SISTEMA DE RECONOCIMIENTO DEL HABLA PARA UN SUBCONJUNTO DE VOCALES DE LA LENGUA NASA YUWE VARIANTE CORINTO LOPEZ ADENTRO CAUCA

LEONARDO JAVIER BASTIDAS MORENO
JAVIER IGNACIO CAICEDO SAMBONI

Monografía para optar al título de
Ingeniero de Sistemas

Director
Ing. Roberto Carlos Naranjo Cuervo

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Sistemas
Popayán, diciembre 10 de 2010

TABLA DE CONTENIDO

INTRODUCCIÓN _____ *1*

OBJETIVO GENERAL _____ 2

OBJETIVOS ESPECÍFICOS _____ 2

1. MARCO TEÓRICO _____ *4*

1.1. PRODUCCIÓN Y RECONOCIMIENTO DEL HABLA, CÓMO SE VE DESDE LA PERSPECTIVA COMPUTACIONAL _____ 4

1.2. FUNCIONAMIENTO DE UN SISTEMA DE RECONOCIMIENTO DEL HABLA _____ 5

1.2.1. ARQUITECTURA DE UN SISTEMA DE RECONOCIMIENTO DEL HABLA BASADO EN PATRONES [5, 6, 7] 5

1.3. PRE PROCESAMIENTO DE LA SEÑAL DEL HABLA _____ 7

1.3.1. MODELOS DE ANÁLISIS ESPECTRALES _____ 8

1.3.1.1. CODIFICACIÓN POR PREDICCIÓN LINEAL (LPC) _____ 8

1.3.1.2. ANÁLISIS DE COEFICIENTES CEPSTRALES DE MEL [17] _____ 10

1.4. APROXIMACIONES EN RECONOCIMIENTO DEL HABLA POR MEDIO DE MÁQUINAS __ 11

1.4.1. ENFOQUE ACÚSTICO FONÉTICO [1, 12] _____ 11

1.4.2. ENFOQUE DE PATRONES [1, 5] _____ 12

1.4.2.1. ALINEAMIENTO DINÁMICO EN EL TIEMPO DTW (DYNAMIC TIME WARPING) [1, 5, 12] __ 14

1.4.2.2. MODELOS ESTADÍSTICOS BASADOS EN MODELOS OCULTOS DE MARKOV [5] _____ 15

1.4.3. ENFOQUE DE LA INTELIGENCIA ARTIFICIAL [22] _____ 17

1.4.3.1. REDES NEURONALES ARTIFICIALES [22] _____ 18

1.5. ESTADO DEL ARTE _____ 20

1.5.1. SISTEMAS DE RECONOCIMIENTO DEL HABLA PARA PALABRAS AISLADAS [26] _____ 20

1.5.2. PROYECTOS RELACIONADOS CON EL TRATAMIENTO DE SEÑALES DE VOZ _____ 22

2. CONSTRUCCIÓN DEL CORPUS _____ *25*

2.1. CRITERIOS DE ELECCIÓN DE LAS VOCALES _____ 25

2.2. NIVELES DE ETIQUETACIÓN _____ 26

2.3.1. CARACTERÍSTICAS LINGÜÍSTICAS _____ 27

2.3.2. CARACTERÍSTICAS SOCIO-LINGÜÍSTICAS DE LOS HABLANTES _____ 27

2.3.3. CARACTERÍSTICAS TÉCNICAS _____ 28

2.4. ESTADÍSTICAS _____ 29

2.5. EXTRACCIÓN DE CARACTERÍSTICAS _____ 30

3. ANÁLISIS Y EXPERIMENTACIÓN DE LAS TÉCNICAS DE RECONOCIMIENTO DEL HABLA 32

3.1. RESULTADOS MEDIANTE LA TÉCNICA DYNAMIC TIME WARPING _____ 34

3.2. RESULTADOS MEDIANTE LA TÉCNICA REDES NEURONALES ARTIFICIALES _____ 35

3.3. RESULTADOS MEDIANTE LA TÉCNICA MODELOS OCULTOS DE MARKOV _____ 38

3.4. RESULTADOS _____ 44

4. ADAPTACIÓN DEL MODELO Y CONSTRUCCIÓN DEL PROTOTIPO _____ 50

4.1. MODELADO DEL SISTEMA DE RECONOCIMIENTO DEL HABLA _____ 50

4.1.1. INICIO _____ 50

4.1.2. ELABORACIÓN _____ 69

4.1.3. CONSTRUCCIÓN _____ 70

4.1.4. TRANSICIÓN _____ 72

5. DIFICULTADES PRESENTADAS Y SOLUCIONES PLANTEADAS DURANTE EL DESARROLLO DEL PROYECTO _____ 75

6. CONCLUSIONES Y RECOMENDACIONES _____ 76

6.1. CONCLUSIONES _____ 76

6.2. RECOMENDACIONES Y TRABAJO FUTURO _____ 77

BIBLIOGRAFIA _____ 78

ÍNDICE DE FIGURAS

FIGURA 1. ESQUEMA DE PRODUCCIÓN Y PERCEPCIÓN DEL PROCESO DEL HABLA [1]..... 4
FIGURA 2. ESQUEMA GENERAL DE UN SISTEMA DE RECONOCIMIENTO DEL HABLA BASADA EN PATRONES [1,5] 5
FIGURA 3. ESQUEMA GENERAL DE LA ETAPA DE ANÁLISIS [1] 7
FIGURA 4. ESQUEMA DE BLOQUES DEL PROCESO DE PREDICCIÓN LINEAL [1] 8
FIGURA 5. CARACTERÍSTICAS DE UN ESPECTROGRAMA DE ONDA OBSERVADAS EN PRAAT [14] 8
FIGURA 6. SÍNTESIS DEL HABLA BASADA EN EL MODELO LPC [1] 9
FIGURA 7. EXTRACCIÓN DE CARACTERÍSTICAS [16] 10
FIGURA 8. ESQUEMA DEL ANÁLISIS CEPSTRAL [17] 10
FIGURA 9. DIAGRAMA ACÚSTICO - FONÉTICO DE UN SISTEMA DE RECONOCIMIENTO DEL HABLA [1] 11
FIGURA 10. PROCESO DE FUNDAMENTAL DE RECONOCIMIENTO DEL HABLA [1] 13
FIGURA 11. TÉCNICA DE PROGRAMACIÓN DINÁMICA CON EL ALGORITMO DTW [8]..... 14
FIGURA 12. ESQUEMA GENERAL DE UN SISTEMA DE RECONOCIMIENTO DEL HABLA BASADO EN MODELOS OCULTOS DE MARKOV [5] 16
FIGURA 13. PROCESO DISCRETO DE MARKOV [5] 16
FIGURA 14. MÓDULOS DE UN SISTEMA EXPERTO, ORGANIZACIÓN JERÁRQUICA [18]..... 17
FIGURA 15. DIAGRAMA EN BLOQUES DEL APRENDIZAJE SUPERVISADO [22] 18
FIGURA 16. MODELO DE UNA RED NEURONAL DE BASE RADIAL [24] 19
FIGURA 17. FORMA DE ONDA DE LA PALABRA ẼS: PIOJO, SELECCIONADA LA Ẽ NASAL ... 26
FIGURA 18. FORMA DE ONDA DE LA PALABRA KYÃDUU: RODEAR..... 26
FIGURA 19. FORMA DE ONDA Y ESPECTROGRAMA DE LA SEÑAL DE VOZ DE LA PALABRA KYÃDUU (RODEAR)..... 27
FIGURA 20. CANTIDAD DE VOCALES ORALES POR HABLANTE..... 29
FIGURA 21. CANTIDAD DE VOCALES NASALES POR HABLANTE..... 29
FIGURA 22. TOTAL DE VOCALES ORALES Y NASALES 30
FIGURA 23. VENTANEO DE LA SEÑAL [11]..... 30
FIGURA 24. CORPUS VOCALES ORALES Y NASALES DE LA LENGUA NASA YUWE 31
FIGURA 25. COEFICIENTES LPC 31
FIGURA 26. RESUMEN DE EXPERIMENTOS 1,2 Y 3 CON DTW 35
FIGURA 27. EXPERIMENTO 1 CON LA TÉCNICA MODELOS OCULTOS DE MARKOV..... 39
FIGURA 28. EXPERIMENTO 2 CON LA TÉCNICA MODELOS OCULTOS DE MARKOV..... 41
FIGURA 29. EXPERIMENTO 3 CON LA TÉCNICA MODELOS OCULTOS DE MARKOV..... 43
FIGURA 30. EXPERIMENTO 1 CON LA TÉCNICA REDES DE BASE RADIAL UTILIZANDO EL 70% PARA ENTRENAMIENTO 45
FIGURA 31. EXPERIMENTO 1 CON LA TÉCNICA REDES DE BASE RADIAL UTILIZANDO EL 80% PARA ENTRENAMIENTO 45
FIGURA 32. EXPERIMENTO 1 CON LA TÉCNICA REDES DE BASE RADIAL UTILIZANDO EL 90% PARA ENTRENAMIENTO 46
FIGURA 33. EXPERIMENTO 2 CON LA TÉCNICA REDES DE BASE RADIAL UTILIZANDO EL 70% PARA ENTRENAMIENTO 46
FIGURA 34. EXPERIMENTO 2 CON LA TÉCNICA REDES DE BASE RADIAL UTILIZANDO EL 80% PARA ENTRENAMIENTO 47

FIGURA 35. EXPERIMENTO 2 CON LA TÉCNICA REDES DE BASE RADIAL UTILIZANDO EL 90% PARA ENTRENAMIENTO 47
FIGURA 36. EXPERIMENTO 3 CON LA TÉCNICA REDES DE BASE RADIAL UTILIZANDO EL 70% PARA ENTRENAMIENTO 48
FIGURA 37. EXPERIMENTO 3 CON LA TÉCNICA REDES DE BASE RADIAL UTILIZANDO EL 80% PARA ENTRENAMIENTO 48
FIGURA 38. EXPERIMENTO 3 CON LA TÉCNICA REDES DE BASE RADIAL UTILIZANDO EL 90% PARA ENTRENAMIENTO 49
FIGURA 39. VISTA GLOBAL DEL SISTEMA 55
FIGURA 40. CASO DE USO DE ENTRENAMIENTO DEL PROTOTIPO 55
FIGURA 41. CASO DE USO DE PRUEBAS DEL PROTOTIPO 56
FIGURA 43. DIAGRAMA DE CLASES DEL PROTOTIPO DE RECONOCIMIENTO DE LA LENGUA NASA YUWE 59
FIGURA 44. DIAGRAMA DE SECUENCIA DEL ENTRENAMIENTO DEL PROTOTIPO DE RECONOCIMIENTO 61
FIGURA 45. DIAGRAMA DE SECUENCIA DE PRUEBAS DEL PROTOTIPO DE RECONOCIMIENTO 62
FIGURA 46. DIAGRAMA DE FLUJO PARA CARGAR DATOS 63
FIGURA 47. DIAGRAMA DE FLUJO PARA CALCULAR LAS CARACTERÍSTICAS DE LA VOCAL 64
FIGURA 48. DIAGRAMA DE FLUJO PARA CALCULAR LOS PATRONES 65
FIGURA 49. DIAGRAMA DE FLUJO DE LA CLASIFICACIÓN DE LAS ONDAS DE PRUEBA 66
FIGURA 50. DIAGRAMA DE FLUJO PARA CALCULAR EL ERROR DE RECONOCIMIENTO 67
FIGURA 51. ADAPTACIÓN DE LA ARQUITECTURA DE FILTROS Y TUBERÍAS PARA EL PROTOTIPO DE RECONOCIMIENTO 68
FIGURA 52. INTERFAZ DEL PROTOTIPO DEL SISTEMA DE RECONOCIMIENTO 71
FIGURA 53. APLICACIÓN DE LA PRUEBA (1) 74
FIGURA 54. APLICACIÓN DE LA PRUEBA (2) 74

