responsabilidad técnica y da lugar a las consecuencias legales que para tal efecto se determinen.

Información adicional sobre este reporte técnico podrá obtenerse en:

La Subdirección Académica de la Escuela Superior de Cómputo del Instituto Politécnico Nacional, situada en Av. Juan de Dios Bátiz s/n Teléfono: 57296000, extensión 52000.

Agradecimientos

Juan Becerra Becerra: Aprovecho este espacio para darle las gracias a todas las personas que a lo largo de estos años contribuyeron en que yo culminara esta meta en mi vida, a mi padre porque fue el único que siempre creyó y tuvo confianza en mí, a mi madre por preocuparse por mi todos los días, a mis hermanos por su apoyo incondicional y muchos ratos de diversión, a mis amigos por los momentos divertidos que pasamos juntos y que hicieron más amenas la estancia en la escuela, a los profesores por su dedicación y entrega y al IPN por darme la oportunidad de hacer realidad un sueño que comenzó desde aquel lejano agosto de 2007 cuando ingrese a la vocacional.

Diego Alberto Farías Pineda: Quiero agradecer principalmente a mis padres Beatriz y Javier, mi familia y mis amigos. A cada persona dentro y fuera de la escuela que me dejó algún aprendizaje, a mis compañeros Juan y Víctor y a todos aquellos que hacen posible la educación pública en México. Millones de gracias.

Víctor Martínez Sánchez: Quiero agradecer a mi familia por el apoyo que me han brindado estos años, en especial a mis padres Froylan y Laura por ser un gran ejemplo a seguir. a los maestros y compañeros que me ayudaron a desarrollarme profesionalmente, a mis compañeros Juan y Diego por trabajar arduamente, y por ultimo a mis directores Víctor Hugo y Rubén por compartirnos su conocimiento, gracias.

Abstract

In this paper the development of a prototype able to recognize keywords within voice conversations is proposed that in order to gather information that may be relevant and can be used to develop other applications. To achieve this it is planned to use techniques of digital signal processing in the preprocessing of digital voice signal, neural networks as a technique for classifying the feature vectors of the speech signal and pattern recognition in order to achieve the correct identification words issued within the voice conversation.

Contenido

ADVERTENCIA **4**

AGRADECIMIENTOS **5**

ABSTRACT **6**

CONTENIDO **7**

ÍNDICE DE ECUACIONES **9**

ÍNDICE DE IMÁGENES **10**

ÍNDICE DE TABLAS **11**

1. INTRODUCCIÓN **12**

1.1 Panorama general 12
1.2 Sistemas similares que se han desarrollado 12
1.3 Justificación 13
1.4 Objetivo 13
1.5 Productos o resultados esperados 14
1.6 Metodología 14
1.7 Organización del documento 14

2. ESTADO DEL ARTE **16**

2.1 Introducción 16
2.2 Breve historia del reconocimiento de voz por computadora 16
2.2.1 Antecedentes 16
2.2.2 Los inicios: años 50 17
2.2.3 Los fundamentos: años 60 – Comienzo en Japón (NEC labs) 17
2.2.4 Las primeras soluciones: años 70 - El mundo probabilístico. 17
2.2.5 Reconocimiento del Habla Continua: años 80 18
2.2.6 Empieza el negocio: años 90 - Primeras aplicaciones: ordenadores y procesadores baratos y rápidos 18
2.2.7 Una realidad: años 2000 - Integración en el Sistema Operativo 18
2.3 Algunas publicaciones relevantes en reconocimiento de voz 18
2.4 Parámetros característicos de una señal de voz 19
2.5 Modelos para el reconocimiento de voz 19
2.6 Aspectos a destacar en un sistema de reconocimiento de voz 19
2.6.1 Tamaño del vocabulario y confusión 20
2.6.2 Sistemas dependientes e independientes del locutor 20
2.6.3 Voz aislada, discontinua y continua 20
2.6.4 Voz aplicada a tareas o en general 21
2.6.5 Voz leída o espontánea 21
2.6.6 Condiciones adversas 21

3. MARCO TEÓRICO **22**

3.1 Esquema a bloques de un sistema de reconocimiento de voz por computadora 22
3.2 Metodología genérica empleada en un reconocimiento de voz 22
3.3 El Corpus de Voces 22
3.3.1 Condiciones para obtener un corpus de calidad 22
3.4 Preprocesamiento de la señal de voz 23
3.4.1 Preénfasis 23
3.4.2 Ventaneo 24

3.5 Detectar inicio y fin de palabra 24
3.6 Parámetros característicos de una señal de voz 26
3.6.1 Cepstrum 26
3.6.2 LPC 28
3.6.3 MFCC 28
3.7 Etapa de reconocimiento 30
3.7.1 Redes neuronales 30
3.7.2 Modelos Ocultos de Markov Discretos 32

4. ANÁLISIS 34

4.1 Análisis de Riesgos 34
4.2 Soluciones a los Riesgos 35
4.3 Estudio de Factibilidad 35
4.3.1 Factibilidad Técnica 35
4.3.2 Factibilidad Económica 36
4.3.3 Factibilidad Operativa 37
4.4 Análisis de Lenguajes de Programación 38
4.4.1 GNU Octave 38
4.4.2 MATLAB 39
4.5 Análisis de Requerimientos 40
4.5.1 Requerimientos Funcionales 40
4.5.2 Requerimientos No Funcionales 40

5. DISEÑO 41

5.1 Adquisición de los archivos de voz. 41
5.2 Los parámetros característicos de voz a emplear para el reconocimiento. 41
5.3 Los modelos de reconocimiento de voz que se probarán. 42
5.4 Preprocesamiento 42
5.5 Preénfasis 44
5.6 Ventaneo 45
5.7 Detección de inicio fin. 46
5.8 Extracción de LPC y MFCC 47
5.9 Diagrama de bloques general de un sistema de reconocimiento de voz. 52

6. IMPLEMENTACIÓN Y RESULTADOS 53

6.1 Caso de Uso 53
6.2 Interfaz 54
6.3 Diagrama de Secuencia 55
6.4 Diagrama de Bloques 55
6.5 Diagrama de flujo del algoritmo de detección de inicio y fin 56
6.6 Pruebas y Resultados 57

7. CONCLUSIONES 59

8. TRABAJO A FUTURO 61

9. REFERENCIAS 62

Índice de Ecuaciones

Ecuación 1Ecuación en diferencias de filtro paso-altas. 23
Ecuación 2Transformada Z de la Ecuación 1 24
Ecuación 3Función de transferencia. 24
Ecuación 4Calculo de real cepstrum. 26

Índice de Imágenes

Imagen 1Esquema de un diagrama de bloques de un Sistema de Reconocimiento del 22
Imagen 2Ejemplo de grabaciones con amplitud: saturada, media y baja. 23
Imagen 3Ejemplo de detección de inicio y fin de palabra con cálculo de cruce de ceros y energía. 25
Imagen 4Cálculo de los coeficientes cepstrales (Cepstrum) de un segmento de señal. 26
Imagen 5Gráficas resultantes en el flujo de cálculo del Cepstrum. 27
Imagen 6Respuestas de frecuencias $|H(w)|dB$ **versus** $|H(w)|dB(p=6{,}15{,}24)$ **..... 28**
Imagen 7Algoritmo para el cálculo de los coeficientes. 29
Imagen 8Arquitectura de una red neuronal de tipo perceptrón multicapas 30
Imagen 9Estructura de un mapa auto-organizado. 31
Imagen 10Modelo de Markov 33
Imagen 11Software utilizado para grabar. 41
Imagen 12Calculo de MFCCs con nuestro software 42
Imagen 13Señal original de la palabra cero. 43
Imagen 14Señal de voz con preénfasis. 44
Imagen 15Ejemplo de ventaneo a una palabra completa. 45
Imagen 16Software con el cual se trabajó para poder determinar los umbrales. 46
Imagen 17Calculo de LPC señal original y con preénfasis. 47
Imagen 18Señal con preénfasis, normalizada y ventaneada. 48
Imagen 19Envolventes obtenidas de los LPC. 48
Imagen 20Señal original. 49
Imagen 21Señal con preénfasis. 49
Imagen 22Señal normalizada. 50
Imagen 23Señal ventaneada y lista para ser procesada. 50
Imagen 24FFT de la señal de voz. 51
Imagen 25Filtros de Mel. 51
Imagen 26MFCC de una señal de voz.7 52
Imagen 27Diagrama de bloques general de un sistema de reconocimiento de voz. 52
Imagen 28 Caso de Uso de la pantalla principal 53
Imagen 29Pantalla principal del prototipo 54
Imagen 30Diagrama de Secuencia 55
Imagen 31Diagrama de bloques del sistemas 55
Imagen 32Diagrama de flujo detección inicio y fin 56
Imagen 33Gráfico de los resultados 57

Índice de Tablas

Tabla 1Comparación de las ventanas comúnmente usadas. 24
Tabla 2Riesgos en el desarrollo del sistema 34
Tabla 3Tabla de Recursos de Hardware 36
Tabla 4Recursos de Software 36
Tabla 5Tabla de Depreciación del Hardware 36
Tabla 6Tabla de Depreciación de Software 37
Tabla 7Tabla de Sueldos 37
Tabla 8Tabla de Gastos de Servicios 37
Tabla 9Tabla de Costo Estimado de Desarrollo 37
Tabla 10Experiencia en los Lenguajes de Programación 37
Tabla 11Requerimientos Funcionales 40
Tabla 12Requerimientos no funcionales. 40
Tabla 13Resultados del prototipo 57
Tabla 14 Pruebas con nuevos resultados 58
Tabla 15Gráfico con muestras nuevas 58

1. Introducción

1.1 Panorama general

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

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

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

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

Los reconocedores estaban capacitados para aceptar un vocabulario más extenso. Durante esta época se logró una mejora con respecto al reconocimiento para palabras aisladas y continuas. Se desarrollaron técnicas como Dynamic Time Warping, modelado probabilístico y el algoritmo de retro propagación.

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

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

En la actualidad se continúa la mejora de las técnicas que comenzaron a desarrollarse hace algunos años. Sin embargo la mayoría de aplicaciones actuales tiene como objetivo principal desarrollar interfaces centradas en las necesidades del usuario, algunos ejemplos son los ya clásicos sistemas de comando por voz para dispositivos móviles, sistemas de comandos de voz para controlar robots y sistemas de autoidentificación por voz para restringir el acceso a recintos.

Nosotros pretendemos explotar otra aplicación, el poder recabar información de las conversaciones de los usuarios, con el fin de utilizar la información recabada para otros fines, como lo serian seguridad, marketing, experiencia del usuario, etc.

1.2 Sistemas similares que se han desarrollado

- Búsqueda de información usando entradas de voz [1].

- Robot Dirigido Por Voz [2].
- Verificación de Pronunciación Basada en Tecnología de Reconocimiento de Voz para un Ambiente de Aprendizaje [3].
- Diccionario español/inglés para el aprendizaje de vocabulario utilizando una interfaz de voz. [4]
- Reconocimiento de palabras clave en conversaciones espontáneas en castellano [5].
- Implementación de un reconocedor de voz gratuito a el sistema de ayuda a invidentes Dos-Vox en español [6].

1.3 Justificación

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

Hoy en día el éxito de ciertas aplicaciones en búsqueda de texto provoca interés en la búsqueda de otros medios. Entre estas, la búsqueda de habla es probablemente la más interesante, ya que la mayoría de la comunicación sigue ésta modalidad, y a que el habla debido a su naturaleza, es el proceso de comunicación más eficiente y económica de la sociedad.

Es por esto que el prototipo para reconocer palabras claves en conversaciones de voz que se pretende realizar, implica una importante innovación tecnológica, además de permitir aplicar un sin fin de conocimiento que se han adquirido atreves de la carrera; procesamiento de señales, procesamiento digital de voz, reconocimiento de patrones, redes neuronales., ingeniería de software, bases de datos, etc. Con el fin de generar un prototipo que pueda ser implementarlo en un futuro cercano en un proyecto de mayor envergadura que involucre otras áreas, poniendo especial énfasis en su potencial aplicación en el área de la seguridad (permitiendo obtener información de las conversaciones de presuntos criminales) y marketing (conociendo los gustos de las personas y los principales temas en sus conversaciones).

1.4 Objetivo

Desarrollar un prototipo que sea capaz de reconocer palabras clave dentro de conversaciones de voz, con el fin de recuperar información específica sin la necesidad de escuchar grabaciones completas de las conversaciones. Utilizando técnicas de procesamiento digital de señales para el preprocesamiento de la señal digital de voz, redes neuronales como técnica de clasificación de los vectores de características de la señal de voz y reconocimiento de patrones para la lograr la correcta identificación de las palabras emitidas dentro de la conversación de voz.

1.5 Productos o resultados esperados

- Prototipo capaz de reconocer palabras clave en conversaciones de voz.
- Documentación del prototipo.
- Un artículo de divulgación científica sobre los resultados obtenidos.

1.6 Metodología

Por la naturaleza del proyecto se ha seleccionado el modelo de desarrollo iterativo e incremental, esta metodología nos permite establecer bloques temporales de tiempo llamados iteraciones. Cada iteración se puede entender como un mini proyecto, en cada iteración se repite un proceso de trabajo similar con el fin de producir un resultado completo sobre un producto final. Debido a esta forma de trabajo será posible ir entregando constantemente los avances del prototipo a nuestros directores de trabajo terminal, sinodales y profesores de las materias trabajo terminar I y trabajo terminal II. Para ello cada objetivo o requisito que visualicemos se deberá completar en una sola iteración, deberemos realizar todas las tareas necesarias para completarlo de manera definitiva como lo son documentación y pruebas y así poder seguir avanzando en el desarrollo del prototipo. De esta manera se podrá visualizar un avance progresivo del prototipo sin el riesgo de pasar por desapercibido algún objetivo o requerimiento.

Conforme las iteraciones vayan transcurriendo nuestro producto debe evolucionar de manera significativa, al finalizar cada iteración se añadirán nuevos objetivos y requisitos, con los cuales se trabajaran. Un aspecto fundamental será hacer un análisis previo para poder priorizar los requisitos, de manera que conforme se avance en el desarrollo del prototipo este adquiera más valor significativo con cada iteración.

1.7 Organización del documento

La organización del presente trabajo se divide en 6 capítulos, a continuación se comenta el contenido de cada uno:

Capítulo 1. Introducción. En este capítulo se da una introducción al tema de investigación, se nombran sistemas similares que se han desarrollado, la justificación, el objetivo, los productos o resultados esperados, la metodología y la organización del trabajo para lograr el desarrollo del trabajo terminal.

Capítulo 2. Estado del arte. En este capítulo se explican algunas de las investigaciones que ya han sido realizadas para el reconocimiento de voz y que fueron aplicadas en diferentes lenguas.

Capítulo 3. Marco Teórico. En este capítulo se explica lo que es el procesamiento digital de voz, así como las diferentes de técnicas utilizadas en el reconocimiento voz para la construcción de una metodología adecuada que pueda ser aplicada.

Capítulo 4. Análisis. En este capítulo se explica las herramientas necesarias para cada una de las partes que serán necesarias para desarrollar el trabajo, los algoritmos, técnicas y métodos a utilizar.

Capítulo 5. Diseño. En este capítulo se explica el diseño del sistema, se muestran los diagramas UML que se consideraron pertinentes en base a la metodología de desarrollo utilizada, permitiendo entender de manera rápida el funcionamiento del sistema.

Capítulo 6. Implementación y Resultados. En este capítulo se muestra el prototipo final, aquí se puede observar cómo está construido el prototipo de acuerdo a un diagrama de bloques y como trabaja e interactúa con el usuario final, además se muestran graficas que permiten analizar de manera visual el desempeño del prototipo puesto a prueba en un ambiente controlado.

Capítulo 7. Conclusiones. En este capítulo se pueden ver nuestras observaciones sobre el trabajo realizado de manera objetiva en base al análisis de los resultados obtenidos en las pruebas.

Capítulo 8. Trabajo a Futuro. En este capítulo se habla de los trabajos que se ambicionan desarrollar y que son factibles de acuerdo a los resultados obtenidos en este trabajo terminal.

Capítulo 9. Referencias. En este capítulo están plasmadas todas las referencias bibliográficas consultadas para el desarrollo del trabajo terminal.

Capítulo 10. Glosario. En este capítulo se aborda un pequeño glosario de palabras muy utilizadas en este documento y que no son conocidas por personas que trabajan directamente con el procesamiento de señales de voz.

2. Estado del arte

2.1 Introducción

En los últimos años se ha avanzado en gran medida en el desarrollo de sistemas que simplifiquen la interacción entre el hombre y la máquina. Uno de estos desarrollos es por medio de la voz. La utilización de la voz, y en el caso que nos ocupa, el Reconocimiento de Habla, como medio para dar órdenes a sistemas controlados por voz ofrece varias ventajas respecto al método tradicional de comunicación entre el usuario y la máquina [7]:

Hace está comunicación más rápida, y más agradable para los nuevos usuarios, ya que al ser la forma natural de comunicarse no se necesita ninguna habilidad especial.

- Permite el tener las manos libres para utilizarlas en alguna otra actividad, a la vez que se van dando órdenes por medio de la voz.

- Permite movilidad, ya que la voz se puede enviar a distancia y ser recogida por un micrófono, a diferencia de un teclado que no se puede mover de la mesa de trabajo.

- Permite acceso remoto, al usar redes de comunicación ya establecidas como la red telefónica o el Internet, que son las redes de comunicaciones más extendidas.

La posibilidad de hacer que, un sistema de cómputo interprete lo que se dice, ha sido una expectativa muy fuerte en el ámbito computacional. Un SARH (Sistemas Automáticos de Reconocimiento del Habla), 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, luego son digitalizadas e interpretadas por el sistema, el cual extrae patrones o parámetros característicos de esta señal, para finalmente realizar una clasificación y el reconocimiento

2.2 Breve historia del reconocimiento de voz por computadora

Los avances conseguidos en el ámbito de las tecnologías del habla son cada vez más significativos. En el campo del reconocimiento automático de voz, los sistemas actuales manejan cada vez vocabularios más grandes y logran menores tasas de error gracias al uso de algoritmos más eficientes, a la aparición de equipos más potentes y baratos.

No obstante, a pesar de los grandes avances realizados, se está todavía muy lejos de un sistema de reconocimiento automático de voz universal que funcione bien en cualquier aplicación a la que sea destinado. En general, el diseño y las características de los actuales sistemas de reconocimiento automático de voz dependen fuertemente de la aplicación a la que van a ser destinados y a las condiciones de funcionamiento.