ÍNDICE DE TABLAS

TABLA 1. FONEMAS DEL NASA YUWE (VOCALES) [39] 25
TABLA 2. CARACTERÍSTICAS SOCIO-LINGÜÍSTICAS DE LOS HABLANTES 28
TABLA 3. EXPERIMENTO 1 CON LA TÉCNICA DYNAMIC TIME WARPING 34
TABLA 4. EXPERIMENTO 2 CON LA TÉCNICA DYNAMIC TIME WARPING 34
TABLA 5. EXPERIMENTO 3 CON LA TÉCNICA DYNAMIC TIME WARPING 34
TABLA 6. EXPERIMENTO 1 CON LA TÉCNICA REDES NEURONALES ARTIFICIALES 35
TABLA 7. EXPERIMENTO 2 CON LA TÉCNICA REDES NEURONALES ARTIFICIALES 36
TABLA 8. EXPERIMENTO 3 CON LA TÉCNICA REDES NEURONALES ARTIFICIALES 36
TABLA 9. EXPERIMENTO 1 CON LA TÉCNICA MODELOS OCULTOS DE MARKOV 38
TABLA 10. EXPERIMENTO 2 CON LA TÉCNICA MODELOS OCULTOS DE MARKOV 40
TABLA 11. EXPERIMENTO 3 CON LA TÉCNICA MODELOS OCULTOS DE MARKOV 42
TABLA 12. RESUMEN CON LAS TÉCNICAS DTW, RNA Y HMM 44
TABLA 13. REQUISITO CREAR MODULO CAPTURA DE SEÑAL DE VOZ 50
TABLA 14. REQUISITO CREAR MODULO DE PREÉNFASIS 51
TABLA 15. REQUISITO CREAR MODULO DE CLASIFICACIÓN 51
TABLA 16. REQUISITO CREAR INTERFAZ GRAFICA DE ASR 51
TABLA 17. RIESGOS POR REQUISITO 52
TABLA 18. RIESGOS POR TIPOS 52
TABLA 19. USUARIOS DEL SISTEMA 54
TABLA 20. CASO DE USO REAL GRABAR SEÑAL DE VOZ 57
TABLA 21. CASO DE USO REAL RECONOCIMIENTO DE LA SEÑAL DE VOZ 58
TABLA 22. DISEÑO DE PRUEBAS EN VIVO PARA EL PROTOTIPO DE RECONOCIMIENTO 72

INTRODUCCIÓN

Este trabajo se planteó debido a que no se encontraron antecedentes en lo que se refiere al reconocimiento del habla para la lengua Nasa Yuwe. Al no haber un estudio de reconocimiento automático del habla ASR (Automatic speech reconigtion) para esta lengua. La presente propuesta servirá para futuros estudios y desarrollo de aplicaciones enfocadas al reconocimiento automático del habla para la lengua Nasa Yuwe.

El reconocimiento del habla ha sido un objetivo de varios investigadores, por más de cinco décadas. Dentro de los avances logrados se observa una interacción interdisciplinaria que se aproxima a la solución, desde diferentes puntos de vista al problema, pero a la vez tantos puntos de vista a tener en cuenta, también son la causa que hace difícil el avance de las investigaciones. Entre las disciplinas más importantes que se ocupan de este problema, según Rabiner [1, 2] se pueden nombrar:

1. Procesamiento de la señal: donde se extrae información relevante, usando análisis espectral y variación de características en el tiempo.

2. Física (acústica) que trata de entender cómo se relaciona la señal física del habla con la fisiología mecánica del tracto vocal humano.

3. Reconocimiento de patrones: por medio de la utilización de algoritmos sobre datos para de esta forma crear prototipos de patrones semejantes y así realizar comparaciones.

4. Comunicación y teoría de la información: aporta los procedimientos para la estimación de parámetros en el uso de modelos estadísticos y de esta forma detectar la presencia de patrones particulares del habla, igualmente, establece una forma de codificación y decodificación moderna de algoritmos (programación dinámica, decodificación de Viterbi) usando las mejores correspondencias de secuencias de palabras reconocidas.

5. Lingüística: relaciona los sonidos (fonología) las palabras en un orden (sintaxis) el significado de las palabras (semántica) y el sentido y uso (pragmática).

6. Fisiología: entendiendo el alto nivel de los mecanismos que se relacionan con el sistema nervioso central con la producción de la voz en los humanos.

7. Ciencias de la computación: trata el estudio eficiente de la implementación de algoritmos en software o hardware para ser usados en prácticas de reconocimiento del habla.

8. Psicología: la ciencia que trata de entender los factores de cómo una tecnología pueda ser usada por humanos en tareas prácticas.

Dentro de las disciplinas anteriormente nombradas, es de principal interés para este trabajo, el desarrollo de sistemas de reconocimiento del habla que se desarrolla por medio de técnicas de reconocimiento de patrones e inteligencia artificial. Además, en este proceso se involucran técnicas de análisis de señal del habla, que son producidas por

medio del tracto vocal, de estas señales se extraen características que sirven para analizar desde un fonema hasta un discurso; en este sentido algunas de las características que se pueden ver son la frecuencia, la energía, los formantes, entre otras. Debido a que en este proyecto se analiza la producción del habla de la lengua Nasa Yuwe, proceso en el que se tiene que hacer referencia a las vocales y dentro de éstas a una clasificación, haciendo énfasis en las orales y nasales simples, sobre las cuales se han realizado mayores análisis por parte del *Grupo de estudios lingüísticos de la Universidad del Cauca*.

En el siguiente trabajo se presentan los resultados de la investigación y experimentación, la cual se realizó con el fin de dar respuesta a la siguiente pregunta: *¿Cuál es el modelo computacional que basado en una técnica de reconocimiento de patrones presente un menor nivel de error al reconocer un subconjunto de vocales de la lengua Nasa Yuwe, para desarrollar un prototipo software de reconocimiento del habla que reconozca las vocales seleccionadas y de esta manera contribuir a la revitalización de esta lengua?*

Para resolver este interrogante se plantearon los objetivos del proyecto teniendo en cuenta los conocimientos suministrados por el grupo de estudios lingüísticos en cuanto a la lengua Nasa Yuwe las tendencias actuales en el desarrollo de sistemas de reconocimiento de patrones que cubrieran de forma general el interrogante planteado.

Objetivo General

Adaptar un modelo computacional basado en una técnica de reconocimiento de patrones que menor nivel de error presente al reconocer un subconjunto de vocales de la lengua Nasa Yuwe y construir un prototipo software que soporte dicho modelo.

Objetivos Específicos

- Construir un corpus básico que contenga un subconjunto de fonemas característicos de las vocales seleccionadas de la lengua Nasa Yuwe.
- Adaptar un modelo computacional basado en una técnica de reconocimiento de patrones para reconocimiento del habla de un subconjunto de vocales de la lengua Nasa Yuwe, a partir del estudio de técnicas para reconocimiento del habla y seleccionar la que menor error presente.
- Desarrollar un prototipo software de reconocimiento de habla que soporte el modelo propuesto.
- Verificar el modelo adaptado utilizando el corpus construido mediante el prototipo software desarrollado.

Para esto se desarrolló el trabajo de investigación el cual se encuentra dividido en los siguientes capítulos:

Marco teórico, en este primer capítulo se realiza el estudio y comprensión del funcionamiento de un sistema de reconocimiento del habla, incluyendo el proceso de preénfasis que aplica técnicas de pre-procesamiento de señales de voz para obtener características como la frecuencia fundamental [3] que serán usadas en el proceso de reconocimiento de patrones, estadístico o neuronal. Como complemento se nombran unos de los proyectos de reconocimientos del habla relacionados y desarrollados en otras lenguas, en los cuales se usa un proceso de utilización de técnicas de reconocimiento de patrones.

En el segundo capítulo se hace referencia al proceso de recolección de las señales de audio, selección de las muestras y posterior construcción del corpus. En el tercer capítulo se hace referencia a las técnicas de reconocimiento elegidas y previamente estudiadas en el marco teórico, con las cuales se realiza la experimentación y comparación de la que menor error presente. Para proseguir con el cuarto capítulo que se trata del desarrollo de un prototipo con la técnica escogida, finalizando con una prueba piloto con integrantes del resguardo de López Adentro Caloto y sus resultados.

1. MARCO TEÓRICO

1.1. Producción y reconocimiento del habla, cómo se ve desde la perspectiva computacional

Una de las facultad es más impresionantes que se encuentra en los humanos es su capacidad de comunicar sus ideas por medio del habla. Esta es la capacidad que ha llevado al hombre al desarrollo de la sociedad. Pero esa capacidad innata en el hombre es un verdadero problema a enfrentar desde la perspectiva computacional. El ser humano desde la antigüedad se ha visto atraído por la construcción de máquinas, que sean capaces de producir y reconocer el habla. Entre las cuales podemos ver los ASR [1] (Automatic Speech Recognition).Estos son sistemas capaces de decodificar la señal producida por el tracto nasal y vocal de un hablante en una secuencia de unidades lingüísticas que contienen el mensaje a comunicar. El objetivo principal de un sistema de reconocimiento del habla es permitir la comunicación humano-máquina, posibilitando un acceso más rápido a información relevante. En el estudio de esta área se han realizado análisis de cómo se produce el habla por un hablante y su posterior reconocimiento por parte de un escucha. A continuación se ilustra una representación del proceso de producción del habla entre el hablante y el oyente y su contraparte en la máquina (Ver Figura 1).

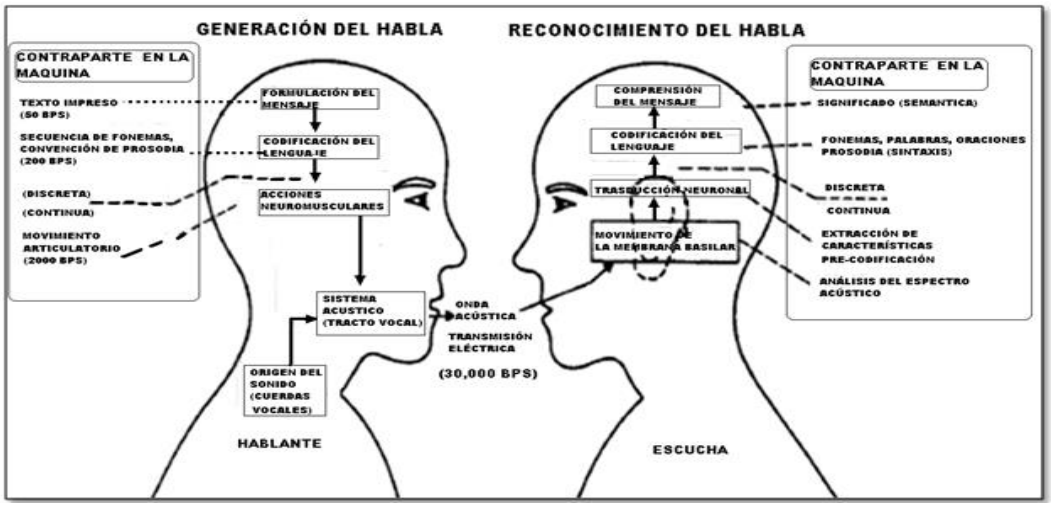


Figura 1. Esquema de producción y percepción del proceso del habla [1].

Como podemos observar el proceso comienza cuando el hablante formula en su mente el mensaje que quiere transmitir al escucha por medio del habla, en contraparte en la máquina se crea un mensaje tipo texto que expresa un mensaje en palabras. El próximo paso es convertir el mensaje en lenguaje de código, en la máquina el equivalente es la formación de secuencias de fonemas que corresponden a los sonidos de las palabras, posteriormente con la prosodia se denota la pronunciación de los sonidos y la duración, cuando el lenguaje es codificado el hablante realiza una serie de comandos neuromusculares que hacen vibrar a las cuerdas vocales para que vibren apropiadamente según los sonidos a ser creados. Las señales son propagadas hacia la membrana basilar

en el interior del oído del escucha, que le permite tomar la señal de llegada y convertirla en una señal que tiene características extraíbles por medio del nervio auditivo. Esto corresponde aproximadamente en la máquina al proceso de extracción de características, prosiguiendo se produce una actividad neuronal que finalmente se convierte en código de lenguaje que se procesa en el interior del cerebro para finalmente resultar con la comprensión del mensaje.

1.2. Funcionamiento de un sistema de reconocimiento del habla

El objetivo principal de sistema de reconocimiento del habla es la comunicación hombre–máquina [2]. Tradicionalmente las tres áreas de trabajo desde el punto de vista del proceso de la señal son la Codificación, Síntesis y Reconocimiento [4]. Para tener un mejor entendimiento de cómo funciona un sistema de estas características, se deben dividir sus procesos en módulos. A continuación se verá una arquitectura general donde se explica su funcionalidad.

1.2.1. Arquitectura de un sistema de reconocimiento del habla basado en patrones [5, 6, 7]

Los sistemas de reconocimiento del habla están compuestos por un conjunto de módulos que se pueden describir de forma sencilla desde de la producción del habla hasta su posterior reconocimiento como se observa en la Figura 2.

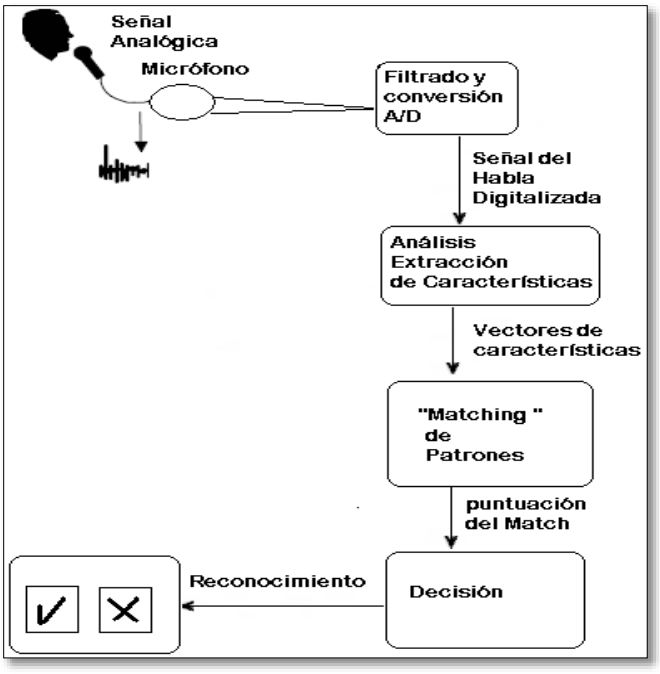


Figura 2. Esquema general de un sistema de reconocimiento del habla basada en patrones [1,5]

A continuación se describen los módulos correspondientes a un sistema de reconocimiento del habla basado en patrones de acuerdo a las fuentes [1, 2, 8, 5, 7]

- **Análisis y extracción de características:** Teniendo la señal de voz digitalizada, se extrae las características necesarias en términos de parámetros espectrales, la intención es hacer las características más evidentes. Se realiza un proceso de filtrado donde se limpia y reduce la dimensión de las características para facilitar su clasificación.

- **Correspondencia de Patrones:** la información espectral es evaluada por medio del patrón de reconocimiento utilizado, cubriendo todas las variedades fonéticas y palabras consideradas en el proceso de reconocimiento.

- **Decisión:** el reconocimiento de la palabra o frase es obtenido por medio de combinación de la información obtenida desde la etapa anterior, con conocimiento de la sintaxis y semántica del proceso de reconocimiento.

Además de los módulos del sistema de reconocimiento, es necesario contar con un conjunto de señales de audio, que representan el conjunto de datos llamado “corpus” [9] estas señales se utilizarán para realizar el análisis y posterior comparación como patrones de entrenamiento y prueba.

Los sistemas de reconocimiento del habla se pueden clasificar por las unidades lingüísticas que usen, tal vez muchos pensarían que la unidad lingüística más natural es la palabra, pero existen problemas, en el momento de manejar un gran conjunto de unidades lingüísticas o un amplio vocabulario [8]. Otra alternativa es dividir las unidades lingüísticas como subpalabras, que corresponden a los fonemas básicos del lenguaje. El problema con estos fonemas es no poder modelar el *efecto coarticulatorio* [10] *(se refiere a la influencia de un fonema sobre otro)* entre fonemas consecutivos. Con el fin de tratar este problema se han propuesto unidades tales como difonemas, sílabas, demisílabas o trifonemas [5].
Para medir la eficiencia de un sistema de reconocimiento de palabras aisladas se calcula una tasa de error (ER), que es el porcentaje de palabras erróneamente reconocidas. Los diferentes tipos de error que se pueden presentar son:

- **Sustitución:** una palabra de la frase original aparece sustituida por una palabra diferente en la frase.
- **Supresión:** una palabra de la frase original no se encuentra reconocida en la frase.
- **Inserción:** una nueva palabra es encontrada en dos palabras de la frase original.

La medida de eficiencia más comúnmente usada está dada por la diferencia de error, entre el porcentaje correcto (PC), la tasa de palabras de error (WER) y la precisión de reconocimiento (WAcc) de las palabras dadas, que se definen como [5]:

$$PC = 100 \times \frac{C - I}{N} = \frac{N - (D + S)}{N}$$

Fórmula 1.

$$WER = 100 \times \frac{S + D + I}{N}$$

Fórmula 2.

$$WAcc = 100 - WER = 100 \times \frac{N - (S + D + I)}{N}$$

Fórmula 3.

Donde *C* es el número de palabras correctamente reconocidas, *S* es el número total de sustituciones, *D* es el número total de supresiones, *I* el número total de inserciones y *N* el número total de palabras evaluadas. La evaluación de *C, S, D I*, se hace usualmente alineando la frase reconocida con la frase original; este procedimiento se puede hacer por medio de técnicas de programación dinámica [11]

1.3. Pre procesamiento de la señal del habla

La importancia principal del pre procesamiento de señales del habla es interpretar como ellas son designadas y como ellas funcionan. Existen diferentes formas de parametrizar una señal. Entre ellas se pueden nombrar, las técnicas de codificación por predicción lineal y coeficientes Cepstrales de Mel [1], [5], [11], [12].La primera operación que suele llevarse a cabo es el pre-énfasis que consiste en filtrar la señal con un filtro digital, posteriormente las señales son divididas en pequeños segmentos denominados marcos (frames) y posiblemente superpuestos. Dentro de un marco la señal se considera cuasi estacionaria de tal manera que los parámetros que caracterizan a la señal (características) pueden considerarse constantes dentro de ese marco.

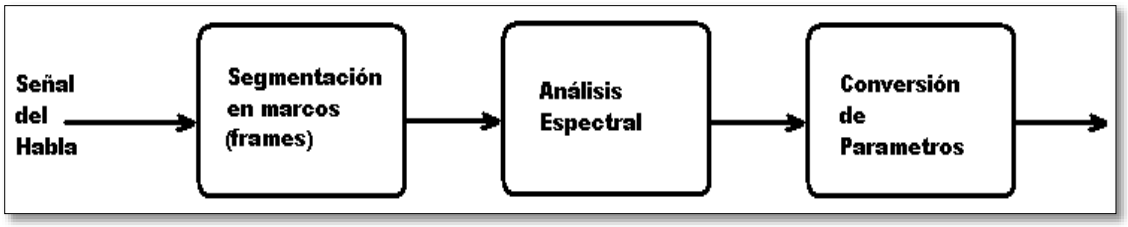


Figura 3. Esquema general de la etapa de análisis [1]

La secuencia de marcos (frames) o fotogramas se obtiene al desplazar una ventana de análisis a través de la señal original. Una ventana muy común en los sistemas de ASR es la *ventana de Hamming* [5] esta función matemática permite dividir señales evitando la discontinuidad de bloques analizados, permitiendo de cada bloque se extraer un conjunto de características, que se consideran constantes en ese momento.

El objetivo del módulo de análisis es que cada marco sea representado por un *vector de características* X que contenga el análisis de los parámetros. El conjunto de características por lo general es muy grande, éstas pueden ser información sobre la energía y características dinámicas (como la frecuencia fundamental, formantes, pitch) [1], [2], [5]. En conclusión el resultado final de este módulo de análisis es un vector de características tales como $X = (x_1, x_2, ..., x_T)$ que representan a la frecuencia fundamental, formantes, pitch entre otras [5]:

1.3.1. Modelos de Análisis Espectrales

Para realizar el análisis de señales del habla, se debe tener en cuenta como se puede realizar la observación de los parámetros de medida utilizados en la fase de pre-énfasis, pues como se ha visto anteriormente en la arquitectura de un ASR, este es el primer módulo del modelo del reconocedor donde se observarán marcadas diferencias en los parámetros que se pueden medir en una señal del habla. Existen diferentes técnicas para realizar el procesamiento de las señales, pero los métodos de análisis espectrales más conocidos y actualmente usados por sus buenos resultados [6] son la transformada de Fourier en su forma discreta y mejorada, Análisis Cepstral y Codificación por Predicción Lineal. Siendo estas dos últimas las que se utilizaron con técnicas de patrones en este proyecto. La razón para utilizar estas técnicas es que permiten parametrizar una señal con un número pequeño de patrones, esto es ideal cuando se cuenta con un conjunto de señales reducido, otra característica es que la técnica LPC tiene un coste computacional bajo [1] y por último se pueden observar resultados con un alto porcentaje de reconocimiento para los dos técnicas [13].

1.3.1.1. Codificación por predicción lineal (LPC)

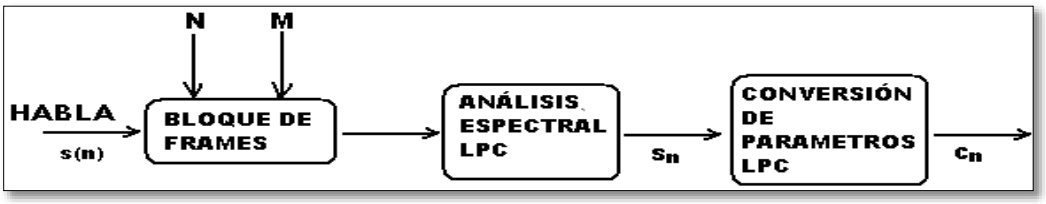


Figura 4. Esquema de bloques del proceso de predicción lineal [1]

Este método sirve para estimar resonancias en variación del tiempo en el tracto vocal. Esto permite obtener la información sobre cómo se representan los sonidos del habla de forma aproximada, por medio de la estimación de los formantes, que son los picos de intensidad de los sonidos, en el caso de las vocales se pueden encontrar más de un formante en diferentes partes del espectro como se ve en la Figura 5.