A continuación se presenta una línea de tiempo referente a la historia del reconocimiento de voz [8].

2.2.1 Antecedentes

- 1870 Alexander Graham Bell Quería construir un dispositivo que hiciera el habla visible a las personas con problemas auditivos. Resultado: el teléfono.

- 1880 Tihamir Nemes Intenta desarrollar un sistema de transcripción automática que identifique secuencias de sonidos y los imprima (texto). El proceso es rechazado por no ser realista.

- 1910 AT&T Bell Laboratories Construye la primera máquina, basada en plantillas, capaz de reconocer voz de los 10 dígitos del Inglés. Requiere un extenso entrenamiento a la voz de una persona, pero una vez logrado tiene un 99%.

2.2.2 Los inicios: años 50

- RCA Labs. Reconocimiento de 10 sílabas mono-locutor.

- University College in England. Reconocedor fonético.

- MIT Lincoln Lab. Reconocedor de vocales independiente del hablante.

2.2.3 Los fundamentos: años 60 – Comienzo en Japón (NEC labs)

- Dynamic Time Warping (DTW – Alineación Dinámica en Tiempo -). Vintsyuk (Soviet Union).

- El proceso es muy lento

- Empiezan a reducir los alcances y se centran en sistemas más específicos:
  - Dependientes del locutor
  - Flujo discreto de habla (con espacios / pausas entre palabras)
  - Vocabulario pequeño (menor o igual a 50 palabras)

- Estos sistemas empiezan a incorporar técnicas de normalización del tiempo.

- Se minimiza la diferencia en la velocidad del habla.

- IBM y CMV (Carnegie Mellon University). Trabajan en Reconocimiento del Habla Continua. HAL 9000. Los resultados no llegan hasta 1970.

2.2.4 Las primeras soluciones: años 70 - El mundo probabilístico.

- Se produce el primer producto de reconocimiento de voz, el VIP100 de Threshold Technology Inc.

- Reconocimiento de palabras aisladas.

- IBM: desarrollo de proyectos de reconocimiento de grandes vocabularios.

- Gran inversión en los EE. UU.: proyectos DARPA.

- Gracias al lanzamiento de grandes proyectos de investigación y financiamiento por parte del gobierno norteamericano, se precipita la época de la inteligencia artificial.

- Los sistemas empiezan a incorporar módulos de: análisis léxico, análisis sintáctico, análisis semántico y análisis pragmático.

- Sistema HARPY (CMU), primer sistema con éxito.

2.2.5 Reconocimiento del Habla Continua: años 80

- Surgen los sistemas con algoritmos para el habla continua de vocabulario amplio más de 1000 palabras.
- Explosión de los métodos estadísticos: Modelos Ocultos de Markov.
- Introducción de las redes neuronales en el reconocimiento de voz.
- Sistema SPHINX.

2.2.6 Empieza el negocio: años 90 - Primeras aplicaciones: ordenadores y procesadores baratos y rápidos

- Sistemas de dictado.
- Integración entre reconocimiento de voz y procesamiento del lenguaje natural.

2.2.7 Una realidad: años 2000 - Integración en el Sistema Operativo

- Integración de aplicaciones por teléfono y sitios de Internet dedicados a la gestión de reconocimiento de voz (Voice Web Browsers).
- Aparece el estándar VoiceXML.
- Empresas importantes actualmente:
  - Philips
  - Lernout & Hauspie
  - Sensory Circuits
  - Dragon Systems
  - Speechworks
  - Vocalis
  - Dialogic
  - Novell
  - Microsoft
  - NEC
  - Siemens
  - Intel

2.3 Algunas publicaciones relevantes en reconocimiento de voz

Las primeras publicaciones documentadas referentes al análisis de voz datan de 1952, con el clásico trabajo de Peterson & Barney [8], realizados en los laboratorios de Bell Telephone, que ha influido fuertemente en las investigaciones posteriores. En este trabajo utilizan grabaciones de 10 vocales pronunciadas por 33 hombres, 28 mujeres y 15 niños, se obtuvieron medidas de los formantes F1, F2 y F3 y del tono fundamental (F0). Se descubrió que existía una considerable variabilidad en las frecuencias de los formantes de los distintos hablantes, a pesar de ello las vocales presentaron un alto grado de acierto en su identificación.

Para 1967 con el trabajo de Lindblom [9] se empieza el análisis para el reconocimiento del habla continua donde se habla sobre de la necesidad de recurrir a un conjunto de información adicional como la evolución de los formantes a lo largo del tiempo.

Los ingenieros de la Bell, Bishnu Atal (1972, 1974) y Aaron Rosenberg y Sanbur (1975) publican sus primeros estudios tomando como base de extracción de datos, coeficientes cepstrum

y coeficientes de predicción lineal o LPC. En esta misma época, son probados estos parámetros acústicos del habla en el diseño de sistemas de reconocimiento automático. En 1972 Wolf analiza combinaciones de hasta veintisiete referencias extraídas de consonantes nasales, espectros de vocales, frecuencia fundamental. Su y Fu en 1973 utilizan como informaciones eficientes los espectros de consonantes nasales. Li y Hughes en 1974 toman como referencia matrices de correlación referidas a fragmentos de habla continua [10].

En 1978 con el trabajo de Rabiner se establece el análisis del habla por porciones, por la naturaleza cambiante de la voz, y a partir de ello se procesan porciones o ventanas de la señal de voz [11].

2.4 Parámetros característicos de una señal de voz

Los parámetros característicos de voz que han dominado el área del reconocimiento del habla han sido:

- Coeficientes Cepstrales Reales (RCC), Oppenheim (1969)
- Coeficientes de Predicción Lineal (LPC), Atal y Hanauer (1971)
- Coeficientes Cepstrales de Predicción Lineal (LPCC), Atal (1974)
- Coeficientes Cepstrales en Frecuencia en escala de Mel (MFCC), Davis y Mermelstein (1980)
- Perceptual Linear Prediction Coefficients (PLP), Hermansky (1990)

2.5 Modelos para el reconocimiento de voz

- Cuantificación Vectorial, Robert M. Gray (1984) y Linde, Buzo, Gray (.1980).
- NNA (Redes Neuronales Artificiales), Lippman (1988).
- Mixturas Gaussianas, Rabiner (1989).
- HMM (Cadenas Ocultas de Markov), Rabiner (1989).
- Mixturas Gaussianas. Asociadas con HMM, Rabiner (1989).

2.6 Aspectos a destacar en un sistema de reconocimiento de voz

Al ser la señal de voz variante en el tiempo, se hace indispensable su digitalización para ser tratada por los recursos computacionales con los que se puede hacer posible su reconocimiento.

Es indudable que dentro de los sistemas de reconocimiento de voz se requieran parámetros específicos que permitan realizar una obtención de información apropiada.

Existen varios factores a analizar en dichos sistemas, dentro de los cuales se encuentran los siguientes:

2.6.1 Tamaño del vocabulario y confusión

Los sistemas conforme más palabras se deseen que reconozcan tienden a incrementar los índices de error. Se tienen reportes de un aceptable porcentaje de reconocimiento cuando se trabaja con números de palabras menores a 1000, pero el problema se agrava cuando este número se incrementa. Originando que el porcentaje de reconocimiento se vea gravemente afectado, pues con frecuencia el sistema tiende a caer en inestabilidades y por ende a perder características de confiabilidad.

2.6.2 Sistemas dependientes e independientes del locutor

La gran controversia dentro de los sistemas de reconocimiento de voz se ve plasmada en los sistemas de reconocimiento dependientes y no dependientes del locutor. A lo largo de la historia que tiene el advenimiento de los SARH, los sistemas dependientes del locutor se han gestado y son los que han hecho una realidad el alcanzar un alto índice de reconocimiento.

Sin embargo y en contraparte, se encuentran los sistemas independientes del locutor donde se hace evidente la necesidad de implementar mecanismos cada vez más sofisticados; representando un problema aún en nuestros días. A pesar de esto, los avances no se han hecho esperar y aunque actualmente podemos hablar de reconocimiento de voz para un grupo determinado de personas, es una realidad que la extensión de estos esquemas para que cubran a toda una población resulta difícil.

2.6.3 Voz aislada, discontinua y continua

Gran parte del desarrollo de este trabajo se ve enfocado en estos elementos. Se entiende por voz aislada aquella que podemos percibir como unidades del habla que forman un núcleo elemental de entendimiento dentro de la estructura lingüística en donde se gesten.

Cabe hacer la mención de que este hecho es importante porque la sílaba y el fonema pueden entrar dentro de esta clasificación. Sin embargo, existe un problema con el fonema; como tal y pieza independiente carece de sentido, siendo totalmente abstracto y sin relevancia cuando se manifiesta de manera independiente. En contraparte, la sílaba es totalmente autónoma sin necesidad de compartir espacios temporales con algún otro medio lingüístico; por tal motivo, podemos decir que su contenido de información es vasto y enorme.

La voz discontinua es una manifestación en donde las palabras o secuencias sonoras se encuentran entrelazadas por una pauta que no permite que haya continuidad entre una estructura anterior y la consecuente, más bien es simplemente el intermedio entre lo continuo y lo individual.

Finalmente el habla continua es por naturaleza la forma que los seres humanos tienen para comunicarse con los demás, es importante observar la forma en la que los elementos anteriores se presentan en la vida del ser humano. El ser humano en su proceso de adaptación permite que los sonidos ayuden al equilibrio óptimo de las funciones básicas del cerebro que lo acompañaran durante toda la vida.

Al pasar el tiempo, el individuo comienza a coordinar sus estructuras sonoras de tal forma que las acopla al medio que le rodea; esto es, no importa si haya nacido en México por ejemplo o si es trasladado a otra región del mundo. Este tiende a aprender y a acoplarse al medio donde se encuentre pues en esos momentos no importan mucho las nacionalidades.

2.6.4 Voz aplicada a tareas o en general

Los sistemas de reconocimiento de voz se encuentran altamente ligados al tipo de aplicación que se esté llevando a cabo en su implementación, es decir, muchos sistemas se encuentran limitados en contexto por la tarea que tienen que realizar mientras que otros quedan completamente abiertos. Piénsese en un SARH destinado a gestionar conversaciones telefónicas, reservaciones de vuelos aéreos, etc., como es de suponerse, la cantidad de elementos que tiene este vocabulario se encuentra limitada a unas pocas palabras.

2.6.5 Voz leída o espontánea

Los SARH existentes hasta estos momentos se manifiestan en dos grandes vertientes y sobre todo cuando se habla de bases de datos destinadas a tal fin. Los corpus de voces almacenados para estudio se diferencian en el hecho de que sus grabaciones se encuentran hechas por personas que pronuncian las frases cuando las leen o cuando se encuentran en una charla normal.

TIMIT es una base de datos que demuestra este hecho, gran parte de las muestras de voz que en ella se encuentran almacenadas son realizadas por personas que se encontraban en charlas de oficina o de lugares concurridos; en donde la voz que se percibe es totalmente continua y espontánea. Esto es, que no existió un esquema de conversación preestablecido.

Caso contrario sucede con las muestras de voces leídas, en donde la muestra de voz es obtenida de una secuencia de frases preestablecidas (lecturas, formatos, etc.) y por ende el hablante pone más cuidado en lo que está diciendo y la claridad se nota en gran parte del texto.

2.6.6 Condiciones adversas

Este tema se refiere específicamente a las perturbaciones que una señal de voz puede recibir por causas del medio ambiente.

Cabe recalcar que es importante tomar este hecho pues si se desea tener un sistema de reconocimiento de voz que opere bajo ciertas características (lugar de trabajo, condiciones atmosféricas, etc.) se deberán de tener las muestras de voz extraídas bajo las mismas condiciones en que funcionará el sistema.

3. Marco teórico

3.1 Esquema a bloques de un sistema de reconocimiento de voz por computadora

Finalmente, los sistemas de reconocimiento basados en cualquier unidad lingüística, se ilustra en el diagrama de bloques.

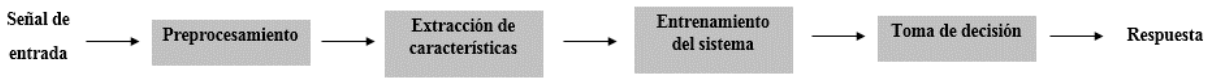

Imagen 1Esquema de un diagrama de bloques de un Sistema de Reconocimiento del

3.2 Metodología genérica empleada en un reconocimiento de voz

Todo trabajo de reconocimiento de voz requiere de un conjunto de tareas, las cuales se ejecutarán de preferencia en secuencia e independientemente del objetivo a alcanzar.

Estas tareas se enlistan a continuación:

- Preparar y crear un Corpus de Voces a utilizar en el trabajo.
- Definir, preparar y realizar el pre-procesamiento que será aplicado a los archivos del

Corpus.

- Seleccionar, diseñar y programar los algoritmos y métodos que se utilizarán para la extracción de parámetros a utilizar en el reconocimiento de voz.
- Entrenar el sistema.
- Verificar el grado (%) de aceptación del sistema (reconocimiento).

3.3 El Corpus de Voces

El Corpus de Voces es el conjunto de archivos con información de voz digitalizada, del cual se extraen el conjunto de parámetros con los cuales se entrena y comprueba el sistema desarrollado.

3.3.1 Condiciones para obtener un corpus de calidad

Es muy importante, durante la etapa de grabación de sonidos, tener en cuenta las condiciones del lugar; distancia al micrófono; ajuste de la ganancia y sensibilidad de dicho micrófono.

En la figura se tienen tres ejemplos de grabaciones de números del cero al nueve, la primera está saturada, con mucho ruido y poca separación entre palabras, la segunda es adecuada y la tercera grabación tiene un volumen muy bajo.

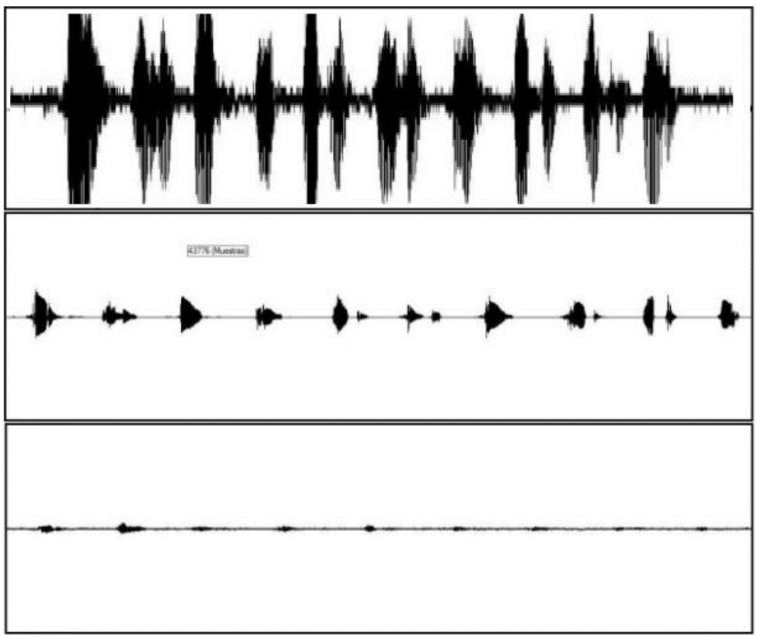

Imagen 2Ejemplo de grabaciones con amplitud: saturada, media y baja.

3.4 Preprocesamiento de la señal de voz

3.4.1 Preénfasis

La descripción del modelo fuente-filtro para la producción de voz mostrado, indica que el espectro de sonidos sonoros posee una pendiente de -6 dB/octava, mientras se incrementa la frecuencia. Esta es una combinación de una pendiente de -12 dB/octava debido a la fuente de excitación de la voz y una pendiente +6 dB/octava debida a la radiación producida por la boca. Esto significa que por cada vez que se duplica la frecuencia, la señal en amplitud, y por tanto la respuesta del tracto vocal medida, son reducidas por un factor de 16. Es por tanto deseable compensar la caída de -6 dB/octava pre-procesando la señal de voz para darle un levantamiento de +6 dB/octava en el rango apropiado de tal forma que el espectro medido tenga un rango dinámico similar a través de toda la banda de frecuencias. Esto se llama preénfasis. En un sistema digital de procesamiento de señales, el preénfasis puede ser implementado ya sea de primer orden, usando un filtro analógico paso-altas con una frecuencia de corte a 3 dB en algún lugar entre 100 Hz y 1 kHz (la posición exacta no es crítica) que precede al filtro de anti-aliasing y el convertidor A/D, o bien usando un filtro digital paso-altas que procesa la señal de voz digitalizada. El filtro paso-altas puede ser logrado digitalmente usando la ecuación en diferencias:

$$y[n] = x[n] - ax[n-1]$$

Ecuación 1Ecuación en diferencias de filtro paso-altas.

Donde y[n] denota la muestra actual de salida del filtro de preénfasis, x[n] es la muestra actual de entrada, x [n-1] es la muestra anterior de entrada y a es una constante usualmente elegida entre 0.9 y 1. De nuevo, el valor elegido no es crítico. Tomando la transformada z de la ecuación da:

$$Y(z) = X(z) - az^{-1}X(z) = (1 - az^{-1})X(z)$$

Ecuación 2Transformada Z de la Ecuación 1

Donde z^{-1} denota el operador de retardo de muestra unitario. La función de transferencia H (z) del filtro es por tanto:

$$H(z) = Y(z) / X(z) = 1 - az^{-1}$$

Ecuación 3Función de transferencia.

3.4.2 Ventaneo

El mecanismo de ventaneo muestra como a cada porción de la señal de voz (de un tamaño predefinido por el usuario), se le asigna una ventana, de tal forma que las muestras queden ponderadas con los valores de la función escogida. En este caso, las muestras que se encuentran en los extremos de la ventana tienen un peso mucho menor que las que se hallan en el medio, lo cual es muy adecuado para evitar que características de los extremos del bloque varíen la interpolación de lo que ocurre en la parte central, la cual es la más significativa, de las muestras del segmento seleccionado.

La colocación de las ventanas puede realizarse de tal forma que existan solapamientos y, aunque ello repercutirá en los tiempos de respuesta del sistema reconocedor, proporcionará una mejor calidad en los resultados obtenidos.