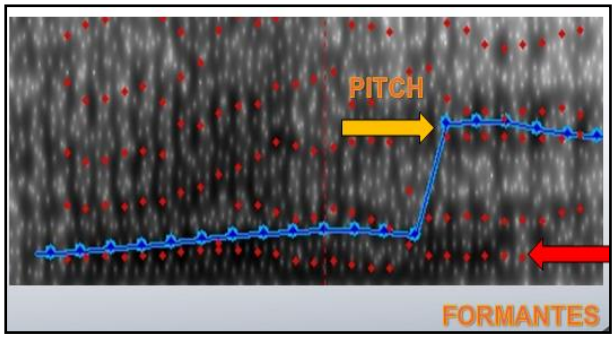


Figura 5. Características de un espectrograma de onda observadas en Praat [14]

En la figura anterior se observa de color rojo los formantes (F1, F2, F3) siendo el primero (F1) la frecuencia fundamental [15] y de color azul el alto de tono (pitch). Las frecuencias de los formantes son las resonancias del tracto vocal, que poseen las diferencias entre los sonidos, estas son las características más útiles para realizar el reconocimiento del habla. El pitch es esencialmente importante en el reconocimiento del tono siendo

importante para eliminar ambigüedad entre homófonos. Para ver más información relacionada vea [7].

En el análisis de la representación espectral está dada por la Fórmula 4 [1].

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + + a_p z^{-p}$$ **Fórmula 4**

En la Formula 4, se observa un polinomio de orden *p-ésimo* con transformada *(z).* El orden *p,* es llamado el orden de análisis LPC. La salida del análisis espectral de LPC es un vector de coeficientes (parámetros LPC) que especifica las mejores similitudes de señales del espectro sobre un periodo de tiempo, en el que los fotogramas de las muestras de habla fueron acumulados.

La codificación del modelo de predicción lineal se puede apreciar en la Figura 6:

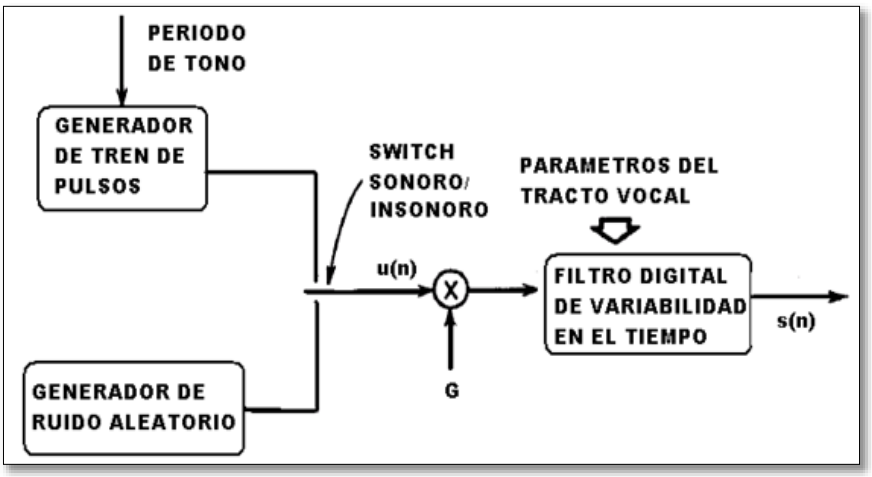


Figura 6. Síntesis del habla basada en el modelo LPC [1]

En la Figura 6, se puede apreciar que para producir la señal de voz *s(n)* se tiene en cuenta también dos fuentes una formada por un tren de pulsos para sonidos sonoros y otra compuesta por un generador de ruido para sonidos sordos, estas fuentes están controladas por un interruptor que escoge la fuente de la señal del habla.

Finalmente de la técnica de LPC podemos obtener un vector de características representadas por un conjunto de coeficientes que representan las resonancias del aire en el tracto vocal como se observa en la Figura 7.

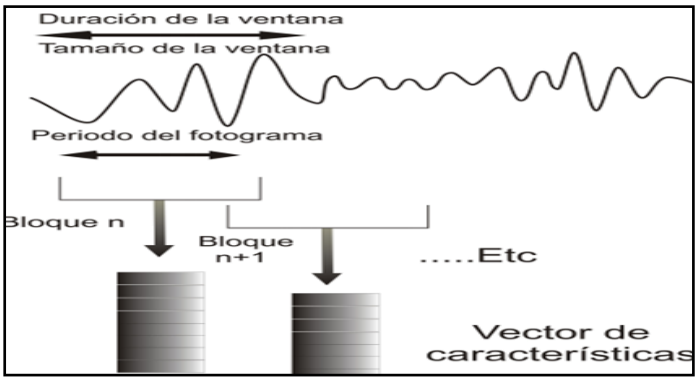


Figura 7. Extracción de características [16]

1.3.1.2. Análisis de Coeficientes Cepstrales de Mel [17]

Los coeficientes espectrales son usualmente transformados para obtener una representación apropiada de reconocimiento, esta técnica realiza una derivación conocida como *cepstrum* [7]. El cepstrum es la función temporal obtenida como la transformada inversa del logaritmo del espectro. Las muestras de cepstrum son usualmente conocidas como *coeficientes cepstrales*. Para una tarea de reconocimiento del habla solo son tomados los primeros coeficientes cepstrales. El proceso de extracción de características se puede ver descrito a continuación en la Figura 8.

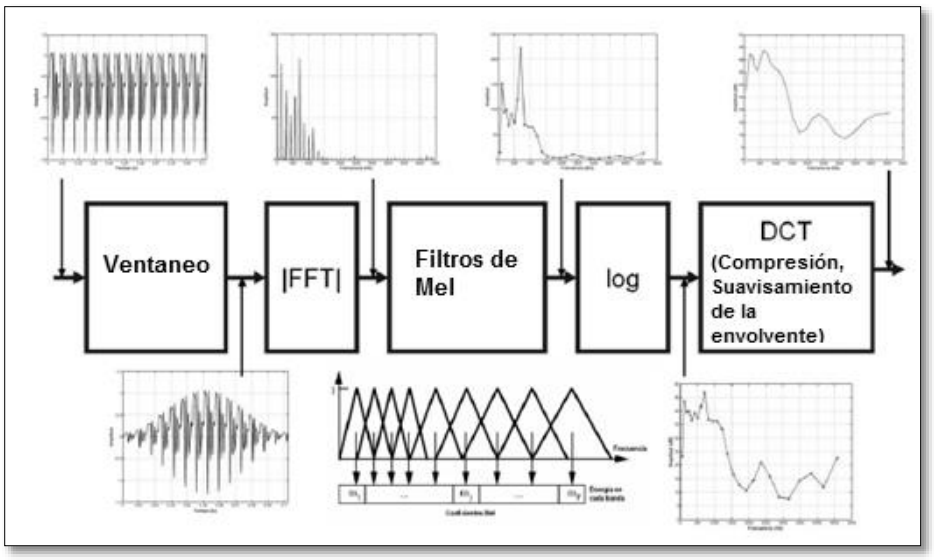


Figura 8. Esquema del análisis Cepstral [17]

En la Figura 8, se pudo observar el que el Cepstrum [7] es resultado de calcular la Transformada rápida de Fourier FFT del espectro de la señal que ha sido dividida en un conjunto de fotogramas por medio de una ventana triangular, seguidamente se aplica el logaritmo a su espectro de potencia y por último se lleva al dominio del tiempo en un proceso inverso a través de la Transformada Discreta de Fourier Inversa (IDTF) [8].

1.4. Aproximaciones en reconocimiento del habla por medio de máquinas

Las diferentes técnicas propuestas para resolver el problema del reconocimiento del habla están agrupadas en tres categorías principales:

- Enfoque acústico fonético.
- Enfoque de patrones.
- Enfoque de inteligencia artificial.

1.4.1. Enfoque acústico fonético [1, 12]

En este enfoque se tiene en cuenta las características de la señal vistas como unidades fonéticas que contienen el lenguaje hablado, entre estas unidades fonéticas tenemos a los fonemas que forman palabras, que a la vez forman las frases que sirven para comunicar ideas. Este enfoque puede ser modelado a través de sistemas expertos [1, 48] formalizando el lenguaje por medio de reglas. Un sistema Acústico Fonético se divide en las siguientes partes. (Ver Figura 9)

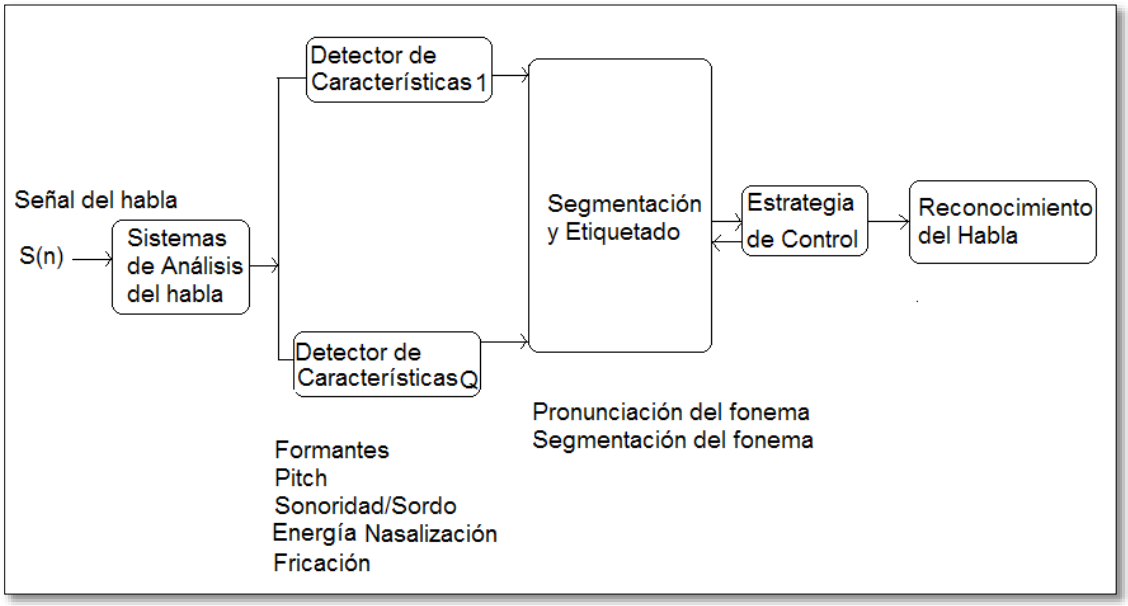


Figura 9. Diagrama Acústico - Fonético de un sistema de reconocimiento del habla [1]

- **Sistemas de Análisis del habla:** el proceso comienza cuando se obtiene la señal de voz por medio de un micrófono, a esta señal se le aplica un procesado inicial, que permite transformar la señal de su *dominio temporal* a un dominio de *frecuencia* por medio de la *transformada de Fourier*. Los resultados obtenidos se denominan *características acústicas,* dentro de estas características, las más representativas son *los formantes*, *el pitch* (alto de tono), *fricación* (observaciones de obstrucción del aire en el tracto vocal) entre otras.
- **Detección de Características Fonéticas:** ya realizado el análisis acústico se prosigue a realizar la extracción de características fonéticas, éstas contienen

unidades fundamentales para diferenciar entre unidades. Las características más usuales que se encuentran son: *la frecuencia fundamental, los formantes* y el *grado de sonoridad (pitch)* [7].

- **Segmentación y etiquetado:** la señal de voz es dividida en zonas con características similares para ser asignadas a una o varias características fonéticas (formantes, pitch, entre otras).
- **Estrategia de control y Reconocimiento del habla:** se realiza a partir de la decodificación de las palabras pronunciadas teniendo en cuenta las características obtenidas en las fases previas, se usan reglas semánticas y sintácticas que se obtienen mediante la señal de voz estudiada.