Las funciones de ventaneo más comunes se muestran a continuación:

Tipo de Ventana	Amplitud pico del lóbulo lateral (relativa)	Ancho aproximado del lóbulo principal	Error pico aproximado $20\log_{10}\delta$ (dB)	Ventana de Kaiser equivalente β	Ancho de la transición de la ventana de Kaiser equivalente
Rectangular	**-13**	**4π/(M+1)**	**-21**	**0**	**1.81π/M**
Bartlett	**-25**	**8π/M**	**-25**	**1.33**	**2.37π/M**
Hanning	**-31**	**8π/M**	**-44**	**3.86**	**5.01π/M**
Hamming	**-41**	**8π/M**	**-53**	**4.86**	**6.27π/M**
Blackman	**-57**	**12π/M**	**-74**	**7.04**	**9.19π/M**

Tabla 1Comparación de las ventanas comúnmente usadas.

3.5 Detectar inicio y fin de palabra

La detección de inicio y fin de palabras, se realiza tomando en cuenta la actividad de energía y cruce de ceros de la señal y(n), con respecto a los valores que se tienen en condiciones de silencio, ruido ambiente. La variación del cruce de ceros es principalmente motivo de la emisión de una señal explosiva (“p‟, “t‟, “k‟, “b‟) o ruido aleatorio (“s‟, “f”, “ch‟, “j‟, “z‟). La variación de

energía ocurre en presencia de vocales (“a‟, “e‟, “i‟, “o‟, “u‟), semivocales (“m‟, “n‟, “ñ‟, “w‟, “d‟, “l‟, “g‟, “y‟).

La señal que se muestra en la figura 5, es una muestra de una detección de inicio y fin, corresponde a la palabra “seis‟, donde se aprecia el inicio del sonido de la silbante “s‟, detectado mediante el cruce de ceros. Las vocales “e‟ e “i‟, se caracterizan por el nivel de su energía y nuevamente la “s‟ al final. Al inicio y fin de la grabación se aprecia el efecto del ruido ambiente que corresponde al silencio.

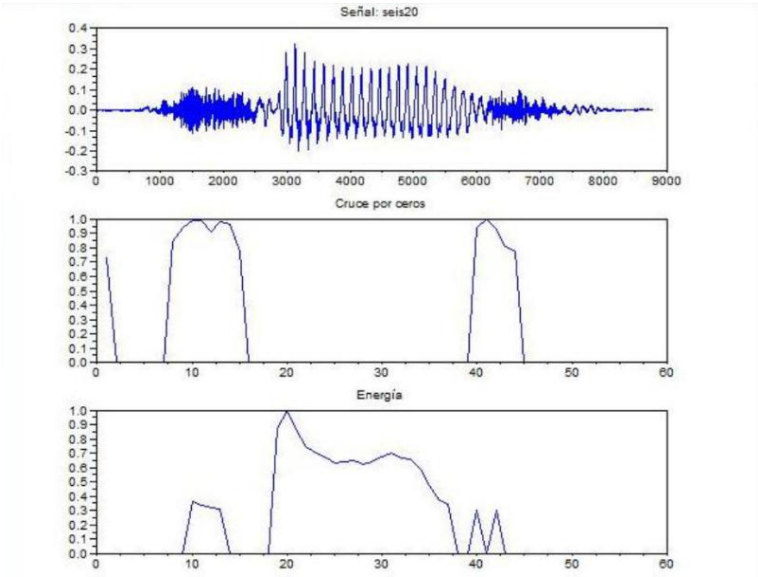

Imagen 3Ejemplo de detección de inicio y fin de palabra con cálculo de cruce de ceros y energía.

3.6 Parámetros característicos de una señal de voz

A través de la historia las diferentes investigaciones referidas al reconocimiento de la voz nos han dejado que los parámetros que mejores resultados han dado son: el LPC, CEPSTRUM, MFCC, entre otros.

3.6.1 Cepstrum

El Cepstrum es llamado por algunos autores, la transformada de la transformada, Lo cual es casi cierto. A continuación se presenta el flujo de procesamiento para obtener los coeficientes Cepstrales de una señal.

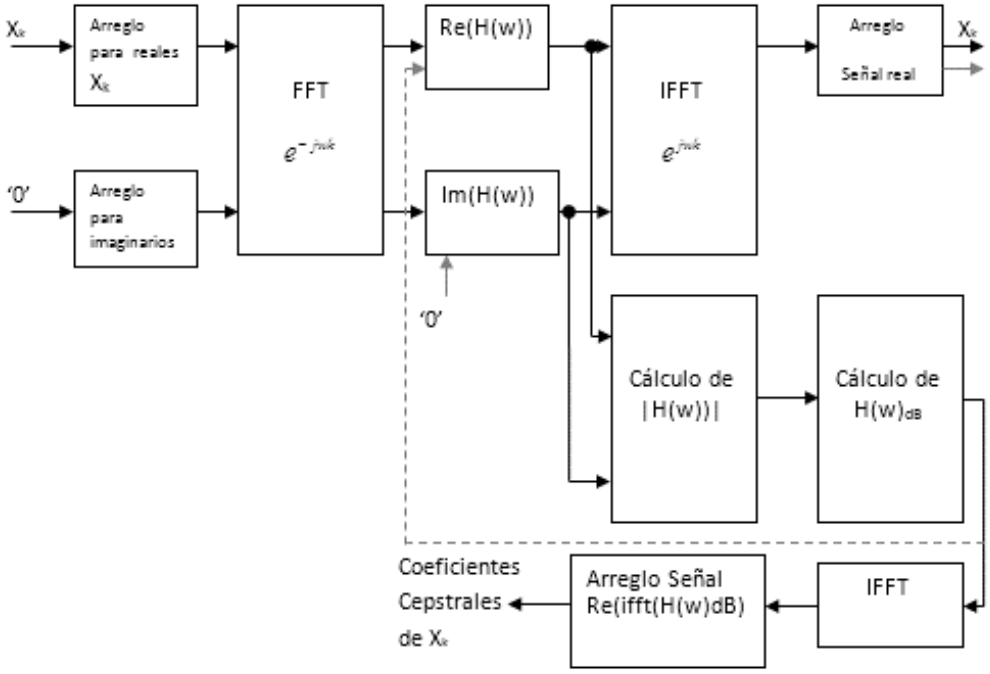

Imagen 4Cálculo de los coeficientes cepstrales (Cepstrum) de un segmento de señal.

Podemos observar que la operación es idéntica a llevar el resultado de $H(w)_{dB}$ a la entrada del módulo que realiza la IFFT y tomar los valores reales de esta operación, siendo el resultado $H(w)_{dB}$ el conjunto de los números reales a tener en cuenta y haciendo el conjunto de los números imaginarios igual a CERO (0). En ese caso en lugar de obtener los valores de X_k, estamos obteniendo los valores de los coeficientes cepstrales de la señal.

El cálculo del real Cepstrum se efectúa de la siguiente manera:

$$rceps=real(ifft(log(abs(fft(x)))))$$

Ecuación 4Calculo de real cepstrum.

El Cepstrum de una señal es llamado la cuasifrecuencia de la señal, lo que es equivalente a decir que es parecido al periodograma de la señal, ya que estamos realizando la transformada inversa de Fourier al espectro de Potencia de la Frecuencia de la señal y la respuesta corresponde al dominio

del tiempo. El componente sobre el valor de la ordenada 1 equivale a la frecuencia de muestreo dividida por 2 (frecuencia de análisis de la señal más alta, según la teoría del muestreo) o lo que es lo mismo, el período que se puede analizar de la señal más pequeño.

Para señales que tienen definida su composición espectral o rango de respuesta de frecuencia, se puede seleccionar el rango de coeficientes Cepstrum que le corresponden, eliminando el resto por considerarlos ruidos o no pertenecientes a la señal bajo análisis.

La cantidad de coeficientes Cepstrum dependen de la frecuencia de muestreo de la señal, ya que el Δf de la $H(w)_{dB}$, está en función de la misma como se explicó anteriormente. A mayor frecuencia de muestreo, mayor cantidad de coeficientes Cepstrales. Para algunas aplicaciones de reconocimiento de patrones, como es el caso de reconocimiento de voz, el número de coeficientes Cepstrales para los patrones de la señal voz se toman bajo este criterio.

Respuesta interesante en el Cepstrum, es la aparición de un valor alto, pico, para el análisis de segmentos de datos donde existe una frecuencia fundamental F0. En la figura 16 se presenta un pico cerca del valor 100 en el eje de las abscisas. La distancia al origen determina el valor de T0 en función de la frecuencia de muestreo f_s, luego se puede conocer F0.

Otra respuesta interesante para el caso de señales de voz, es que tomando los valores Cepstrales que se encuentran por debajo del pico a partir del origen y en una cantidad no muy alta, máximo 40 valores, al aplicarles a los coeficientes CPP el cálculo de $H(w)_{dB}$, se observa que la respuesta se corresponde con la envolvente de la transformada $H(w)_{dB}$ del segmento de señal bajo análisis, que es la respuesta a la información inteligente que la señal de voz porta, producida por la modulación en el tracto vocal; y si se toman los componentes después del pico, obtenemos como respuesta $H(w)_{dB}$ la parte de la señal de voz que se corresponde con la excitación del tracto vocal. Ambas respuestas están juntas en la señal de voz, pero la más importante es la de la envolvente que es donde está la información consciente del habla. De ahí que los coeficientes Cepstrales iniciales forman parte del repertorio de parámetros que se utilizan para el reconocimiento de voz.

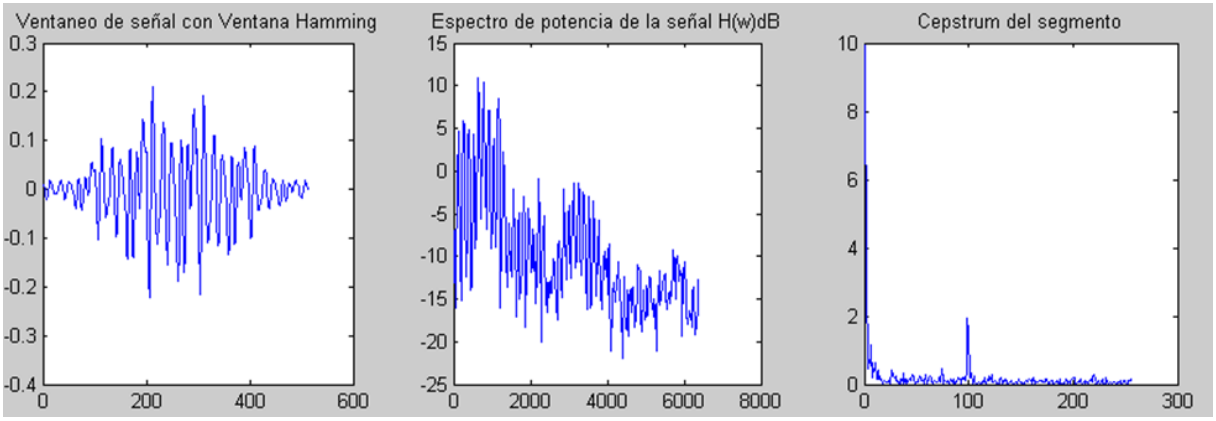

Imagen 5Gráficas resultantes en el flujo de cálculo del Cepstrum.

3.6.2 LPC

El modelo de filtro "todo polos" es el más utilizado y el análisis de auto recursividad de la señal es la herramienta para hallar los coeficientes del filtro en cuestión. La funcionalidad matemática que da como resultado los coeficientes del filtro, es una relación matricial que se obtiene al realizar la predicción lineal de las muestras de las señales, con lo cual se obtienen los coeficientes de predicción lineal (LPC), también llamados parámetros LPC, que no son otra cosa que los coeficientes del filtro "todo polos" buscado.

El cálculo de parámetros LPC es un procedimiento que se realiza en el dominio del tiempo, y está basado en la consideración de que el término enésimo s(n) de una secuencia, puede ser estimado a partir de los términos anteriores, considerando la sumatoria de los mismos con un peso asociado a cada uno de ellos. ŝ(n) es el término estimado de la secuencia. Este modelo es llamado de predicción lineal (LPC) o auto recursivo (AR).

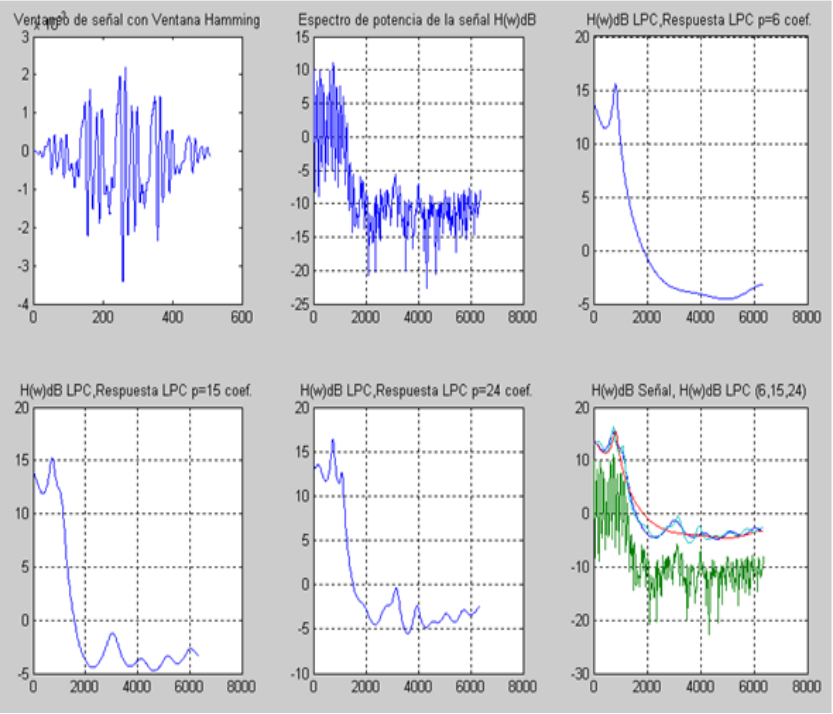

Imagen 6Respuestas de frecuencias $|H(w)|\,dB$ ***versus*** $|H(w)|\,dB(p=6{,}15{,}24)$.

3.6.3 MFCC

Un método más eficiente para extraer características y que es el más utilizado actualmente en reconocedores comerciales son los Coeficientes Cepstrales en Escala de Mel (MFCC), este método es robusto, además hace uso de la Transformada de Fourier para obtener las frecuencias de la señal. El objetivo es desarrollar un conjunto de características basadas en criterios perceptuales, diversos experimentos muestran que la percepción de los tonos en los humanos no está dada en una escala lineal, esto hace que se trate de aproximar el comportamiento del sistema auditivo.

Los coeficientes Cepstrales en Frecuencia en Escala de Mel (MFCC) son una representación definida como el Cepstrum de una señal ventaneada en el tiempo que ha sido derivada de la aplicación de una Transformada Rápida de Fourier, pero en una escala de frecuencias no lineal, las cuales se aproximan al comportamiento del sistema auditivo humano.

El cálculo de los coeficientes Mels utiliza dos de las herramientas más conocidas en el análisis de señales:

La transformada de Fourier para la representación del contenido espectral de una señal.

El diseño de un banco de filtros para permitir la selección de bandas de frecuencia de la señal bajo análisis.

Con la transformada de Fourier se conoce el contenido en frecuencia (espectro) de la señal y con los filtros diseñados (sintetizados), se logra obtener las componentes de frecuencia que a cada banda les aporta la señal analizada.

El principio de ponderar la energía que aporta a cada banda de frecuencias la señal bajo análisis y luego calcular en términos de un coeficiente para cada valor de energía en banda de frecuencia, es a lo que llamamos coeficientes Cepstral.

El algoritmo o método para el cálculo de estos coeficientes, usando las dos herramientas mencionadas es lo que se describe a continuación.

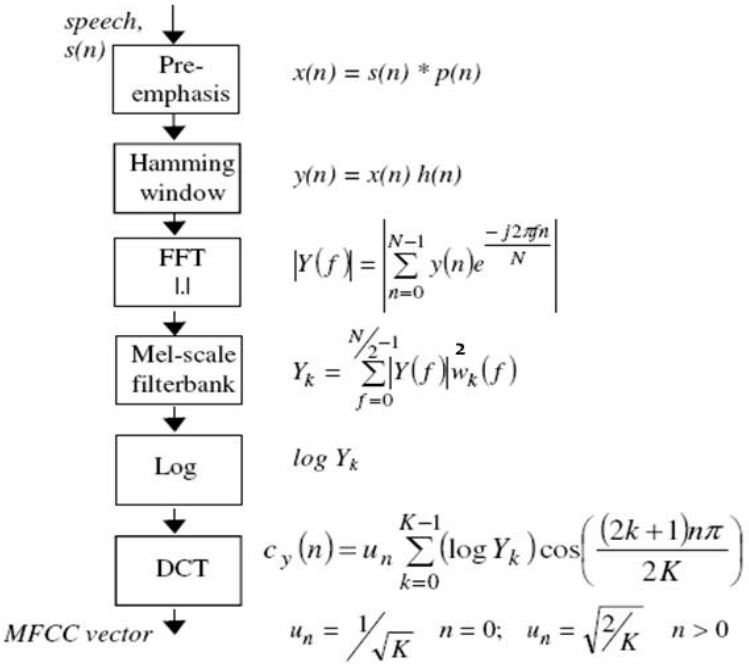

Imagen 7Algoritmo para el cálculo de los coeficientes.

3.7 Etapa de reconocimiento

3.7.1 Redes neuronales

Perceptrón

Los sistemas de reconocimiento 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. Esa forma de interconexión de las unidades de proceso es especialmente útil en aplicaciones que requieren una gran potencia de cálculo para evaluar varias hipótesis en paralelo, como sucede en los problemas de reconocimiento de voz.

Las unidades de proceso pueden ser de varios tipos; las más simples (y utilizadas) disponen de varias entradas, y la salida es el resultado de aplicar alguna transformación no lineal a la combinación lineal de todas las entradas. Otro tipo de neuronas un poco más elaborado se caracteriza por disponer de memoria; en ellas la salida en cada momento depende de entradas anteriores en el tiempo.

La forma en que las neuronas se conectan entre sí define la topología de la red, y se puede decir que el tipo de problemas que una red neuronal particular soluciona de forma eficiente, depende de la topología de la red, del tipo de neuronas que la forman, y la forma concreta en que se entrena la red.