1.4.2. Enfoque de patrones [1, 5]

La característica más importante de este enfoque es un marco matemático bien definido [16], que se basa en establecer representaciones de patrones del habla, por medio de un conjunto de modelos captados automáticamente en fases de entrenamiento, los modelos de entrenamiento captados pueden usarse para realizar comparaciones directas con una muestra de voz desconocida (muestra a reconocer). La representación de los patrones puede ser una plantilla (*template*) o un modelo estadístico (*HMM, Modelos ocultos de Markov*) que pueden ser aplicados a sonidos pequeños (fonemas), palabras o frases. Puede resumirse en dos fases como se muestra a continuación.

La ventaja principal del reconocimiento de patrones con otro enfoque es poder seleccionar las características más adecuadas para describir los objetos, hacer una buena selección de características relevantes permitirá mejorar la clasificación y aumentar la velocidad de procesamiento [1, 7]. Otras características son [5, 19]:

- El aprendizaje, ya que a la técnica se le introduce un conjunto de datos que a la vez indican la respuesta esperada.
- Auto organización, que tiene su propia estructura interna.
- Tolerancia a fallos, puede seguir funcionando de forma eficiente aunque falten o fallen algunos datos.
- Flexibilidad, al poder manejar datos con cambios no importantes en los datos de entrada, por ejemplo señales con ruido.
- Tiempo real, puesto que se pueden obtener respuestas en tiempo real.

Un sistema de reconocimiento basado en patrones puede resumirse en dos fases: *Una fase de entrenamiento*, en donde se generan los modelos de referencia y una *fase de reconocimiento*, en donde se hacen comparaciones entre las pronunciaciones producidas por un hablante y las que se tienen como referencia; eligiendo la secuencia que mejor se aproxime a los modelos o plantillas de referencia.

A continuación se muestra el diagrama de bloques que describe el proceso fundamental de reconocimiento del habla (ver Figura 10).

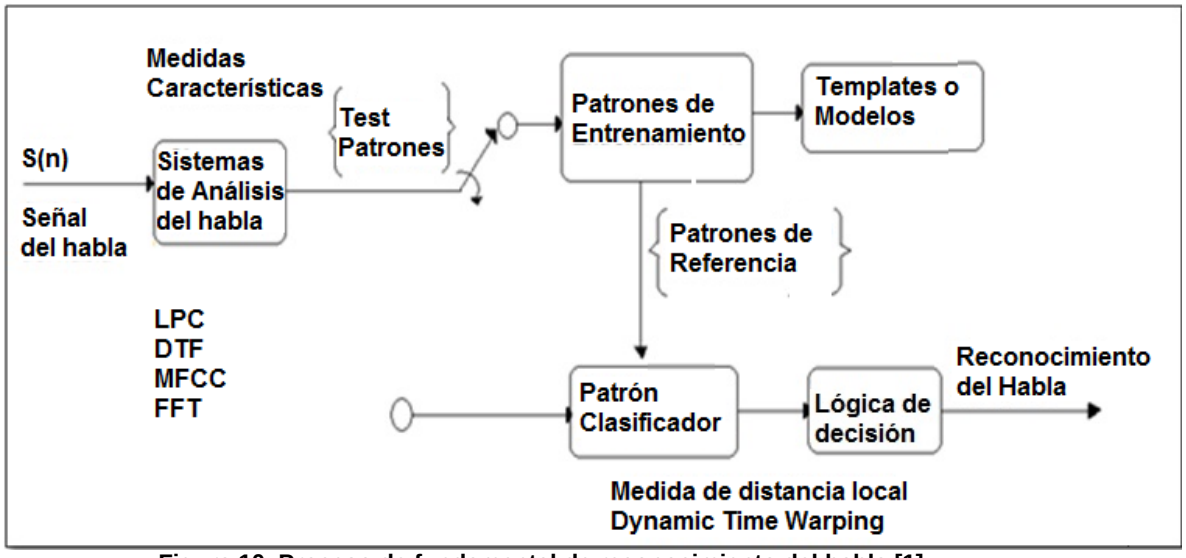


Figura 10. Proceso de fundamental de reconocimiento del habla [1]

El proceso fundamental de reconocimiento del habla se describe en los siguientes pasos:

- **Sistema de análisis de Habla [1]** donde una secuencia de medidas es tomada sobre la señal de voz, para ser definida como el "patrón de voz" [8], de esta señal se extraen algunas características que son obtenidas mediante alguna técnica de análisis espectral, como *Coeficientes Cepstrales de Mel [7], predicción lineal* [11] o *transformada rápida de Fourier* [20].

- **Patrones de Entrenamiento[1, 5, 6]** donde uno o más patrones de prueba correspondientes a un sonido del habla de una misma clase son usados para crear un patrón representativo de la clase, este resultado es usualmente llamado *patrón de referencia* y puede determinarse como *ejemplar o template,* derivado de alguna técnica de promedio o modelo estadístico.

- **Patrón Clasificador [5]** en el que un patrón de prueba no reconocido es comparado con el de referencia, y se calcula la correspondencia (distancia) entre el patrón de prueba y cada uno de los patrones de referencia computados, al comparar estos patrones del habla (que se resume en un conjunto de *vectores espectrales)* se mide la *distancia local* definida como distancia "*espectral*" entre dos vectores espectrales bien definidos y un procedimiento de tiempo global (DTW).

- **Lógica de Decisión [16]** en que los patrones de referencia con mayor puntaje son usados para decidir el mejor "*match*" de los patrones no reconocidos de prueba.

La representación por patrones se puede ver por medio de plantilla (Template) o un modelo estadístico como HMM siendo estos aplicables a sonidos pequeños como fonemas, hasta palabras y frases.

1.4.2.1. Alineamiento Dinámico en el Tiempo DTW (Dynamic Time Warping) [1, 5, 12]

Dentro de este enfoque de reconocimiento de plantillas podemos observar la técnica DTW que se basa en la aproximación de las medidas de distancia entre una entrada de prueba y una secuencia de referencia, el problema es: ¿cómo medir distancias entre secuencias de diferente longitud? Como caso común en el reconocimiento del habla, para resolver este problema es necesario alinear los objetos de entrada con los objetos de referencia, este alineamiento se puede lograr por medio de esta técnica que aplica *Programación Dinámica*(citado en [5] tomado de [21]), el proceso es el siguiente:

Teniendo una *matriz d(i, j) (i = 1, . . . , I; j = 1, . . . , J),* con todas las posibles distancias entre los vectores de entrada de objetos con *I fotogramas y los objetos de referencia* que representan los patrones ya construidos con *J fotogramas* se lleva a cabo el alineamiento para encontrar la ruta óptima en *la matriz d(i, j)* (ver Figura 11).

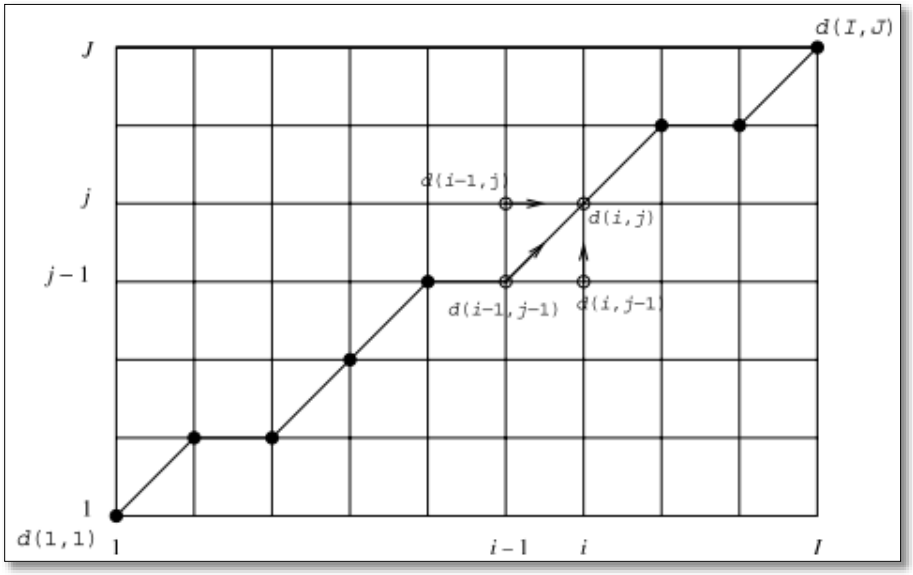


Figura 11. Técnica de Programación Dinámica con el algoritmo DTW [8]

La ruta óptima es entonces el resultado de calcular distancia más significativa (el menor coste) de las distancias computadas así:

$$g(i,j)=\min\begin{cases} g(i-1,j)+d(i,j) & \text{Inserción de fotograma} \\ g(i-1,j-1)+2d(i,j) & \text{Progresión normal} \\ g(i,j-1)+d(i,j) & \text{Eliminación de fotograma} \end{cases}$$

Fórmula 5

Entonces la distancia entre ambos objetos es:

$$D=\frac{g(I,J)}{(I+J)}$$

Fórmula 6

El valor de la distancia resultante puede ser utilizado en la clasificación de un objeto de entrada, para comparar con cada uno de los objetos de referencia que representan las posibles pronunciaciones y seleccionar la que provee la distancia mínima. El proceso no sólo provee la distancia entre dos objetos, sino que, también *la ruta más óptima* (correspondencia entre *fotogramas* particulares) [5].

1.4.2.2. Modelos estadísticos basados en Modelos Ocultos de Markov [5]

Las técnicas de *Modelización* son herramientas usadas para resolver tareas de predicción, reconocimiento o identificación. Su uso es aplicable a modelos de señales porque limpia la señal de ruido, además el modelo puede ayudar al entendimiento de la fuente de la señal y su proceso de generación.
Suponiendo*W = {Wi}* como el conjunto posible de frases de un lenguaje dado y que se desea obtener la frase *W(X)* correspondiente a una evidencia acústica *X,* aplicándole la regla de decisión del Máximo A Posterior (MAP Maximum A Posteriori) [1, 8] el reconocimiento de la frase obtenida es vista como:

$$W(x) = \arg_j \max P(W_j \mid X)$$ **Fórmula 7**

Esta maximización requiere del cálculo de probabilidades $P(W \mid X)$la clásica aproximación es la descomposición por medio de la regla de Bayes.

$$P(W \mid X) = \frac{P(X \mid W)P(W)}{P(X)}$$ **Fórmula 8**

La probabilidad condicional $P(X \mid W)$ es obtenida por el *modelo acústico* y $P(W)$ por el modelo del lenguaje.

En el área de reconocimiento del habla el modelo estadístico por excelencia ha sido los Modelos Ocultos de Markov (HMM, Hidden Model Markov) a continuación se presenta el esquema general de esta técnica y los procesos para llevar al reconocimiento de la señal del habla enviada como entrada (ver Figura 12).

Definición de Modelos Ocultos De Markov

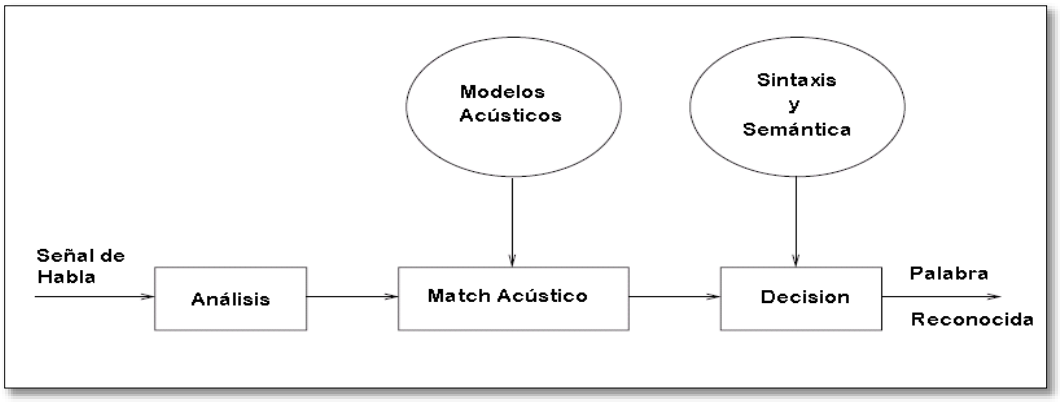


Figura 12. Esquema General de un sistema de reconocimiento del habla basado en Modelos Ocultos de Markov [5]

Un modelo oculto de Markov es obtenido como un proceso de Markov (ver Figura 13), descrito por un conjunto de N estados *{s_1, s_2,…,s_N}.* Cada estado es un cierto evento u observación. El sistema cambia de un estado a otro *(transición)* en cada intervalo de tiempo, llamado el estado q_t en el tiempo t.

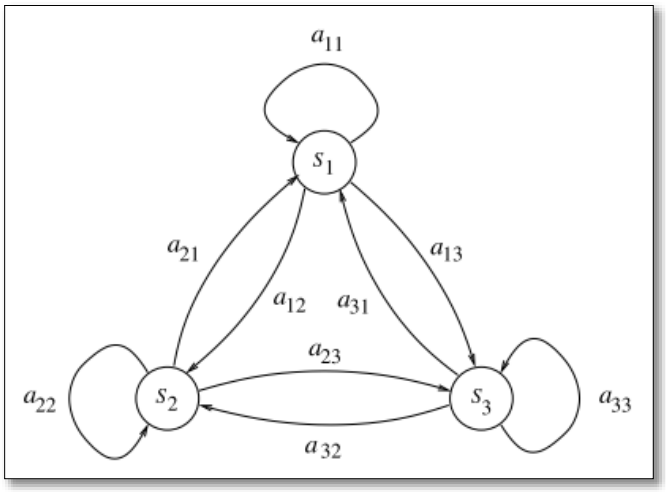


Figura 13. Proceso Discreto de Markov [5]

Los procesos de Markov se caracterizan por la dependencia del estado actual con respecto al estado anterior, en otras palabras el proceso tiene *"memoria"*, el estado actual solo depende del estado previo, independientemente de las consideraciones del tiempo, este proceso es totalmente descrito por las *probabilidades de transición,* de un estado a otro.

Un Modelo Oculto de Markov Discreto se caracteriza por los siguientes elementos:

Una notación habitual es la representación como una tupla *(Q,V,π,A,B):*

- El conjunto de estados $Q = \{1,2,\ldots, N\}$. El estado inicial se denota como q_t. Cada valor de t hace referencia a la posición de la palabra en la frase.
- El conjunto V de posibles valores $\{v_1, v_2,\ldots,v_M\}$ observables en cada estado. M es el número de palabras posibles y cada v_k hace referencia a una palabra diferente.
- Las probabilidades iniciales $\pi = \{\pi_i\}$, donde π_i es la probabilidad de que el primer estado sea el estado Q_i
- El conjunto de probabilidades A = $\{a_{ij}\}$ de transiciones entre estados.
    - $a_{ij} = P(q_t = j \mid q_t - 1 = i)$, es decir, a_{ij} es la probabilidad de estar en el estado j en el instante t si en el instante anterior t − 1 se estaba en el estado i.
- El conjunto de probabilidades B = $\{b_j(v_k)\}$ de las observaciones.
    - $b_j(v_k) = P(o_t = v_k \mid q_t = j)$, es decir, la probabilidad de observar v_k cuando se está en el estado j en el instante t.
- La secuencia de observables se denota como un conjunto $O = (o_1, o_2,\ldots,o_T)$.

Es muy común encontrar un grupo de diferentes problemas asociados al uso de HMM (citado en [5] tomado de [1]). Estos son:

- *Problema de la evaluación:* dada una secuencia de observación O, encontrar la probabilidad $P(O \mid \lambda)$ de la secuencia *O* generada por λ modelo.
- *Determinación de ruta óptima problema:* Dada una secuencia de O y λ modelo, encontrar la trayectoria óptima P.
- *Problema de estimación:* dada una secuencia de observación O, encontrar el conjunto de parámetros λ que mejora la secuencia de estimación.

1.4.3. Enfoque de la inteligencia artificial [22]

El cerebro humano consta de una gran cantidad de neuronas que se comunican entre sí mediante conexiones sinápticas, estas conexiones pueden *excitarse* o *inhibirse;* cada neurona recibe entradas a través de las conexiones y emite una salida. Igualmente las redes neuronales artificiales son circuitos, algoritmos de computadora, o representaciones matemáticas de un conjunto de neuronas conectadas masivamente que forman una red, simulando las neuronas biológicas. Éstas son útiles en el reconocimiento de patrones, procesamiento de señales, estimación y control de problemas. El objetivo de las redes neuronales artificiales es conseguir respuestas similares a las del cerebro humano.

En este enfoque se intenta automatizar el procedimiento de reconocimiento de acuerdo a la forma en que una persona aplica su inteligencia en la visualización, análisis y caracterización de la voz basada en un conjunto de características acústicas. Algunas técnicas que se emplean son los sistemas expertos, donde se organizan estructuras de forma jerárquica (ver Figura 14), dividiendo el trabajo en varios bloques de proceso concatenado. Cada uno de los bloques tiene como entrada la salida del anterior bloque, de esta manera el procesador acústico fonético analiza la forma de onda y produce varias secuencias de fonemas correspondientes a un grado de probabilidad determinado a la transcripción fonética de la señal de entrada del sistema. Los módulos de procesamiento semántico, sintáctico y morfológico se encargan de limpiar y recortar dejando solo las secuencias de palabras gramaticalmente correctas (citado en [18] tomado de [23])

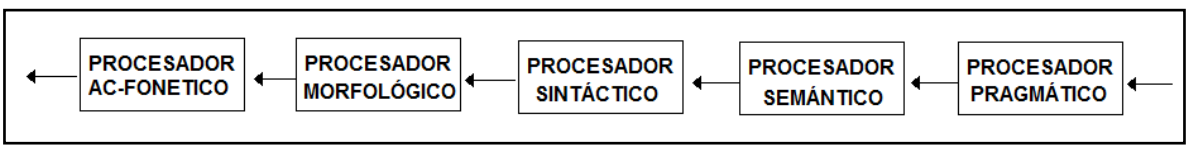


Figura 14. Módulos de un sistema experto, organización jerárquica [18]

Dentro de los enfoques de inteligencia artificial se presenta el aprendizaje supervisado y no supervisado. En la figura 15 se puede observar el tipo de aprendizaje supervisado.

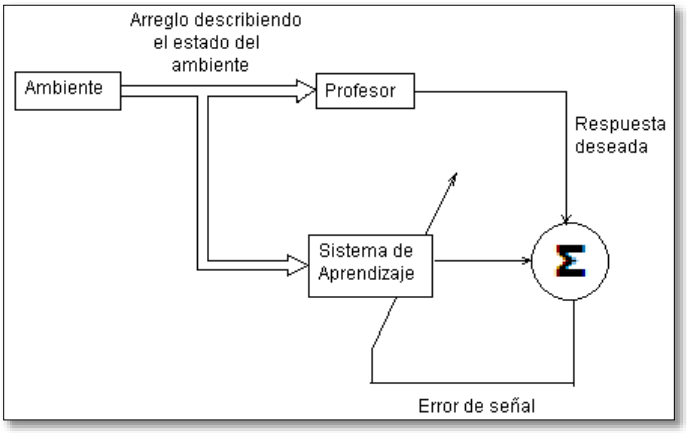


Figura 15. Diagrama en bloques del aprendizaje supervisado [22]

El aprendizaje supervisado tiene como ingrediente esencial la disponibilidad de un experto externo [22]. En términos conceptuales, se puede pensar en el experto como el que tiene el conocimiento del contexto que es representado por un conjunto de ejemplos de entrada y salida. El contexto es sin embargo, desconocido a las redes neuronales de interés. En este tipo de aprendizaje, el conocimiento del experto es útil a la red neuronal ya que este conoce la respuesta objetivo o deseada para el vector de entrenamiento, la respuesta deseada representa la acción a ser desempeñada por la red neuronal. Luego, los parámetros de la red son ajustados bajo la influencia combinada del vector de entrenamiento y la señal de error; la señal de error es definida como la diferencia entre la respuesta actual de la red y la respuesta deseada. Este ajuste es llevado iterativamente paso a paso con el objetivo de hacer que la red neuronal emule al experto.

1.4.3.1. Redes neuronales artificiales [22]

- **Redes de función de base radial [22, 24]**

A diferencia de la disposición que se tiene en la funciones de activación que permite construir modelos de entrenamiento mediante *backpropagation [22]*, estas nuevas redes basadas en RBF (por sus siglas en inglés de Radial Basis Function) construyen sus modelos con funciones de activación que son diferente tanto en la capa oculta como la de salida. Esto es, una red RBF está diseñada con neuronas en la capa oculta, activadas mediante funciones radiales de carácter no lineal con sus centros gravitacionales propios y en la capa de salida mediante funciones lineales. Estas redes están construidas por una arquitectura rígida de tres capas: la de entrada, la oculta y la de salida [22].

El funcionamiento de estas redes neuronales es el siguiente:

- La capa de entrada que sirve para los ejemplos o patrones de entrenamiento y prueba.

- La capa oculta completamente interconectada entre todos sus nodos con la capa de entrada y activada a través de la función radial (Gaussiana) y

- La capa de salida, también completamente interconectada a la capa oculta y activada a través de una función lineal continúa.

El entrenamiento de una red de base radial, a diferencia de una red que usa backpropagation, es solamente hacia delante. De este modo, la salida *z* de una red RBF, en general, está influenciada por una transformación no lineal originada en la capa oculta a través de la función radial y una lineal en la capa de salida a través de la función lineal continua (ver Figura 16).

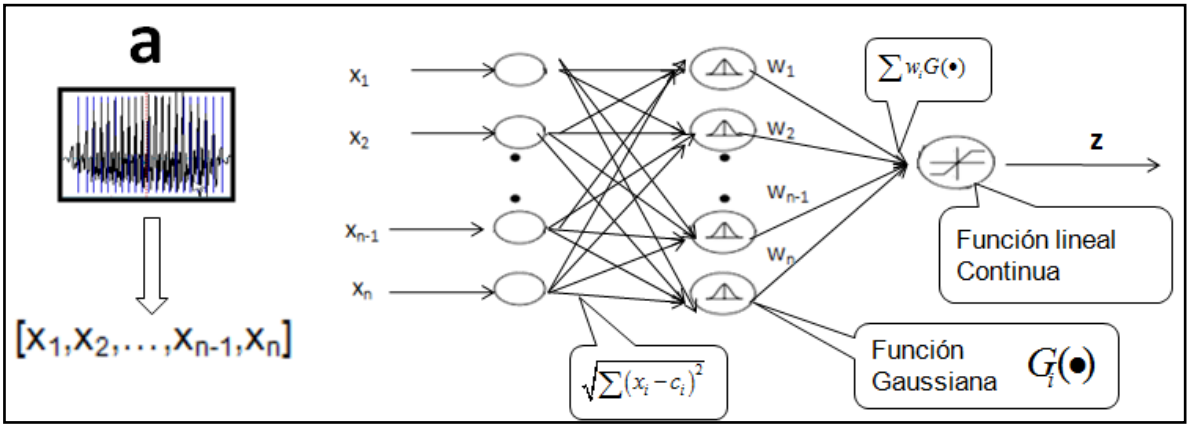


Figura 16. Modelo de una red neuronal de base radial [24]

De acuerdo a su topología tenemos que:

- Los nodos ocultos contienen una función base radial, la cual tiene como parámetros a centro y ancho.
- Existe un *centro* para cada función radial involucrada en la capa oculta. Regularmente, definen un vector de la misma dimensión del vector de entrada y hay normalmente un centro diferente por cada nodo de la capa oculta.
- Por otro lado, el *ancho* es el término empleado para identificar a la amplitud de la campana de Gauss originada por la función radial. Es decir, la desviación estándar de la función radial.