La red neural que mejores resultados está dando hasta este momento en reconocimiento automático del habla es la denominada "perceptrón multicapa". La figura muestra su topología: las neuronas se disponen por "capas"; hay una capa de entrada, que opera directamente sobre los vectores de observación o puntos de las plantillas, una capa de salida que apunta la palabra reconocida, y una o más capas intermedias. Cada capa está compuesta por varias unidades de proceso, que se conectan con la siguiente capa por una serie de enlaces a los que se da un cierto peso específico wij.

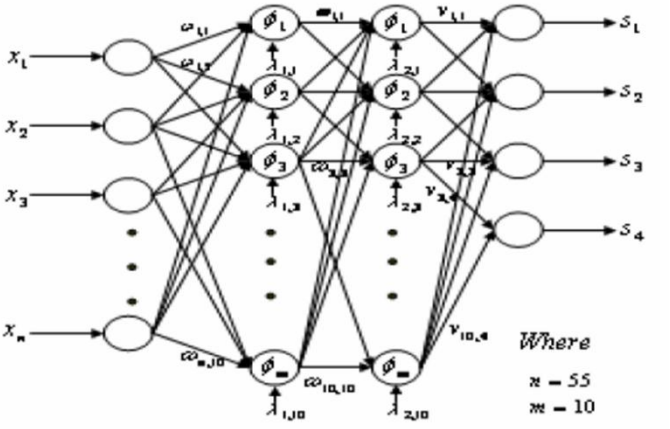

Imagen 8Arquitectura de una red neuronal de tipo perceptrón multicapas

El conjunto de vectores de características entra en la capa de neuronas de entrada, y posteriormente es propagado a las capas siguientes. En cada célula de proceso se calcula la suma ponderada (por los pesos wij) de las señales de entrada, y posteriormente se procesa en la neurona

con su sistema no lineal. Si el resultado de esta operación supera un cierto umbral, la neurona reacciona, transmitiendo señal a las neuronas siguientes de la capa superior.

En la fase de entrenamiento, dada una entrada conocida (p.ej. un conjunto de vectores que representa el dígito 1), la salida de la red es comparada con la salida esperada (y conocida de antemano), calculándose un error. Ese error se propaga hacia abajo, ajustándose de esta manera los pesos de las conexiones entre neuronas. Efectuándose este proceso varias veces se consigue que la red "aprenda" que respuesta debe dar para cada entrada en la fase de reconocimiento.

Mapas auto-organizados

Los mapas autoorganizados o SOM (Self-Organizing Maps), también llamados redes de Kohonen son un tipo de red neuronal no supervisada, competitiva, distribuida de forma regular en una rejilla de, normalmente, dos dimensiones.

Su finalidad es descubrir la estructura subyacente de los datos introducidos en ella. A lo largo del entrenamiento de la red, los vectores de datos son introducidos en cada neurona y se comparan con el vector de peso característico de cada neurona.

La neurona que presenta menor diferencia entre su vector de peso y el vector de datos es la neurona ganadora (o BMU) y ella, y sus vecinas verán modificados sus vector es de pesos.

Este tipo de mapas permiten reducir la dimension de los vectores de entrada para representarlos mediante una matriz de distancias unificada (U-matriz) generalmente consistente en una matriz 2D, apta para la visualización como una imagen plana.

Su estructura está dada por:

Matriz de neuronas: Las neuronas se distribuyen de forma regular en una rejilla de dos dimensiones, que pueden ser rectangulares o hexagonales, en las que cada neurona puede tener cuatro o seis vecinos respectivamente.

Espacio de entrada: Los datos de entrada corresponden a un vector de N componentes por cada atributo que queramos comprar, siendo esta dimensión la misma del vector de pesos sinápticos asociado a cada una de las neuronas de la rejilla.

Espacio de salida: Corresponde con la posición (2D) en el mapa de cada neurona.

Relación entre neuronas: Entre todas las neuronas hay una relación de vecindad que es la clave para conformar el mapa durante la etapa de entrenamiento. Esta relación viene dada por una función.

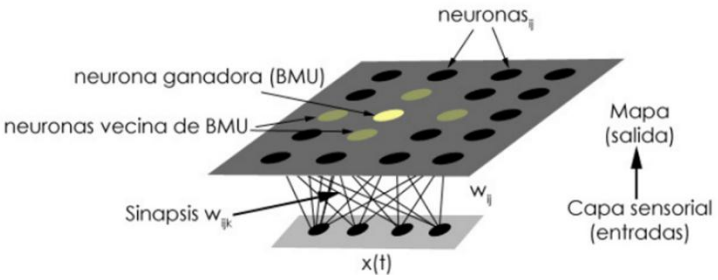

Imagen 9Estructura de un mapa auto-organizado.

Se entrena de la siguiente manera:

Para cada paso del entrenamiento (época) se introduce un vector de datos correspondiente a una entrada seleccionada aleatoriamente y se calcula la similitud entre este vector y el peso de cada neurona.

La neurona más parecida al vector de entrada será la neurona ganadora (Best-Matching Unit ó BMU). Generalmente se usa la distancia euclidiana para medir esta similaridad entre pesos sinápticos.

Tras esto, los vectores de pesos de la BMU y sus vecinos son actualizados de forma que se acercan al vector de entrada.

3.7.2 Modelos Ocultos de Markov Discretos

Otro enfoque alternativo al de medir distancias entre patrones (enfoque topográfico) es el de adoptar un modelo estadístico (paramétrico) para cada una de las palabras del vocabulario de reconocimiento, como son los modelos ocultos de Markov (HMM, del inglés "Hidden Markov Models").

En un modelo estadístico se puede considerar que la señal de voz se puede caracterizar de manera apropiada mediante un proceso aleatorio, y que los parámetros del proceso estocástico pueden ser determinados de manera precisa y bien definida.

Las Cadenas Ocultas de Markov (HMM) es un excelente modelo para este trabajo; la teoría básica de las HMM fue publicada en una serie de artículos clásicos por Baum y sus colegas a finales de los 60 e inicios de los 70 y se utilizó por primera ocasión para el reconocimiento de voz por Baker en CMU, y por Jelinek y colegas en IBM en los años 70.

Estos sistemas han sido posteriores en el tiempo, y hoy día la mayoría de los reconocedores en funcionamiento se basan en esta técnica estadística, ya que aunque sus prestaciones son similares a las de los sistemas basados en QV, requieren menos memoria física y ofrecen un mejor tiempo de respuesta. Tienen como contrapartida, una fase de entrenamiento mucho más lenta y costosa pero como esta tarea se realiza una única vez, y se lleva a cabo en los laboratorios. Es un precio que parece valer la pena pagar.

Un HMM se puede ver como una máquina de estados finitos en que el siguiente estado depende únicamente del estado actual, y asociado a cada transición entre estados se produce un vector de observaciones o parámetros. Se puede así decir que un modelo de Markov lleva asociados dos procesos: uno oculto (no observable directamente) correspondiente a las transiciones entre estados, y otro observable (y directamente relacionado con el primero), cuyas realizaciones son los vectores de parámetros que se producen desde cada estado y que forman la plantilla a reconocer.

Para aplicar la teoría de los HMM en reconocimiento de voz, se representa cada palabra del vocabulario del reconocedor con un modelo generativo (que se calculará en la fase de entrenamiento) y posteriormente, se calcula la probabilidad de que la palabra a reconocer haya sido producida por cada uno de los modelos de la base de datos del reconocedor. Para ello, se asume que durante la pronunciación de una palabra, el aparato fonador puede adoptar sólo un número(finito de configuraciones articulatorias (o estados), y que desde cada uno de esos estados se producen uno o. varios vectores de observación (puntos de la plantilla), cuyas características

espectrales dependerán (probabilísticamente) del estado en el que se hayan generado Así vista la generación de la palabra, las características espectrales de cada fragmento de señal dependen del estado activo en cada instante, y la evolución del espectro de la señal durante la pronunciación de una palabra depende de la ley de transición entre estados.

La representación más usual de un HMM es la utilizada para máquinas de estados finitos, es decir, conjuntos de nodos (que representar a los estados) y arcos (transiciones permitidas entre los estados). Un tipo de HMM especialmente apropiado para reconocimiento de voz son los modelos "de izquierda a derecha"; modelos en los que una vez que se ha abandonado un estado, ya no se puede volver a él. La figura representa un modelo con 'n' estados en el que desde cada estado sólo se permiten tres tipos de transición: al propio estado, al estado vecino y a dos estados más allá (este tipo de saltos que da recogido en una matriz de transiciones tridiagonal).

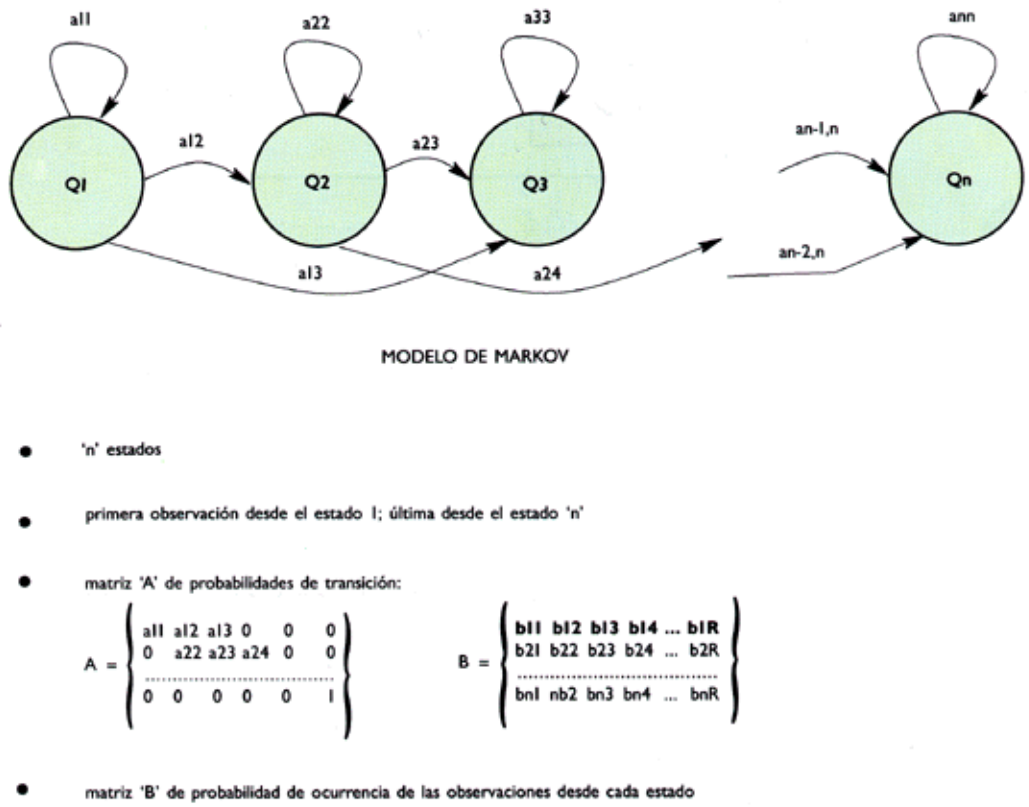

Imagen 10Modelo de Markov

En cuanto a la generación de puntos de la plantilla, en estos modelos se asume que el primer vector de observaciones se produce desde el primer estado, y el último se emite desde el último estado. Recuérdese que la secuencia de estados es la parte oculta del modelo: se conocen los vectores de parámetros, pero no desde qué estado se han producido.

4. Análisis

4.1 Análisis de Riesgos

En la tabla que se muestra a continuación están descritos los riesgos que son susceptibles a presentarse en el desarrollo de éste trabajo, describiendo su posible comportamiento así como una evaluación del resultado que daría al presentarse.

Id	Riesgo	Descripción	Efecto
R1	Control del Ambiente	No es posible controlar el ambiente en el que se obtienen la grabación de voz de un individuo, por lo tanto las grabaciones presentan variaciones.	Serio
R2	Definición de características	Las características que se determinan para diferenciar un fonema o palabra de otros no son lo suficientemente excluyentes.	Serio
R3	Extracción de características	Debido a las variaciones que puedan tener las grabaciones de voz, se dificulta la extracción de las características requeridas.	Serio
R4	Cambios en los requerimientos	Variaciones en los requerimientos iniciales o modificaciones necesarias en el esquema general del proyecto.	Tolerable
R5	Retraso en el avance del proyecto	El desarrollo del proyecto toma más tiempo del que se tenía previsto.	Serio
R6	Pérdida de Información	Los documentos y demás archivos del proyecto se pierden o sufren daños.	Catastrófico
R7	Cambios en la tecnología	La tecnología sobre la cual se desarrolla el proyecto se convierte en obsoleta y requiere un cambio.	Tolerable

Tabla 2Riesgos en el desarrollo del sistema

4.2 Soluciones a los Riesgos

En el siguiente apartado se describen las propuestas que se consideran para lograr minimizar los riesgos que pueden presentarse durante el desarrollo del sistema, identificados en el apartado anterior.

R1: Capturar las grabaciones de voz del individuo teniendo en cuenta las variables de distancia, posición; de ésta manera no se evitan las variaciones pero si se reducen de manera que éstas sean pocas y no afecten al posterior procesamiento.

R2: Identificar la mayor cantidad de características que sean posibles obtener y procesar, además de la que se tenga considerada como característica principal de exclusividad.

R3: Aplicar un procesamiento exhaustivo a la grabación de manera que se puedan obtener la mayor cantidad de características necesarias para ayudar a tener una mejor diferenciación entre los patrones de cada palabra o fonema.

R4: Desarrollar en la mayor medida de lo posible el proyecto de manera modular, de tal forma que al introducir nuevos requerimientos se identifiquen los módulos que requieren ser modificados o eliminados para no tener que cambiar el sistemas totalmente.

R5: Trabajar con la mayor anticipación posible cada fase del proyecto, apegándose a la planeación previamente realizada.

R6: Realizar constantemente respaldos de toda la información perteneciente al proyecto, utilizar también el almacenamiento en la nube como una herramienta adicional para la realización del respaldo.

R7: Investigar y buscar asesoramiento sobre las nuevas tecnologías sobre las que debe adaptarse el proyecto.

4.3 Estudio de Factibilidad

A continuación se describen los análisis de factibilidad técnica, económica y operativa que implican el desarrollo y la implantación del proyecto propuesto, donde se describen las necesidades a cubrir en los ámbitos de hardware, software, costos, beneficios y operacional.

4.3.1 Factibilidad Técnica

En éste ámbito de la factibilidad se analizan los recursos de hardware y software que son necesarios para la implementación y funcionamiento del sistema, en las siguientes tablas se describen estos recursos y se especifica la cantidad requerida para el proyecto.

Cantidad	Recurso	Descripción / Características (mínimas)
3	Equipo de Cómputo	Procesador 1 GHz Memoria RAM 512 Mb Almacenamiento de 1.5 Gb Tarjeta Gráfica Pantalla de Resolución 1366 x 768 Tarjeta de red Tarjeta de audio
1	Conexión a Internet	Acceso a internet ya sea de forma alámbrica o inalámbrica.

Tabla 3Tabla de Recursos de Hardware

Recurso	Descripción
Sistema Operativo	Windows 7 u 8 y Linux
IDE (Programación)	MATLAB R2013b GNU OCTAVE
IDE (Análisis/Diseño)	Microsoft Visio 2013 EDraw

Tabla 4Recursos de Software

4.3.2 Factibilidad Económica

Para el estudio de factibilidad económica se calcula el Costo Estimado de Desarrollo del proyecto que será elaborado en un periodo de 10 meses; se toman en cuenta los gastos de tres rubros principales: hardware/software, sueldos y servicios. En relación al hardware es importante enfatizar que el equipo necesario ya se poseía y no fue adquirido para el desarrollo del proyecto; por lo tanto, únicamente se tomará en cuenta la depreciación del mismo.

Costo	Valor de Salvamento	M.A.D. (Monto A Depreciar)		M.A.D. (año)	M.A.D. (mes)	M.A.D. (10 meses)
Equipo de Cómputo 1	$ 10,000	$ 4,000	$ 6,000	$ 2,000	$ 166.67	$ 1,666.67
Equipo de Cómputo 2	$ 10,000	$ 4,000	$ 6,000	$ 2,000	$ 166.67	$ 1,666.67
Equipo de Cómputo 3	$ 10,000	$ 4,000	$ 6,000	$ 2,000	$ 166.67	$ 1,666.67
Total a Depreciar				**$ 5,000.01**		

Tabla 5Tabla de Depreciación del Hardware

	Costo	Costo Mensual	Costo para los 10 meses
MATLAB	$12,000	$1,000.00	$10,000.00
Microsoft Office 2013	$ 2,200	$ 183.33	$ 1,833.33
Microsoft Visio 2013	$ 4,000	$ 333.33	$ 3,333.33
Total		**$ 15,166.66 * 3 =45,499.98**	

Tabla 6Tabla de Depreciación de Software

Puesto	Cantidad	Sueldo por Hora	Sueldo Mensual	Sueldo para 5 meses	Sueldos
Analista	3	$ 200	$ 19,200	$ 96,000	$ 192,000
Programador	3	$ 150	$ 14,400	$ 72,000	$ 144,000
Total			**$ 336,000**		

Tabla 7Tabla de Sueldos

Costo Mensual		Costo para los 10 meses
Luz	$ 200	$ 2,000
Teléfono	$ 200	$ 2,000
Internet	$ 400	$ 4,000
Papelería	$ 100	$ 1,000
Total	**$ 9,000**	

Tabla 8Tabla de Gastos de Servicios

Concepto	Costo para los 10 meses
Depreciación del Hardware	$ 5,000.01
Gastos del Software	$ 45,499.98
Sueldos	$ 336,000
Gastos de Servicios	$ 9,000
Costo Estimado de Desarrollo	**$ 395,499.56**

Tabla 9Tabla de Costo Estimado de Desarrollo

4.3.3 Factibilidad Operativa

Para este apartado del estudio de factibilidad operativa consideramos la experiencia de los programadores en los lenguajes de programación comúnmente utilizados.

	MATLAB	OCTAVE	C++
Becerra Juan	Media	Media	Media
Martínez Sánchez Víctor	Media	Media	Alta
Farías Pineda Diego	Media	Media	Media

Tabla 10Experiencia en los Lenguajes de Programación

4.4 Análisis de Lenguajes de Programación

4.4.1 GNU Octave

Octave o GNU Octave es un programa libre para realizar cálculos numéricos. Como su nombre indica, es parte del proyecto GNU. Es considerado el equivalente libre de MATLAB. Entre varias características que comparten, se puede destacar que ambos ofrecen un intérprete, permitiendo ejecutar órdenes en modo interactivo. Nótese que Octave no es un sistema de álgebra computacional, como lo es Maxima, sino que está orientado al análisis numérico.

El proyecto fue creado alrededor del año 1988, pero con una finalidad diferente: ser utilizado en un curso de diseño de reactores químicos. Posteriormente, en el año 1992, se decidió extenderlo, y comenzó su desarrollo a cargo de John W. Eaton. La primera versión alpha fue lanzada el 4 de enero de 1993. Un año más tarde, el 17 de febrero de 1994, apareció la versión 1.0.

El nombre surge de Octave Levenspiel, profesor de uno de los autores y conocido por sus buenas aproximaciones, por medio de cálculos mentales, a problemas numéricos en ingeniería química.

Octave está escrito en C++ usando la biblioteca STL.

Tiene un intérprete de su propio lenguaje (de sintaxis casi idéntica a Matlab), y permite una ejecución interactiva o por lotes.

Su lenguaje puede ser extendido con funciones y procedimientos, por medio de módulos dinámicos.

Utiliza otros programas GNU para ofrecer al usuario crear gráficos para luego imprimirlos o guardarlos (Grace).

Dentro del lenguaje también se comporta como una consola de órdenes (shell). Esto permite listar contenidos de directorios, por ejemplo.

Además de correr en plataformas Unix también lo hace en Windows.

Puede cargar archivos con funciones de Matlab (reconocibles por la extensión .m).

La sintaxis es casi idéntica a la utilizada en MATLAB.

Es un lenguaje interpretado.

No permite pasar argumentos por referencia. Siempre son pasados por valor.

No permite punteros.

Se pueden generar scripts.

Soporta gran parte de las funciones de la biblioteca estándar de C.

Puede ser extendido para ofrecer compatibilidad con las llamadas al sistema UNIX.

El lenguaje está pensado para trabajar con matrices, y provee mucha funcionalidad para trabajar con éstas.

Soporta estructuras similares a los "struct"s de C.

Al ser su licencia Licencia pública general de GNU, puede ser compartido y utilizado libremente

4.4.2 MATLAB

ATLAB (abreviatura de MATrix LABoratory, "laboratorio de matrices") es una herramienta de software matemático que ofrece un entorno de desarrollo integrado (IDE) con un lenguaje de programación propio (lenguaje M) y servicio de especie. Está disponible para las plataformas Unix, Windows, Mac OS X y GNU/Linux .

Entre sus prestaciones básicas se hallan: la manipulación de matrices, la representación de datos y funciones, la implementación de algoritmos, la creación de interfaces de usuario (GUI) y la comunicación con programas en otros lenguajes y con otros dispositivos hardware. El paquete MATLAB dispone de dos herramientas adicionales que expanden sus prestaciones, a saber, Simulink (plataforma de simulación multidominio) y GUIDE (editor de interfaces de usuario - GUI). Además, se pueden ampliar las capacidades de MATLAB con las cajas de herramientas (toolboxes); y las de Simulink con los paquetes de bloques (blocksets).

Es un software muy usado en universidades y centros de investigación y desarrollo. En los últimos años ha aumentado el número de prestaciones, como la de programar directamente procesadores digitales de señal o crear código VHDL.

Durante mucho tiempo hubo críticas porque MATLAB es un producto propietario de The Mathworks, y los usuarios están sujetos y bloqueados al vendedor. Recientemente se ha proporcionado una herramienta adicional llamada MATLAB Builder bajo la sección de herramientas "Application Deployment" para utilizar funciones MATLAB como archivos de biblioteca que pueden ser usados con ambientes de construcción de aplicación .NET o Java. Pero la desventaja es que el computador donde la aplicación tiene que ser utilizada necesita MCR(MATLAB Component Runtime) para que los archivos MATLAB funcionen correctamente. MCR se puede distribuir libremente con los archivos de biblioteca generados por el compilador MATLAB.

LabVIEW

GNU Octave, software libre similar a matlab.

SAS

Scilab

Mathcad

SciPy & Numerical Python

Lenguaje R

Álgebra computacional:

4.5 Análisis de Requerimientos

Los requerimientos consisten en aquellas necesidades que debe cubrir el sistema para su funcionamiento. Para esta etapa de análisis del sistema, así como para la de diseño se utilizan las herramientas y conceptos que provee el Leguaje Unificado de Modelado (UML).

4.5.1 Requerimientos Funcionales

En el siguiente apartado se describen los Requerimientos funcionales, es decir, aquellas necesidades básicas que debe cubrir el sistema para su correcto funcionamiento

No.	Requerimientos Funcionales
RF1	Cargar un archivo de audio de voz.
RF2	Procesar el archivo de audio para la eliminación de ruido y otras variaciones.
RF3	Detección de inicio fin de las palabras.
RF4	Extraer las características necesarias de cada muestra y obtener los vectores de características de éstas.
RF5	Identificar cada palabra.
RF6.	Mostrar un resumen de las palabras identificadas en forma de reporte.

Tabla 11Requerimientos Funcionales

4.5.2 Requerimientos No Funcionales

A continuación se muestran los Requerimientos No Funcionales del sistema, es decir, características del sistema que no son indispensables pero que se desean cubrir.

No.	Factor	Requerimientos No Funcionales
RNF1	Mantenibilidad	El sistema debe contar con un manual técnico y de usuario.
RNF2	Escalabilidad	El desarrollo del sistema seguirá un modelo modular, de tal modo que sea posible intercambiar módulos con facilidad.
RNF3	Portabilidad	El sistema puede ser utilizado desde cualquier sistema operativo.
RNF4	Facilidad de Uso	A través de menús representativos y cuadros de diálogo, se busca la sencillez en la interacción entre el sistema y el usuario.

Tabla 12Requerimientos no funcionales.

5. Diseño

5.1 Adquisición de los archivos de voz.

El corpus de voces necesario para el sistema se generó mediante la captura de archivos de audio en formato wav, se optó por este formato debido q que es un formato de audio digital normalmente sin compresión de datos desarrollado y propiedad de Microsoft y de IBM que se utiliza para almacenar sonidos en el PC, admite archivos mono y estéreo a diversas resoluciones y velocidades de muestreo, su extensión es .wav

Cada persona realiza cincuenta repeticiones por cada palabra a reconocer, se realizan dos o más grabaciones extras para asegurar un corpus completo sin errores, se escoge un lugar aislado de ruido y empleando la tarjeta de audio de las computadoras se obtienen los archivos mediante un software desarrollado por nosotros en MATLAB y Octave el cual permite seleccionar la duración de la grabación la ruta donde se guardara, el nombre del archivo y la frecuencia de muestreo deseada, para el desarrollo de los prototipo se utilizaran las dos frecuencias más utilizadas en procesamiento de voz 8000 y 16000 Hz.

Cabe hacer mención que con el objetivo de poder obtener las mejores condiciones para una correcta grabación es necesario considerar condiciones adecuadas de captura, es decir mediante un mínimo de ruido ambiental (con o sin filtro externo), dispositivos de captura de voz de calidad (micrófono mono-aural o estéreo) y así mismo una frecuencia de muestreo adecuada.

Para este trabajo, se utilizó un corpus de grabaciones obtenidas de 3 personas adultas, en formato wav a 8000 y 16000 Hz, 16 bits, Mono, mediante las tarjeras de audio de las laptops.

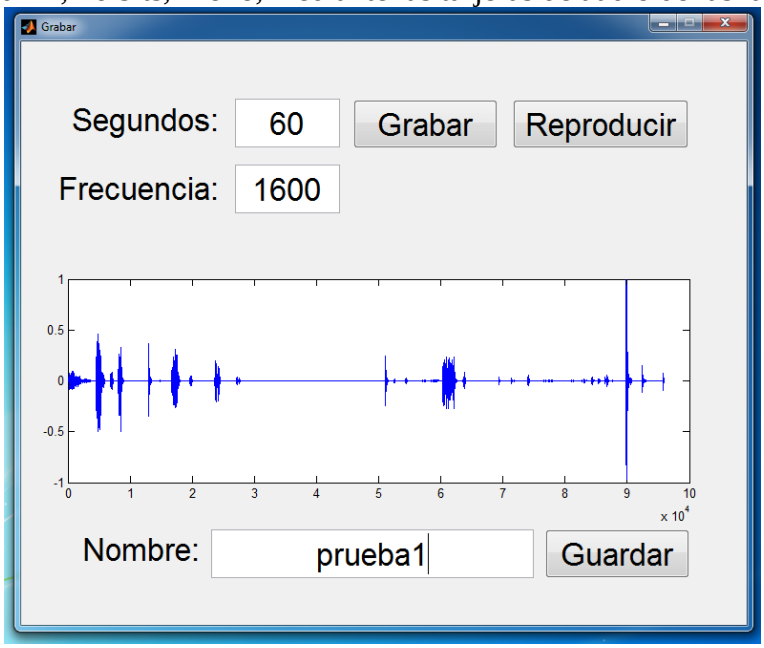

Imagen 11Software utilizado para grabar.

5.2 Los parámetros característicos de voz a emplear para el reconocimiento.

Los parámetros empleados en este trabajo fueron los Coeficientes de Predicción Lineal (LPC), y los Coeficientes Cepstrales en Frecuencia en escala de Mel (MFCC), los primeros debido a su fácil cálculo en el dominio del tiempo y su carácter de modular la producción de la voz humana

simulando el tracto vocal, los segundos por su aproximación a la percepción auditiva natural humana y su mejor representación espectral, tal y como se mencionó en capítulo anterior.

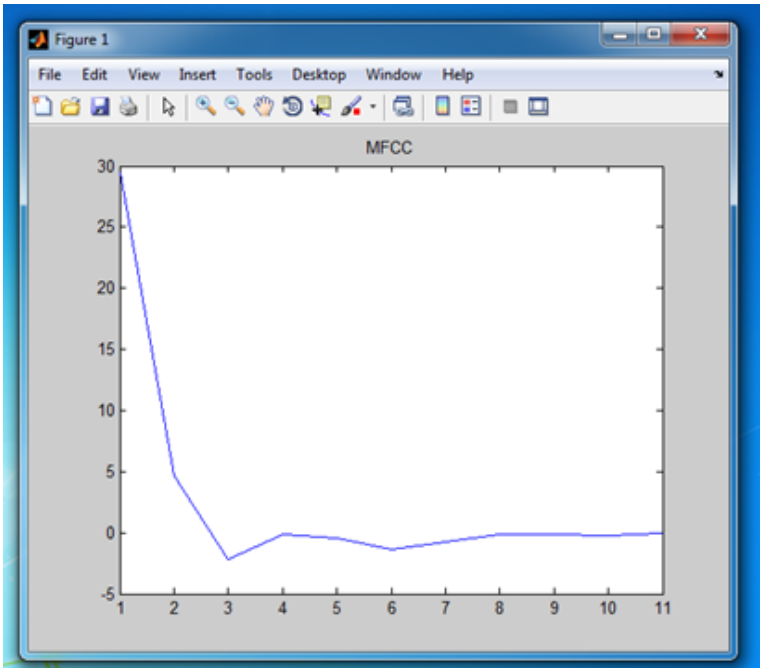

Imagen 12Calculo de MFCCs con nuestro software

5.3 Los modelos de reconocimiento de voz que se probarán.

En el caso de los modelos empleados para clasificar o reconocer los parámetros característicos se en cuentan los más usados en la actualidad, lo que no llevo a pasar desde modelos prácticos con resultados aceptables hasta unos más sofisticados con mejores resultados pero con un costo computacional también elevado.

El primero que se utilizo fue implementando una red neuronal backpropagation, perceptrón multicapa, el cual es ampliamente utilizado en el mundo del desarrollo de sistemas de reconocimiento de voz, especialmente cuando se trata de sistemas que trabajan por palabra y no por secuencia de fonemas.

El segundo modelo que se utilizo es el modelo basado en cadenas ocultas de Markov, este modelo es ampliamente utilizado debido a la capacidad que tiene de identificar entre palabras con un contenido armónico igual, tal es el caso atún y tuna, las cuales para los modelos convencionales que trabajan por voz son prácticamente imposibles de identificar.

5.4 Preprocesamiento

Dentro del procesamiento de señales es de crucial importancia que los datos a ser analizados presenten condiciones óptimas o cercanas a éstas, debido a que factores como el ruido ambiental y en ocasiones los dispositivos de captura pueden generar una baja en la calidad de la señal capturada. Cuando las condiciones de grabación no son óptimas, ya sea porque la acción de captura y/o grabado de voz se lleva a cabo dentro de las habitaciones, las cuales resultan ser sitios con ruidos

ambientales extremos, tales como conversaciones entre personas, ruidos de mascotas, televisores, radios, computadoras entre otras, es necesario emplear la etapa de preprocesamiento la cual realiza un acondicionamiento de la señal para reducir las variaciones de los patrones característicos y aumentar por lo consiguiente el porcentaje de reconocimiento del sistema.

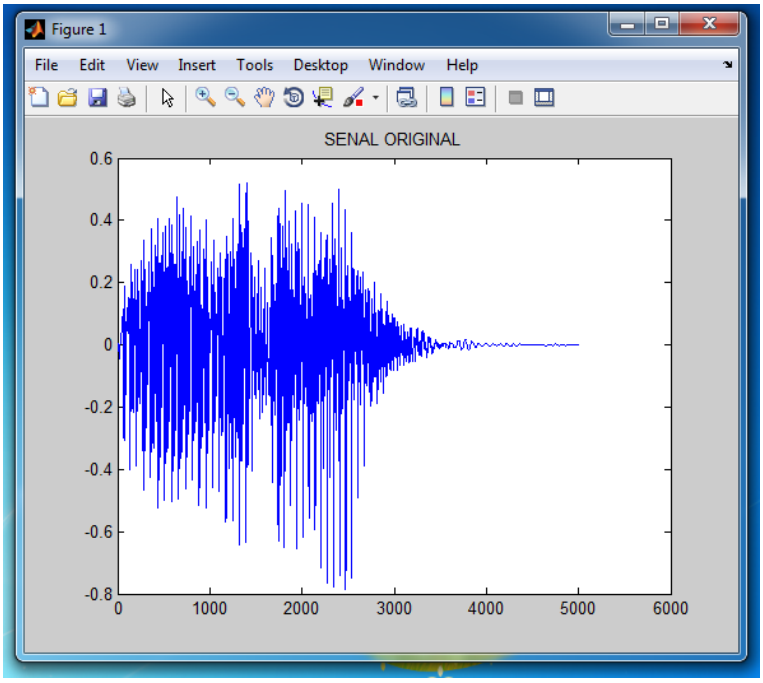

Imagen 13Señal original de la palabra cero.

5.5 Preénfasis

Este corresponde al acondicionamiento necesario para compensar la caída que sufre la señal de voz en las altas frecuencias luego de salir de los labios. El preénfasis es un filtro con características de pasa altas, donde las frecuencias bajas ligeramente salen y las superiores a 1 Khz se amplifican, es decir es posible considerar al preénfasis como el procesamiento al que se somete una señal al pasar por un filtro pasa altas con el objeto de aumentar los niveles de las frecuencias agudas para que no sean despreciadas cuando se calculen las características de la señal.

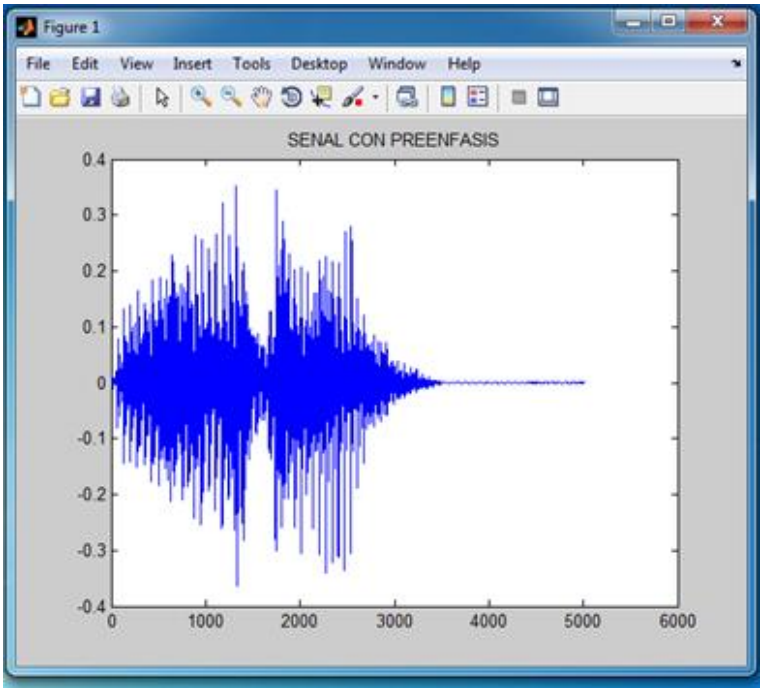

Imagen 14Señal de voz con preénfasis.

5.6 Ventaneo

Como también se menciona, el ventaneo es un procesamiento necesario para adecuar la señal de manera que los paquetes adyacentes no tengan discontinuidad al realizar el traslape de los mismos, ello con el principal interés de obtener un mayor número de segmentos y una menor pérdida de información.

La ventana utilizada en este trabajo fue la de Hamming ya que es la más usada en sistemas de reconocimiento de voz, debido a que tiene la mejor respuesta en el dominio de la frecuencia que elimina las variaciones de la señal en sus bordes.

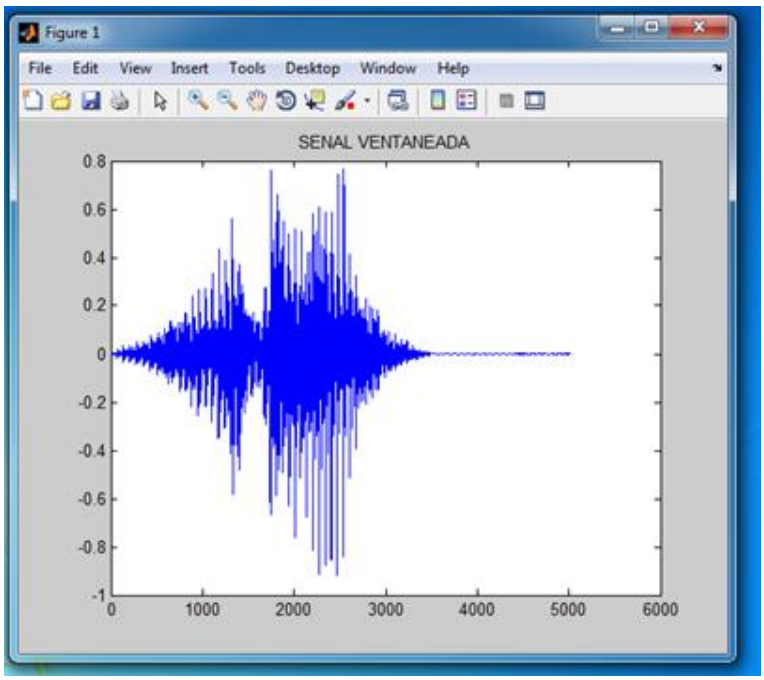

Imagen 15Ejemplo de ventaneo a una palabra completa.