El cálculo efectuado en la capa oculta es hallar en un nodo de la capa oculta la distancia radial (distancia euclidiana) *d* entre el vector de entrada *x*, con n observaciones, a ese nodo en particular y el centro de gravedad *c* de ese mismo nodo. Es decir:

$$d = \|x - c\| = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \ldots + (x_n - c_n)^2}$$ **Fórmula 9**

Este valor *d* es un componente de la entrada para activar la función radial *G (•)*. En cuanto a la función radial *G (•)*, el contenido evaluado en cada nodo es la distancia euclidiana *d* y se estaría trabajando con $\exp(d^2 / a)$ donde *a* es el ancho para ese nodo oculto. Entre la capa oculta y la capa de salida se derivan un conjunto de pesos *w* que se verían afectados de acuerdo al algoritmo de aprendizaje. En este caso

particular sería la combinación lineal entre los pesos y la resultante de cada función radial para determinar la salida *z.* Por lo tanto:

$$z = \sum w_i G(\bullet)$$ **Fórmula 10**

Donde *G (•)* es la salida de la capa oculta y se corresponde con la función radial aplicada a la distancia euclidiana en cada una de las unidades ocultas.

1.5. Estado del arte

La investigación documental sobre el objeto de estudio ha permitido aprender los referentes básicos de sistemas de reconocimiento del habla y orientar este proyecto, como se verá a continuación.
El desarrollo computacional y la posibilidad de capturar información acústica permitieron el desarrollo de nuevas técnicas para el tratamiento de señales. En los 60 se trabajó con unidades lingüísticas pequeñas como fonemas. En los 70 se desarrollaron reconocedores con un vocabulario mayor con mejores técnicas de reconocimiento [7] [25] el proyecto más ambicioso fue el iniciado por DARPA con un sistema de entendimiento del habla. Aunque actualmente se ha avanzado mucho en la utilización de aplicaciones de reconocimiento del habla con interfaces orales pero en conclusión los proyectos desarrollados en esta área de la inteligencia artificial muestran que no existen ningún sistema de reconocimiento y/o entendimiento del habla autónomo y universal para todos los idiomas.

Actualmente los sistemas de reconocimiento automático del habla se han utilizado en diferentes aplicaciones, dentro de las que se destacan: Las de *Dictado* que incluye transcripciones médicas, jurídicas, escolares, en algunos casos los vocabularios son usados para aumentar la exactitud del sistema; las de *Comando y control,* estos sistemas desempeñan funciones a través de expresiones tales como “Abrir terminal” ejecutando el comando preciso; *aplicaciones embebidas* en los teléfonos celulares con la pronunciación “Call home”.
Como se ve el tipo de aplicaciones que se le puede dar a los sistemas de reconocimiento del habla son muy variadas, es por eso, que la revisión sobre antecedentes en esta área se realizó con proyectos realizados dentro de la temática que involucran la utilización de técnicas de reconocimiento de patrones para el habla. Dentro de estos antecedentes se partió de la siguiente clasificación:

1.5.1. Sistemas de reconocimiento del habla para palabras aisladas [26]

En estos sistemas usualmente se requiere que por cada palabra pronunciada exista un silencio, estos sistemas tienen estados de “escucha” y “no escucha”, esto permite localizar las fronteras en cada palabra (para ver otros tipos de sistemas de reconocimiento del habla, ir al Anexo A), los ejemplos de este tipo de sistemas son los siguientes:

a) *Reconocimiento automático de fonemas usando Redes Neuronales* [27]

En este trabajo se hace un recuento de los conceptos básicos de cómo funciona el sistema auditivo, se realiza un estudio básico sobre fonemas y reconocimiento del habla. Para la aplicación de esta temática se hace uso de redes neuronales con aprendizaje no supervisado para el desarrollo de un software básico, utilizando herramientas software de MATLAB. Este proyecto se usa para reconocer el fonema de la letra d y t.

b) Aplicaciones en reconocimiento del habla utilizando HTK [28]

Tesis en la cual se realiza una investigación sobre dos enfoques de reconocimiento: *palabras aisladas y reconocimiento de dígitos conectados*. En el primer enfoque las palabras se pronunciaron entre silencios teniendo en cuenta que las unidades lingüísticas son las palabras de un vocabulario específico para interactuar con el sistema entre las cuales están: prender, apagar, cancelar, sí y no. En el segundo enfoque se trata de reconocer una cadena de 6 dígitos del 0 al 9. Estos sistemas son dependientes del locutor haciendo más fácil la tarea de reconocimiento. Se hizo una interfaz que guía al usuario paso a paso en las diferentes fases de la elaboración de los reconocedores adaptando así la herramienta HTK Toolkit [29] que implementa la técnica de modelos ocultos de Markov. Como conclusión final brinda una adaptación de la metodología para construir ASR.

c) Reconocimiento automático del habla: Interface humano computador para la lengua Kinyarwanda [30]

En este proyecto se muestra como se realizó un sistema de reconocimiento del habla para la lengua Kinyarwanda, los productos del estudio incluyen un corpus de ésta, para utilizarlo en un sistema telefónico que reconoce los dígitos pronunciados (0-9) que se conoce como palabras aisladas, utiliza una herramienta software llamada HTK, donde se puede utilizar la técnica de modelos ocultos de Markov, la razón que se da para el desarrollo de este trabajo es que los Ruandeses no hablan inglés ni francés, lenguas en las que están construidas las herramientas TIC que llegan a Ruanda. El propósito del proyecto fue diseñar y entrenar un sistema de reconocimiento de habla que pudiera ser usado por desarrolladores para implementar una aplicación que tome hablantes de la lengua indígena Kinyarwanda para abordar las actuales tecnologías de información y comunicación.

d) Proyecto de tecnologías del habla 2004 Construcción de un reconocedor del habla para el holandés basado en Modelos Ocultos de Markov [31]

Este proyecto trata sobre la construcción de un sistema de reconocimiento para el idioma holandés usando los modelos ocultos de Markov, el objetivo de este proyecto fue la construcción de un robusto reconocedor de fonemas que sirviera como pilar para continuar desarrollando aplicaciones para esta lengua, y no sólo un ejemplo más basado en el aprendizaje. Para ello el proyecto se restringió a dominios sugeridos por el tutorial del HTK Toolkit, es decir, a acciones que puede ejecutar un teléfono, como por ejemplo: “Marcar 1 2 0 5” o “Marcar *nombre_contacto*”. Como conclusión se destaca que el sistema con el corpus IFA (Instituto de Ciencias Fonéticas de Amsterdam) no afecta la tasa de reconocimiento por la forma como se dividieron las grabaciones, el sistema fue capaz de reconocer hablantes con los que nunca entrenó.

e) Algoritmos y métodos para el reconocimiento de voz en español mediante sílabas [32]

Este trabajo afronta el problema de reconocimiento de fronteras entre fonemas, ya que es difícil identificar las representaciones acústicas de voz. En el trabajo se analiza la sílaba y su alta sensibilidad al contexto; también es importante resaltar que se propone una alternativa a la forma en que los ASR han sido implementados hasta el momento, para ello toman factores como la función de energía total en corto tiempo y el Sistema Basado en Conocimiento, el cual es capaz de realizar la clasificación de la señal de entrada en unidades silábicas. La sílaba, se puede utilizar como elemento primordial en un ASR para el español, ya que se tiene una marcada semejanza en la forma en cómo se pronuncia los fonemas a la forma en que escribe, entonces se toman las reglas de las sílabas y se identifican para el español la existencia de 27 letras y reglas tales como: Que en las sílabas tiene que ir por lo menos una vocal. Los resultados obtenidos hacen uso de las técnicas de cadenas ocultas de Markov. Al final se realiza la implantación usando un Sistema Basado en el Conocimiento y como conclusiones se tienen que: los modelos basados en sílabas pueden ser conducidos a remover las ramificaciones durante la ejecución, además éstos son la unidad de organización natural para reducir la computación redundante y que define el espacio de búsqueda.

f) Aplicación de tecnología del habla en la enseñanza del español [33]

En este proyecto se desarrolló una aplicación que tiene un método para la pronunciación correcta de palabras o frases en español mexicano, siendo útil para la enseñanza y aprendizaje del idioma, asistido por la computadora. Para este proyecto utilizaron la herramienta CLSU Toolkit [34] que permite no sólo hacer el reconocimiento de la voz sino que provee de una serie de herramientas gráficas que permiten crear una interfaz que a su vez muestra un personaje (Agente conversacional llamado Baldi en 3d) que se puede adecuar para interactuar con el usuario. Se debe tener en cuenta que si se pronuncian fonemas que no están en el sistema, éste los rechazará o buscará los más parecidos. Como el sistema es para enseñar español mexicano a estudiantes hablantes de inglés tomaron los fonemas que provienen del inglés norteamericano que no pertenecen a los que se utilizan en el español y se incluyeron. El proceso se completa con el entrenamiento de una red neuronal.

1.5.2. Proyectos relacionados con el tratamiento de señales de voz

a) Estimación de la frecuencia fundamental de señales de voz del suroccidente colombiano aplicando la técnica Wavelet [20]

El objetivo de este proyecto fue implementar un sistema de estimación de la Frecuencia Fundamental en señales de voz del suroccidente colombiano mediante la técnica *Wavelet*. Para ello se llevó a cabo un estudio de la teoría *Wavelet* y sus aplicaciones en el procesamiento de señales de voz, identificando las familias *wavelet* más adecuadas para este tipo de señales; también se desarrolló un algoritmo de estimación de la Frecuencia Fundamental, en señales de voz del suroccidente colombiano, utilizando la técnica *wavelet*. Para finalizar se realiza la simulación y evaluación del desempeño del algoritmo de estimación de la Frecuencia Fundamental de la voz con respecto a otras herramientas disponibles (tal como el software “*Speech Filing System*”). Su aporte principal consistió en la implementación de un sistema de estimación de la frecuencia fundamental de las señales de voz colombianas.

b) *Adecuación de señales de voz para el sistema de estimación de la frecuencia fundamental CGAWAVF* [15]

El objetivo de este proyecto fue adecuar las señales de voz grabadas en exteriores, al sistema de detección de la frecuencia fundamental CGAWAVF (Complex Gaussian Adapted *Wavelet* Family por sus siglas en inglés). Para esto fue necesario llevar a cabo un análisis de las fuentes de interferencia y ruido contaminantes de las señales de voz que son grabadas en exteriores; un análisis de los algoritmos apropiados para la eliminación de interferencias y ruido contaminantes de señales de voz y la respectiva implementación de la etapa de pre-procesamiento para adecuación de la señal de voz al sistema CGAWAVF. Su aporte principal consistió en el desarrollo de dos algoritmos: Un algoritmo para la eliminación de interferencias y ruidos presentes en la señales de voz grabadas en exteriores y un algoritmo de segmentación automática de señales de voz para extracción de sonidos vocálicos, integrados al sistema de detección de la frecuencia fundamental de señales de voz CGAWAVF; con aplicaciones inmediatas en la medicina y en sistemas de comunicaciones y de seguridad.

c) *Compresión y descompresión de voz mediante técnicas de procesamiento digital de imágenes utilizando wavelets* [35]

El objetivo de este proyecto fue desarrollar una técnica para la compresión y descompresión de voz, mediante técnicas de procesamiento digital de imágenes utilizando la Teoría Wavelets para convertir un determinado periodo de voz en una única imagen. Una vez desarrollada esta etapa se aplicó el estándar de compresión de imágenes JPEG2000, este estándar de edición y procesado permiteextraer los bytes comprimidos que forman una determinada región y re-ensamblarlos en una nueva secuencia comprimida sin necesidad de realizar descompresión. En resumen se puede extraer una región de la imagen, recortar la imagen o rotarla sin necesidad de descomprimirla. Sus aportes consisten en analizar el comportamiento que tiene la Teoría de Wavelets en la codificación de tramas de voz, al tratar un archivo de voz como si fuese una imagen.
Como se ve estos últimos proyectos realizados en nuestra alma mater son importantes, puesto que tratan explícitamente la extracción de características en señales de voz, sirviendo como referencia para este proyecto; ya que ésta es una de las fases que se debe seguir para elaborar uno de los módulos de el sistema de reconocimiento; además de realizar este módulo, el sistema a desarrollar tendrá otros módulos entre los que se destaca el módulo de reconocimiento de patrones, que tomará como base los datos procesados con técnicas de extracción de características, y servirán para realizar el proceso de entrenamiento y reconocimiento propios de un sistema de reconocimiento del habla.