5.7 Detección de inicio fin.

Debido a la naturaleza del sistema fue necesario desarrollar un módulo capas de recibir un archivo de audio wav y que como resultado no devolviera todas las palabras que se encontraran en este, para esto fue necesario programar un algoritmo de detección de inicio fin, los algoritmos y técnicas ya están prestablecidos sin embargo no son fácil mente aplicables, se necesitó de constante experimentación para poder lograr una buena separación.

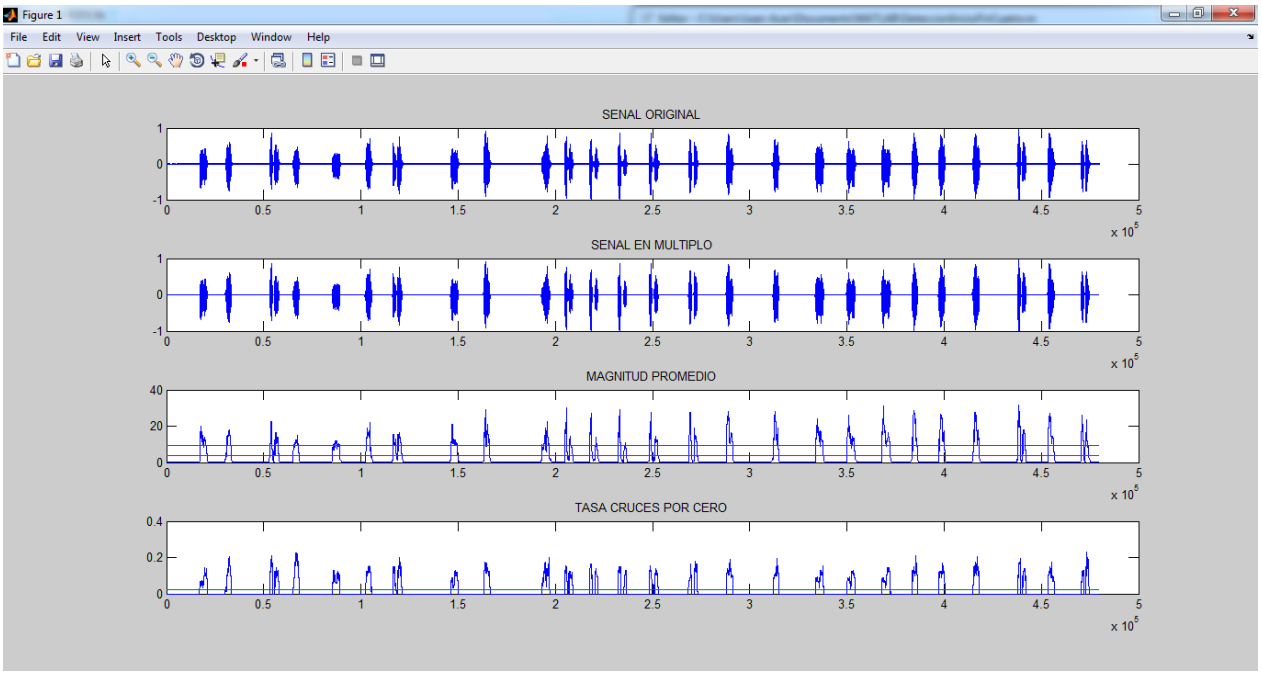

Imagen 16Software con el cual se trabajó para poder determinar los umbrales.

Como se muestra en la figura de arriba la entrada es una grabación donde se encuentran varias palabras de esa señal se calcularon las magnitudes promedios y las tazas de cruces por cero y mediante un exhaustivo análisis se logró la correcta separación.

5.8 Extracción de LPC y MFCC

Para el caso de los prototipos desarrollados por fonema fue necesario hacer uso de una de las características que tiene la señal de voz es que es cuasi-periódica, es decir que no es periódica, pero que si se toman pequeños segmentos de esta señal se puede considerar así, por tanto es común trabajar con segmentos de entre 10 y 20 ms, además de tener un traslape entre estos.

Para este trabajo se trabajaron diferentes tamaños de segmentos, con el fin de encontrar los más adecuados, en cuanto al traslape sólo se manejó del 50% y el número de coeficientes fue dado .por la frecuencia de muestreo dividido entre 1000 + 3

A continuación se mostrara las pruebas realizadas para calcular los LPC y MFCC.

En la primera ventana del cálculo de LPC se muestra la señal original y la señal con preénfasis, este es un procesamiento básico que se debe realizar con la señal antes de calcular los LPC.

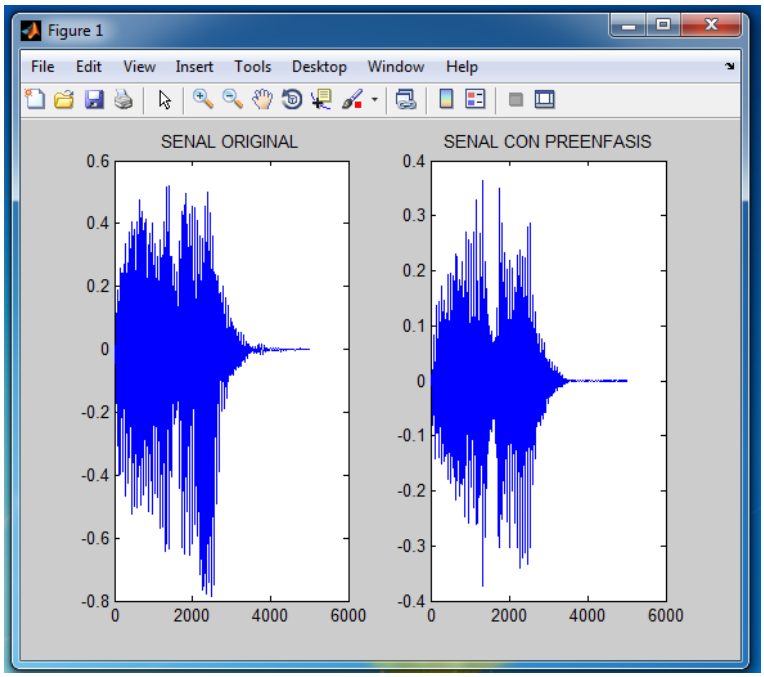

Imagen 17Calculo de LPC señal original y con preénfasis.

Después de aplicar el preénfasis se debe normalizar la señal y se le aplica una ventana.

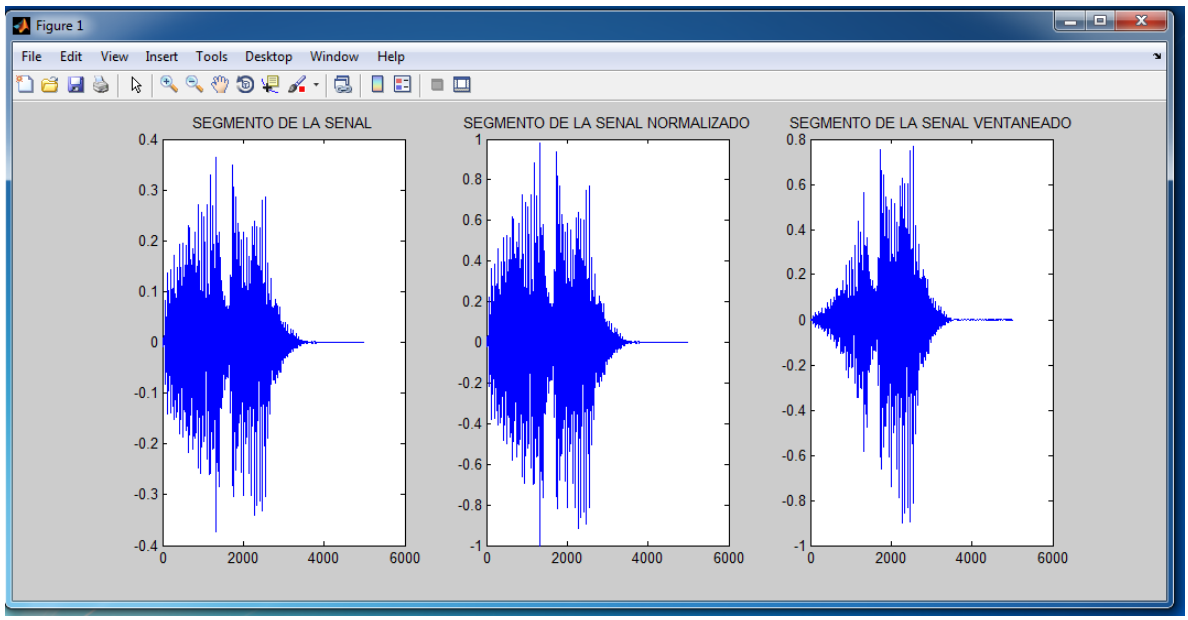

Imagen 18Señal con preénfasis, normalizada y ventaneada.

Por último se calculan los LPC, la última ventana nos muestra como en base de los coeficientes calculados se puede obtener una envolvente realmente similar al Fourier de la señal original.

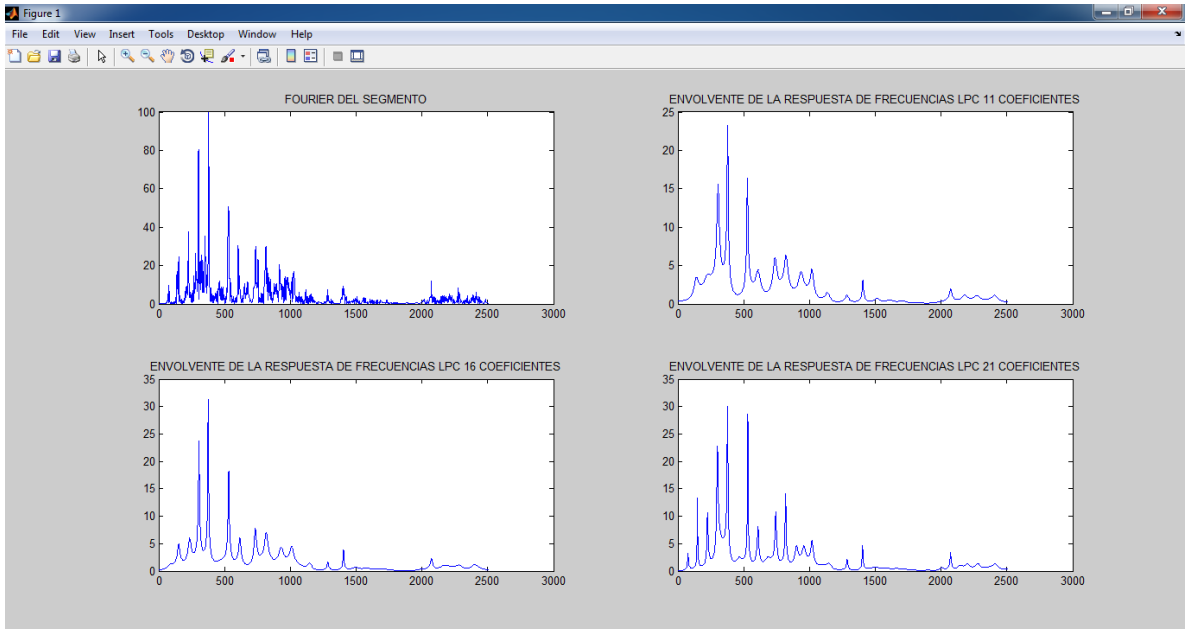

Imagen 19Envolventes obtenidas de los LPC.

Para el cálculo de los MFCC es prácticamente lo mismo, se tiene la señal original.

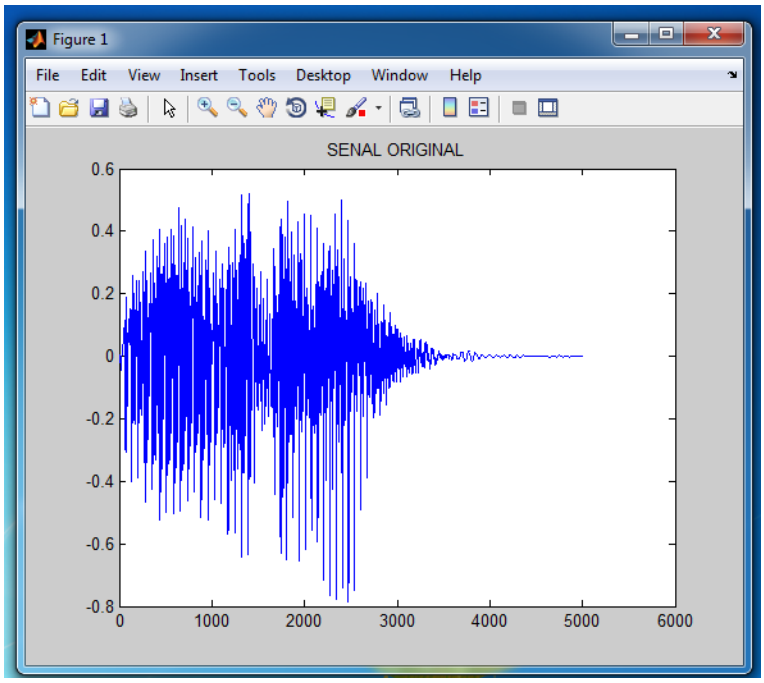

Imagen 20Señal original.

Se aplica un filtro preénfasis para mejorar la señal.

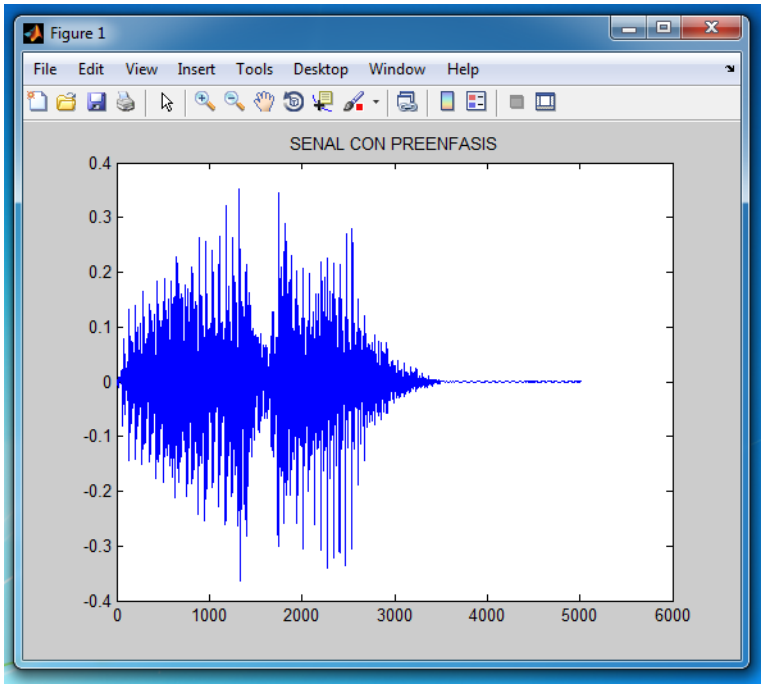

Imagen 21Señal con preénfasis.

Posterior mente se normaliza la señal de voz.

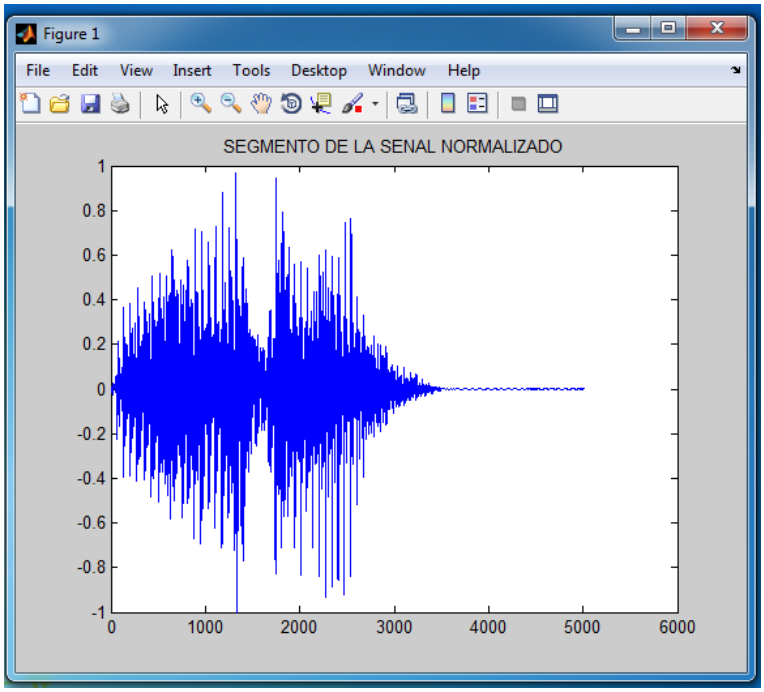

Imagen 22Señal normalizada.

Y se ventanea para evitar perdida de información.

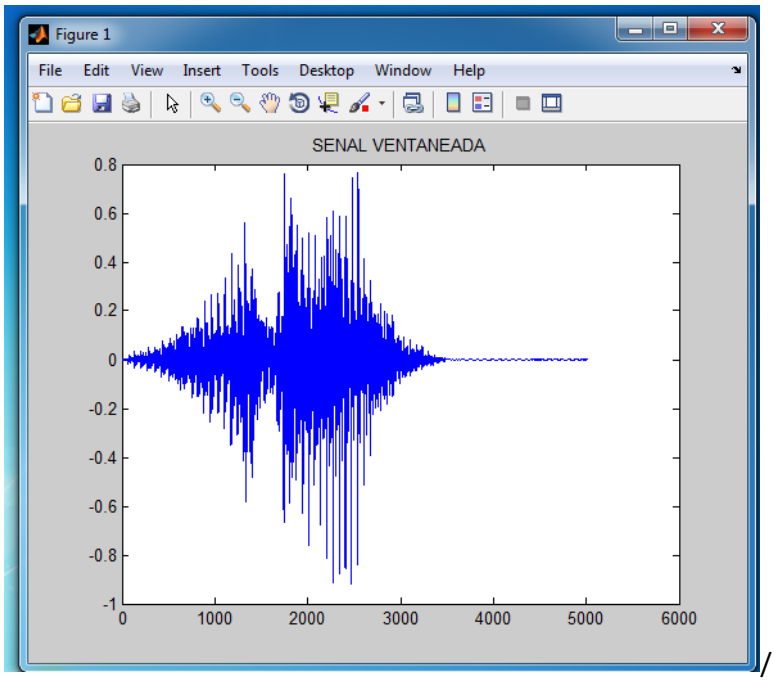

Imagen 23Señal ventaneada y lista para ser procesada.

Aquí se muestra la FFT de la señal.

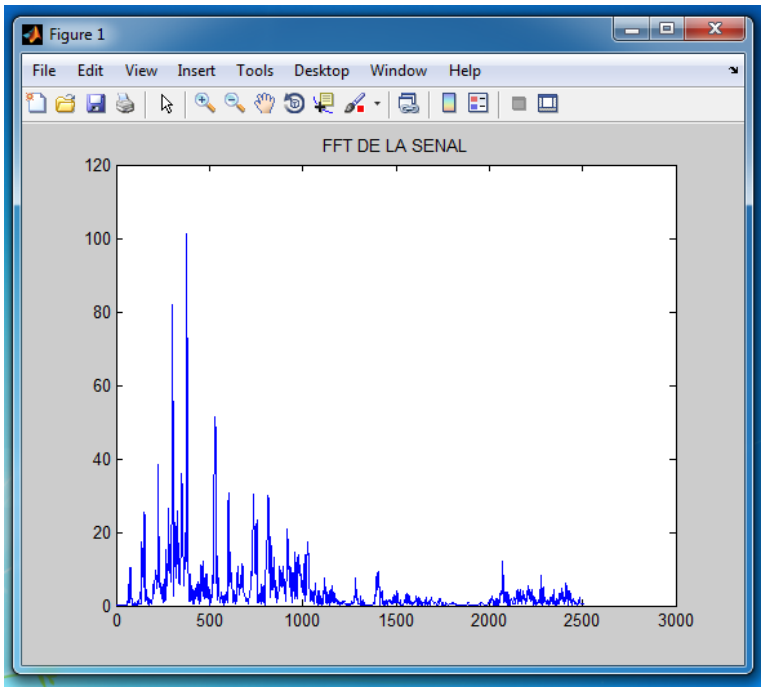

Imagen 24FFT de la señal de voz.

Aquí se calcular los filtros que se utilizaran para calcular los coeficientes.

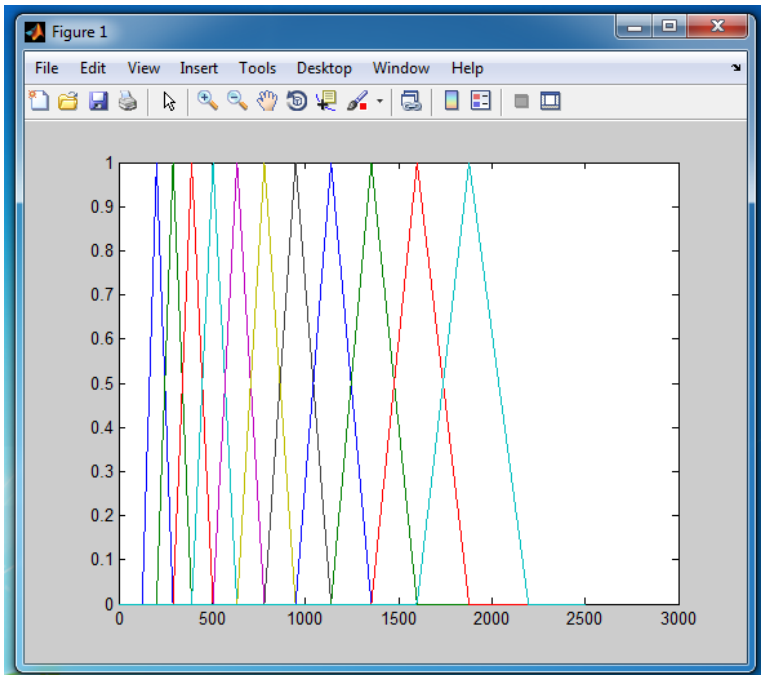

Imagen 25Filtros de Mel.

Aquí se muestran por último los MFCC.

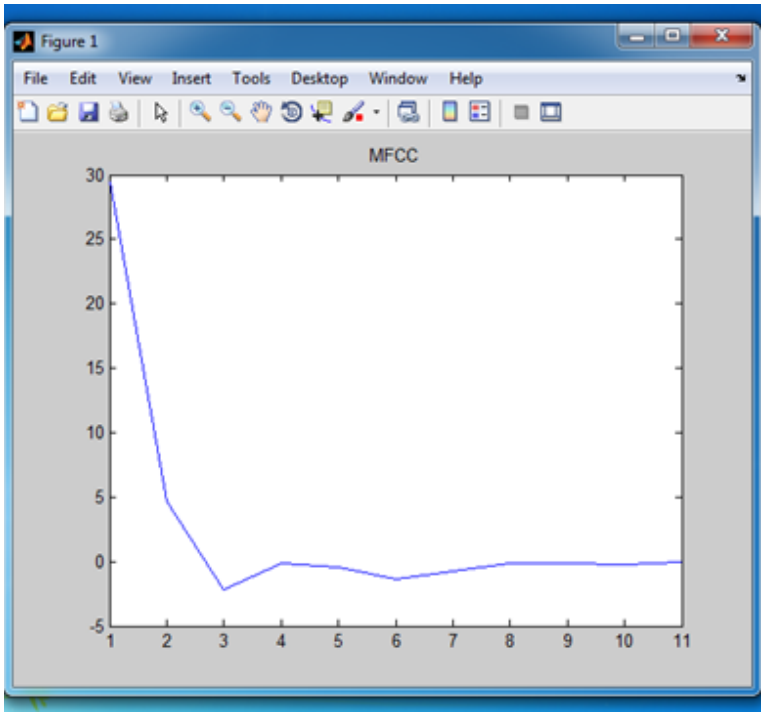

Imagen 26MFCC de una señal de voz.7

5.9 Diagrama de bloques general de un sistema de reconocimiento de voz.

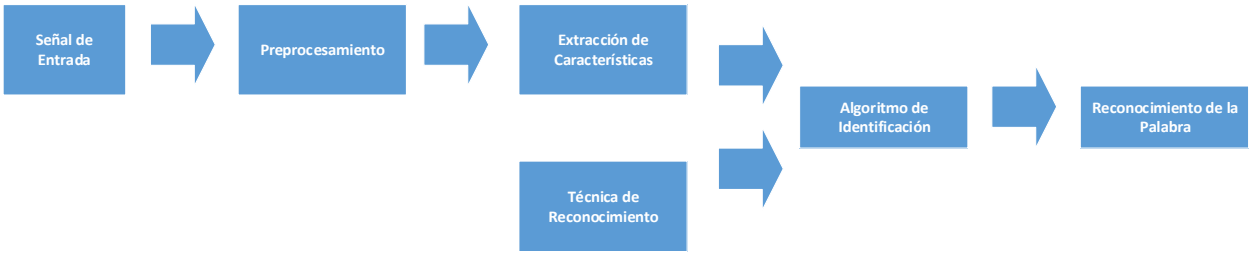

Imagen 27Diagrama de bloques general de un sistema de reconocimiento de voz.

Como podemos ver en el diagrama de bloques los sistemas de reconocimiento de voz son aptos para poder ser trabajados mediante módulos independientes los cuales en conjunto conforman un sistema complejo, el módulo de preprocesamiento permite mejorar la señal entrante, el módulo de extracción de características nos permite trabajar con los patrones que uno desee, ya sea con LPCs o con MFCCs de acuerdo al gusto del equipo de trabajo, y el modulo donde se elige la técnica de reconocimiento puede ser intercambiado de acuerdo al expertis del desarrollador.

6. Implementación y Resultados

6.1 Caso de Uso

1 Nombre

Analizar archivo de audio

2 Meta

Ofrecer una interfaz sencilla al usuario para obtener los resultados del análisis de un archivo de voz.

3 Flujo de eventos

3.1 Ruta principal

[Sistema]: El sistema presenta la interfaz principal del reconocedor de números

[Personal]: El actor selecciona la opción “Examinar”

[Personal]: El actor selecciona el archivo de audio a examinar y da clic en aceptar [flujo de excepción E01] [flujo alterno A01]

[Sistema]: El sistema analiza el archivo y presenta los resultados obtenidos

- Una gráfica con la señal de audio analizada. La señal se resalta de otro color en los intervalos donde se encuentra una palabra.
- Un registro de las palabras reconocidas
    - Palabra
    - Inicio
    - Fin

3.2 Flujos alternos

3.2.1 Opcionales

3.2.1.1 A01 Cancelar

[Personal]: El actor selecciona la opción cancelar

[Sistema]: El sistema regresa a la pantalla principal sin realizar ninguna acción

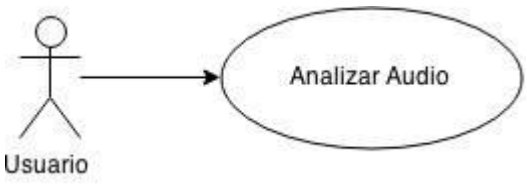

Imagen 28 Caso de Uso de la pantalla principal

6.2 Interfaz

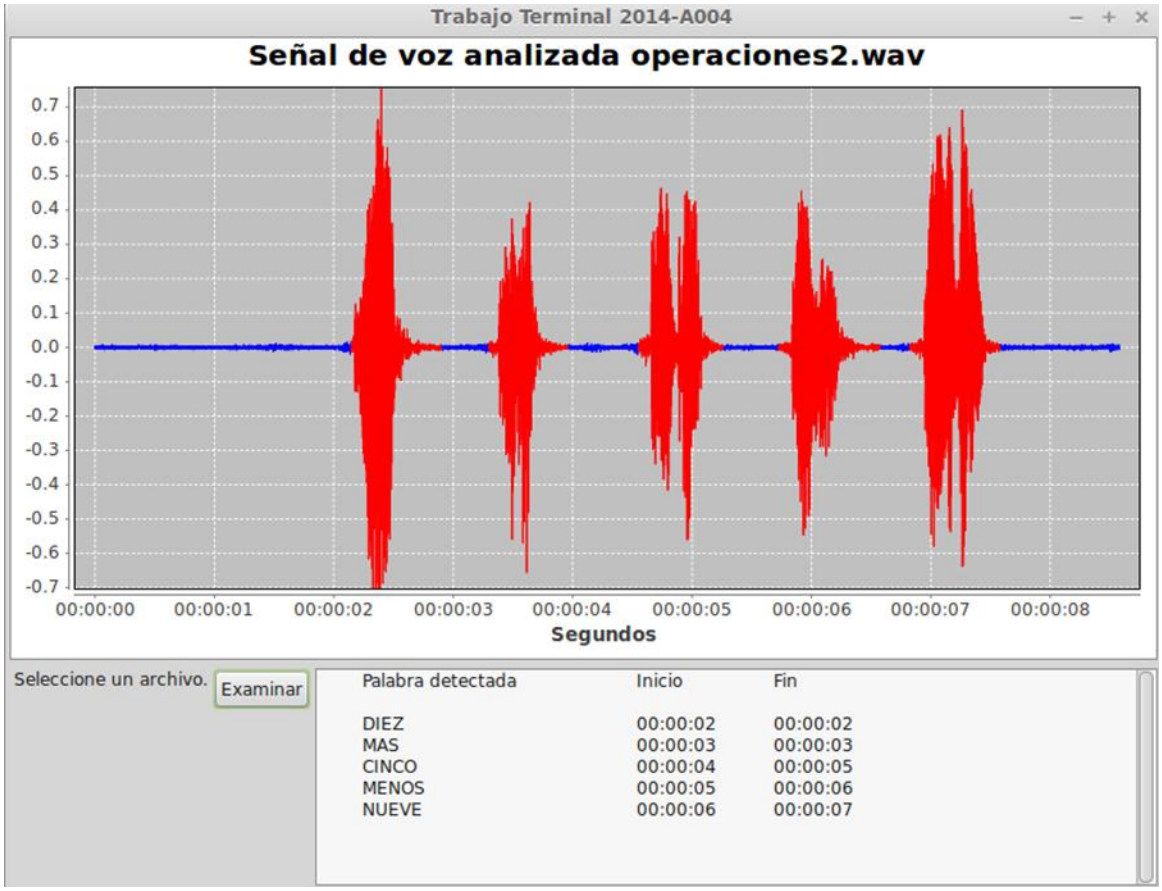

Imagen 29Pantalla principal del prototipo

Descripción

La interfaz de usuario consta de una ventana que muestra un área para mostrar la gráfica de la señal de audio analizada, en la parte inferior se muestra un botón “examinar” que lanza un navegador de archivo para seleccionar la señal de voz que se desea analizar y un área de texto donde se muestran los resultados del procesamiento de la señal.

6.3 Diagrama de Secuencia

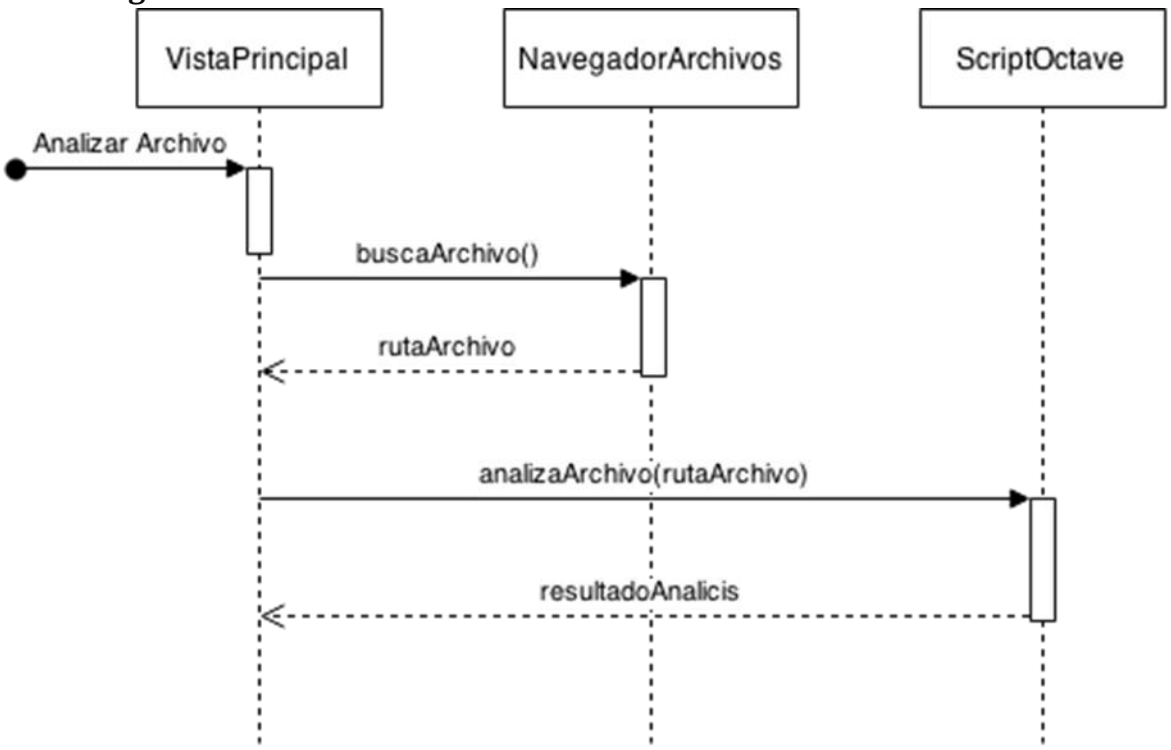

Imagen 30Diagrama de Secuencia

6.4 Diagrama de Bloques

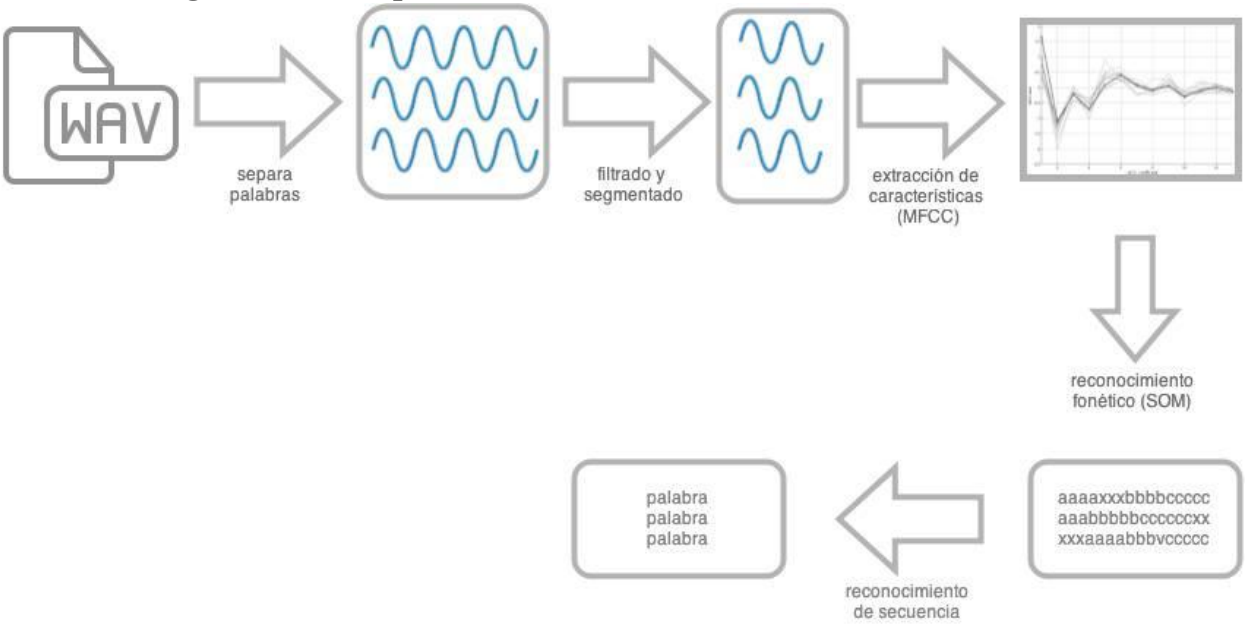

Imagen 31Diagrama de bloques del sistemas

6.5 Diagrama de flujo del algoritmo de detección de inicio y fin

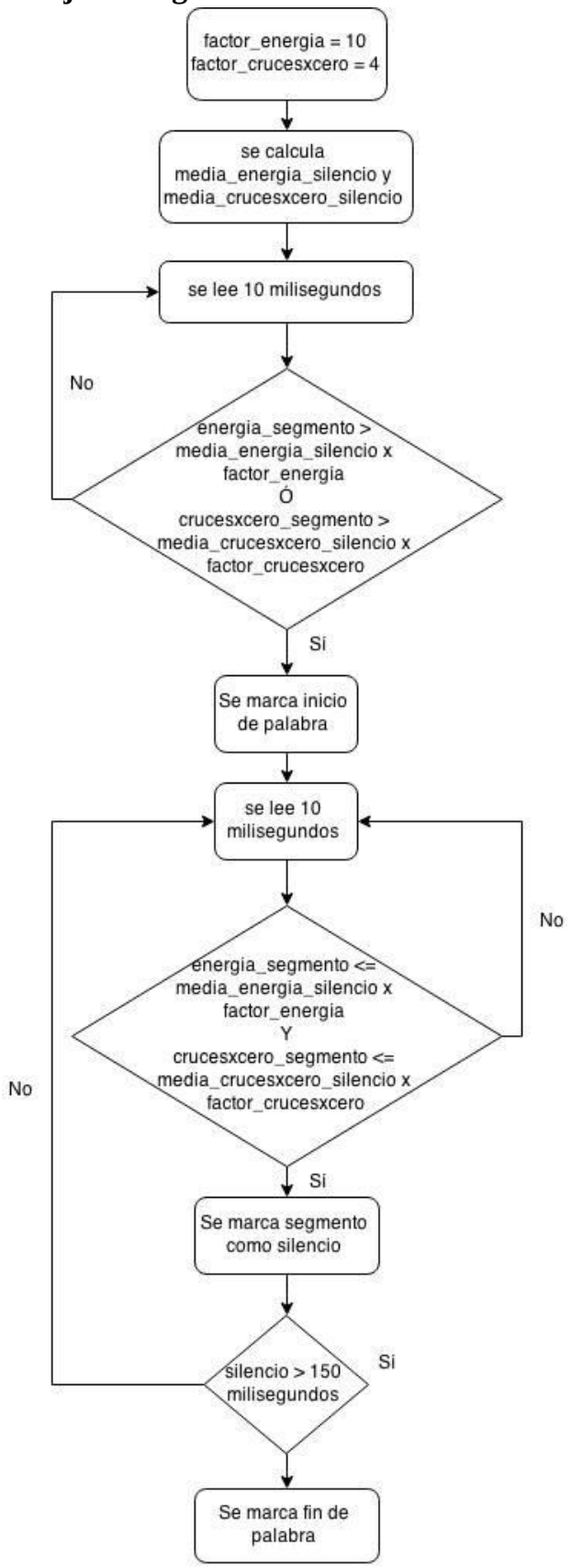

Imagen 32Diagrama de flujo detección inicio y fin

6.6 Pruebas y Resultados

La etapa de reconocimiento de secuencias del sistema se hizo usando modelos ocultos de Markov. Las secuencias se obtienen haciendo uso del reconocedor fonético. El conjunto de muestras de entrenamiento consta de alrededor de 30 muestras por cada palabra que debe reconocer el sistema. Se generaron secuencias de cada una de las muestras y para entrenar se utilizaron sólo 15 de cada palabra. Las secuencias resultantes se utilizaron para validar que el reconocedor de secuencias funciona correctamente.

Para entrenar el sistema se dividió el conjunto de muestras en muestras de entrenamiento y muestras de prueba. Se tomaron 15 muestras por cada palabra para entrenamiento y el resto para hacer pruebas. A continuación se muestran los resultados de las pruebas al reconocedor de palabras.

	CERO	UNO	DOS	TRES	CUATRO	CINCO	SEIS	SIETE	OCHO	NUEVE	DIEZ	MAS	MENOS
CERO	13												
UNO		17	1										
DOS			12										
TRES				8									
CUATRO					19								
CINCO						15		1					
SEIS							15						
SIETE								17					
OCHO			1		2				20				
NUEVE										15			1
DIEZ								1			11		1
MAS												12	
MENOS													17

Tabla 13Resultados del prototipo

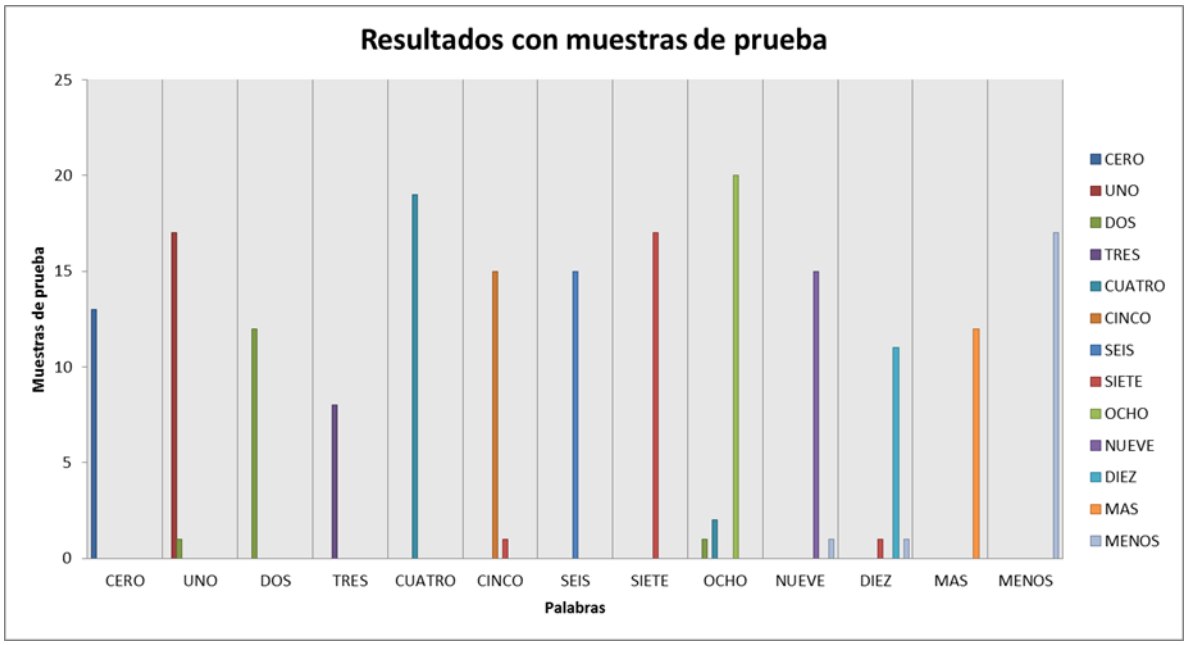

Imagen 33Gráfico de los resultados

Con el sistema entrenado se realizaron pruebas con archivos de audio nuevos. Los archivos de audio nuevos constan de varias palabras que dictan operaciones matemáticas. Los resultados de las pruebas con dichos archivos son los siguientes.

	CERO	UNO	DOS	TRES	CUATRO	CINCO	SEIS	SIETE	OCHO	NUEVE	DIEZ	MAS	MENOS
CERO	0												
UNO		5	1										1
DOS			6		1								1
TRES				7				1					
CUATRO					4								
CINCO						5							
SEIS							4	2					
SIETE								4					1
OCHO					1								
NUEVE								1		1			1
DIEZ											6		
MAS												26	2
MENOS													13

Tabla 14 Pruebas con nuevos resultados

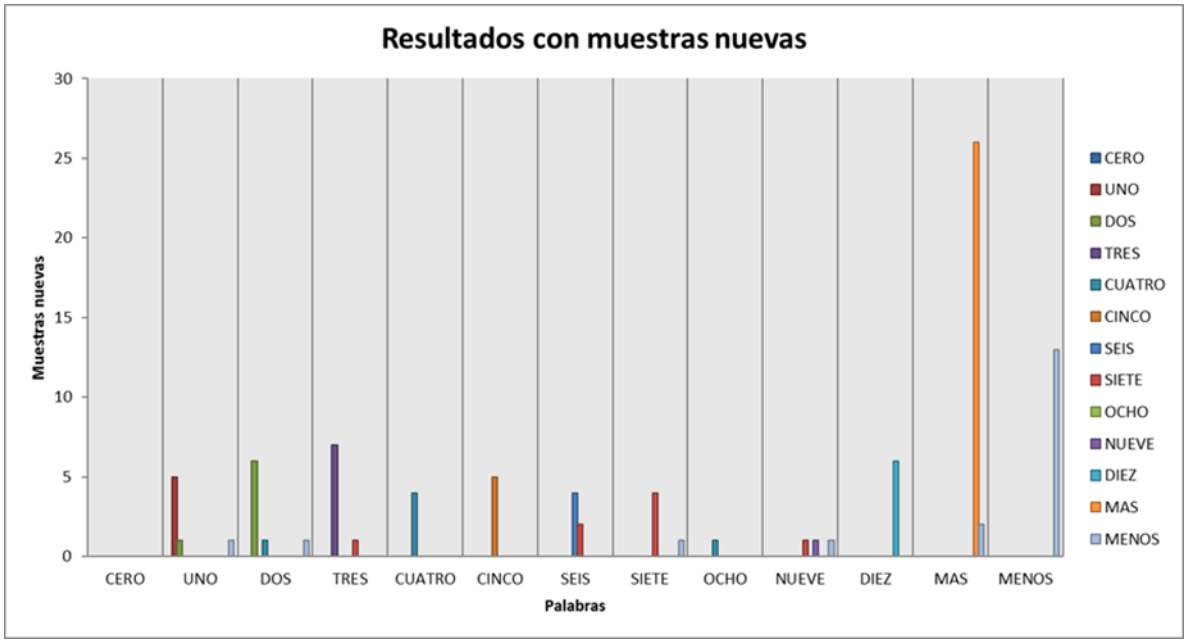

Tabla 15Gráfico con muestras nuevas

7. Conclusiones

He aquí todas nuestras conclusiones sobre el trabajo realizado, las cuales servirán de guía para las personas que en un futuro deseen incursionar en esta interesante pero difícil área de reconocimiento de voz y trabajen con reconocedores de palabras por fonema, todo lo aquí escrito es resultado de horas de trabajo y sufrimiento, este último es lo que precisamente pretendemos que evites.

El primer problema a enfrentar es la nula existencia de bases de datos de voz, en esta área la mayoría de trabajos se realizan mediante bases de conocimiento generadas con el fin de dar solución a un problema en particular, es de suma importancia tomar esto en cuenta a la hora de proponer la aplicación, pues todos los audios necesarios para trabajar tendrán que ser generados por uno mismo lo que dará como resultado tener que cada palabra sea grabada mínimo 80 veces por cada locutor que se pretenda reconocer, si tú diccionario es de n número de palabras trata de grabar 1.5n palabras distintas uno nunca sabe los problemas que se presentaran,

Es muy difícil obtener señales de voz estables, las cuales permitan tener un mejor análisis y reconocimiento. La única forma de poder controlar esto, es trabajar desde el comienzo con un ambiente controlado, el cual sea fácilmente replicable y nos permita simular las condiciones óptimas para realizar la grabación, una vez conocido el diccionario de palabras genera los audios en el ambiente controlado y genera tú base de conocimiento. Es muy importante determina el ancho de banda con el que vas a trabajar pues la tarea de grabar los audios consume mucho tiempo, de no saber con qué ancho de banda trabajar graba con el ancho de banda más alto que puedas y después podrás utilizar herramientas para editar los audios.

Estudia tu diccionario de palabras, ve que tan difícil son los fonemas que componen las palabras de tu diccionario, al inicio esto parece difícil y solo serás capaz de percibir la dificultad de reconocer los fonemas atreves del análisis visual de las señales, mediante herramientas que te permitan ver de manera gráfica los coeficientes LPC y MFCC y comparar los de un fonema con otro, descarta fonemas o incluye nuevos todo esto es experimentación.

Prueba diferentes tipos de coeficientes LPC, MFCC, etc. Los estudios han demostrado que los MFCC generar un mejor reconocimiento, sin embargo es siempre bueno experimentar para poder ver la calidad de tus coeficientes, en nuestro caso tuvimos que generar una variación en este paso haciendo uso de una correlación lo cual nos permitió obtener unos coeficientes que diferenciaran mejor un fonema de otro

A la hora de clasificar hay dos caminos el supervisado y el no supervisado, sea cual sea tu algoritmo de reconocimiento (K-NN, Árbol de decisión, Red neuronal, Maquina Vector Soporte, Algoritmo Probabilístico) no hay formula o teorema que te diga que método clasificara de mejor manera, tendrás que experimentar con todos los que te sean posibles, a veces los resultados no serán los esperados y tendrás que regresar al preprocesamiento de la señal o a la obtención de los coeficientes con el fin de obtener mejores resultados, si trabajas con un método supervisado será más tedioso y tardado pues tendrás que realizar un etiquetado manual de los fonemas y por si fuera poco no siempre los resultados serán mejores, en nuestro caso experimentamos con una red neuronal multicapa, la cual nos dio muy buenos resultados al trabajar con pocos fonemas pero al ir trabajando con un mayor número de fonemas la cosa se dificulto e increíblemente un mapa auto

organizado nos arrojó mejores resultados, el mapa autoorganizado nos permitió automatizar todo el proceso hasta este punto y poder hacer más pruebas y experimentos.

Al final de todo este trabajo ya se tendrá el libro código y se podrán generar las cadenas de caracteres de los audios de voz, ahora tendrás que elegir un método para reconocer las palabras en nuestro caso utilizamos Markov e hicimos pruebas con DTW, los resultados arrojados por Markov fueron muy buenos y decidimos quedarnos con este.

Todos los algoritmos desarrollados pueden ser optimizados, el trabajo con señales de voz dependiendo de las técnicas que se utilizan y el ancho de banda seleccionado puede ser muy pesado computacionalmente hablando, nosotros logramos mejorar el rendimiento de nuestra aplicación haciendo uso de un paradigma de programación aceptado por Octave, programación vectorizada, en algunos casos será necesario el uso de hilos o procesos para mejorar el rendimiento de las aplicaciones.

No queda más que decir que como en cualquier trabajo en el cual esta inmiscuido el reconocimiento de patrones, el experimentar por uno mismo es la mejor forma de entender el problema y encontrar la mejor solución, los algoritmos y metodologías están en los libros y son fieles guías de lo que uno debe hacer, sin embargo no siempre son replicables en el mundo real, herramientas como Matlab o Octave facilitan el desarrollo de prototipos funcionales que con otras herramientas consumirían demasiado tiempo y deben ser ampliamente tomados en cuenta para realizar trabajos en esta área.

Por ultimo nos dimos cuenta que el campo de investigación en esta área es amplio y de gran proyección, hoy en día más que nunca es un área que merece mucho trabajo y dedicación con el fin de obtener mejores resultados y crear aplicaciones más robustas, la inversión por grupos privados crece cada día más y en comparación con otras áreas falta mucho que investigar lo cual genera una área de oportunidad.

8. Trabajo a futuro

En trabajos a futuro que se pretenden desarrollar se encuentran varios puntos interesantes que harían más robusta la herramienta, a continuación algo de lo que se pretende realizar:

Extender el vocabulario: Es uno de los puntos clave dentro de cualquier sistema de reconocimiento de voz, el que tan extenso sea el vocabulario, al desarrollar el prototipo en base a Markov permite ir agregando los modelos de las palabras sin la necesidad de tener que volver a entrenar todo el sistema, esto lo hace altamente escalable.

Extenderlo a dos locutores: Es una de las opciones más ambiciosas pues esto permitiría analizar conversaciones en ambientes controlados con alta fuente de información por ejemplo llamadas telefónicas, locuciones de radio o conversaciones en lugares cerrados.

Identificación de locutores: Esto permitiría saber quién dice que y cuando, permitiría dar seguimiento a conversaciones y obtener información muy valiosa.

Optimización mediante hilos: Este trabajo se centraría en optimizar los tiempos de análisis de las grabaciones de audio, punto central y de gran aporte al sistema, seria impresionante poder analizar conversaciones de horas en cuestiones de segundos.

Módulo de entrenamiento para locutores: Que el sistema solo sea entrenado en cuestiones de segundos y que pueda ser utilizado por dos nuevas personas haría que el sistema adquiriera gran valor comercial y pudiera ser explotado.

Mas robusto a ruido: Permitiría que funcionara fuera de ambientes controlados, lo que generaría una amplia gama de nuevas aplicaciones.

9. Referencias

[1] I. d. J. B. Jiménez, Búsqueda de información usando entradas de voz, México, 2007.

[2] J. N. A. Rivero, Robot Dirigido Por Voz., México, 2007.

[3] N. A. García, Verificación de Pronunciación Basada en Tecnología de Reconocimiento de Voz para un Ambiente de Aprendizaje, México, 1999.

[4] A. A. Larios, Diccionario español/inglés para el aprendizaje de vocabulario utilizando una interfaz de voz., México, 1999.

[5] V. P. García, Reconocimiento de palabras clave en conversaciones espontáneas en castellano, España , 2008.

[6] B. M. M. Avendaño, Implementación de un reconocedor de voz gratuito a el sistema de ayuda a invidentes Dos-Vox en español, México, 2004.

[7] D. A. F. Rodríguez, Estado del arte en el reconocimiento, Colombia, 2005.

[8] H. B. G.E. Paterson, «Control methods used in a study of the vowels,» de *Jornal of the Acustic Society of America*.

[9] M. S.-K. B. Lindblom, «On the role of formant transitions in vowel recognition,» *Jornal of the Acustic Society of America,* vol. 42, pp. 830-843, 1967.

[10] R. F. y. G. H. José R. Calvo, *Métodos de Extracción, Selección y Clasificación de Rasgos Acústicos para el Reconocimiento del Locutor,* Habana, Cuba.

[11] L. R. R. &. R. W. Schafer, Digital Procesing of Speech Signals, New Jersey: Prentice-Hall, 1978.