1.5.3. Otros proyectos

a) *Elaboración de un corpus balanceado para el cálculo de modelos acústicos usando la web* [36]

En él se presenta una metodología para la elaboración de un corpus balanceado fonéticamente para el español mexicano. En este proyecto se tienen en cuenta los elementos necesarios para la construcción de este reconocedor, se resalta la importancia de contar con una colección de grabaciones que servirá de base para el cálculo de los modelos acústicos pertinentes. Dicha colección de grabaciones debía cuidar ciertos aspectos para que el reconocedor fuera lo más robusto posible. Dos de estos aspectos

fueron abordados en el trabajo en mención: primero el corpus oral debe ser rico, es decir, debe contener todos los fonemas del español mexicano, y debe ser balanceado, lo que quiere decir, debe conservar la distribución fonética del español mexicano. *En el caso de este proyecto no se cuenta con construcciones previas de corpus, lo cual tendrá que realizarse como parte del proyecto.*

En los proyectos referenciados anteriormente se puede observar que se han realizado diversos estudios para reconocimiento automático en otras lenguas, pero no para lengua Nasa Yuwe, también es observable que para los proyectos mencionados se ha predispuesto el uso de una técnica predeterminada, en el caso de este proyecto se realizara una comparación entre las técnicas más usadas en reconocimiento del habla y escoger la que menor error de reconocimiento presente. Igualmente se ha identificado el pre procesamiento de las señales de voz que implica un trabajo especial para la extracción de características, éstas se obtiene de un conjunto de datos de audio que se denomina corpus que debe estar formalizado por características técnicas, lingüísticas, sociales que se explicaran a continuación.

Para finalizar este capítulo podemos establecer que las técnicas de reconocimiento del habla elegidas fueron redes neuronales, Modelos ocultos de Markov y alineamiento dinámico en el tiempo, con las cuales se realizaran las comparaciones de menor nivel de error de reconocimiento, usando técnicas de pre procesamiento como son LPC y MFCC, por sus características de rendimiento y viabilidad en el uso de un conjunto de datos pequeño.

Cabe resaltar que dentro de los antecedentes de reconocimiento del habla estudiados, en ninguno se enuncia estudios comparativos de estas técnicas o construcción de corpus como se mostrará en este proyecto.

2. CONSTRUCCIÓN DEL CORPUS

La construcción del corpus es de vital importancia para este proyecto, ya que no existe antecedente alguno para esta lengua. Para su elaboración nos basamos en proyectos que sirven de guía para su elaboración [36] y estudio en lenguas minoritas [37] como es el caso de esta lengua.

Cuando se habla de corpus se hace referencia a una colección de textos orales o escritos de una lengua, los cuales han sido seleccionados a partir de unos criterios lingüísticos explícitos [38] que se pasan en formato electrónico mínimamente procesados, estos se utilizan como muestras representativas para el abastecimiento de datos de un estudio sistemático. Los corpus se encuentran clasificados de acuerdo a unos criterios [37] como puede ser el soporte original de los datos, en esta clasificación están los *corpus escritos o corpus orales.* Es resaltable que para este proyecto se cuenta con avances en la construcción de un alfabeto [39] y diccionario de palabras para esta lengua [40] lo que permitió no solo tener datos orales sino ya escritos.

Otro tipo de clasificación está dada por el número de lenguas que lo constituyen, aquí se clasifican en corpus *monolingües y multilingües,* en este caso se clasifica como un corpus monolingüe con variación lingüística. Siendo el estudio sobre la lengua Nasa Yuwe reciente, los logros que se han obtenido permiten crear la base para la primera aproximación de un corpus básico, dentro de la construcción de este se puede ver partes como son la unificación de un alfabeto [39, 40, 41] del que obtendremos la escritura de las palabras que serán parte del diccionario de pronunciación de un conjunto de palabras monosílabas y bisílabas que contienen las vocales orales y nasales de esta lengua, entendiendo que fueron escogidas dentro de un alfabeto que comprende 38 consonantes y 32 vocales.

A continuación en la tabla 1 se visualiza el conjunto de vocales, en la primera columna de la izquierda se encuentran los modos de articulación y en la primera fila los 2 conjuntos de vocales que se presentan en la lengua Nasa Yuwe, en cuanto al modo de articulación hace referencia a la postura que adoptan los órganos cuando se producen los sonidos [42].

	Vocales orales				Vocales nasales			
Simples	a	e	i	u	Ã	ẽ	ĩ	ũ
glotalizadas	a’	e'	i'	u'	ã'	ẽ’	ĩ'	ũ'
Aspiradas	ah	eh	ih	uh	Ãh	ẽh	ĩh	ũ h
Alargadas	aa	ee	ii	uu	Ãa	ẽe	ĩi	ũu

Tabla 1. Fonemas del Nasa Yuwe (vocales) [39]

2.1. Criterios de elección de las vocales

Dentro de las características que se buscaban para obtener las vocales de la lengua Nasa Yuwe [39] se llegó a la construcción de un diccionario de 72 palabras donde se pueden

observar la presencia de vocales orales y nasales simples. La configuración de estas palabras es la siguiente:

Para palabras monosílabas, tenemos configuraciones de palabras que empiezan con una vocal y terminan con una consonante (VC) un ejemplo de estos es la palabra ***ab (surco)*** que contiene una a oral simple y la palabra ***Ēs (piojo)*** que contiene una Ē nasal simple, también encontramos palabras que empiezan por una consonante y terminan con una vocal (CV) por ejemplo encontramos la palabra ***Khī (comprometerse a)*** también se encuentran configuraciones como la combinación consonante-vocal-consonante (**CVC)**de la palabra ***KhĀg (hormiga)*** y por último encontramos palabras como **Twaçe (el cafetero)**donde se observan dos vocales orales simples. Las anteriores configuraciones son el tipo de palabras que se usan para la creación del sistema de reconocimiento; para ver el diccionario completo puede observar el Anexo B.

2.2. Niveles de etiquetación

Dentro del conjunto de vocales nasales y orales simples que se mencionaron anteriormente, se tomaron la forma de escritura creada para la unificación del alfabeto de la lengua Nasa Yuwe [39]. En consecuencia se crearon 2 niveles de etiquetación, un nivel fonológico, que corresponde a los fonemas y un nivel ortográfico, que corresponde a la escritura de las palabras.

Una de las herramientas que se utilizaron para realizar este proceso fue Audacity [43], para segmentar las palabras y vocales de forma manual teniendo en cuenta la señal de audio donde se aprecia las características de tiempo y forma de la onda, donde se puede ver que las vocales muestran un nivel de energía más destacable que las consonantes. En la Figura 17, se puede observar una onda de la vocal Ē seguida por la consonante S palabra (Ēs).

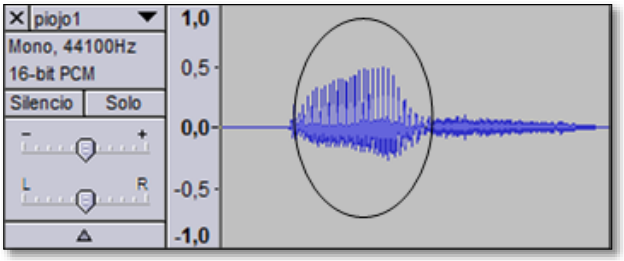


Figura 17. Forma de onda de la palabra Ēs: Piojo, seleccionada la Ē Nasal

En este proceso de etiquetado y segmentación de la señal a veces se torna difícil pues no es muy claro ver donde empieza y termina una vocal como se ve en la Figura 18 ya sea por encontrarse en medio de dos vocales o una consonante sonora como la palabra *kyĀduu*: rodear

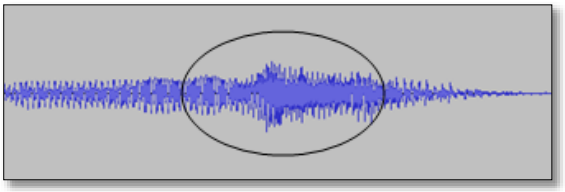


Figura 18. Forma de onda de la palabra kyĀduu: rodear

Cuando los fonemas no son tan fáciles de diferenciar es necesario utilizar herramientas que suministren otras características como es el caso de Praat [14], que nos permiten ver información visual del espectro de onda de los fonemas que hacen parte de la onda. En la Figura 19 podemos observar de una forma más detalla la información del espectro de la letra *Ā* de la palabra *kyĀduu*de esta forma se puede extraer mejor la parte de la señal haciendo caso a características como el pulso glotal, la intensidad de la señal, la visualización de formantes, el pitch y su intensidad en la vista de espectro.

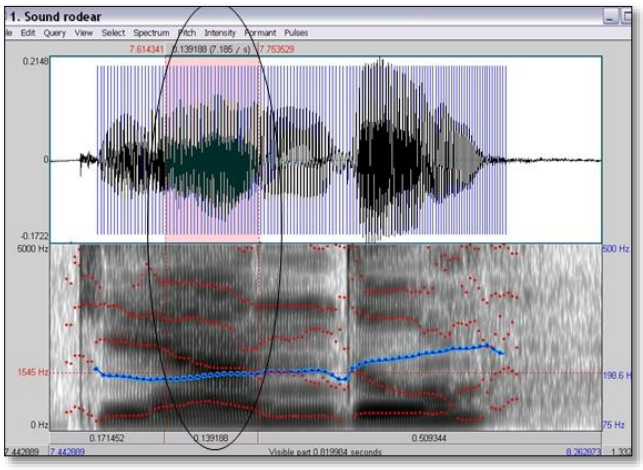


Figura 19. Forma de Onda y Espectrograma de la señal de voz de la palabra kyĀduu (rodear)

2.3. Características del Corpus

Las vocales se describen con unos determinados rasgos, apertura de la boca, posición de la lengua, posición de los labios, entre otras [42]. En el caso de Nasa Yuwe existen rasgos de alargamiento, de nasalidad, de aspiración y una combinación de estas, además desde el punto de vista fonético la información es más manejable si se usan vocales nasales y orales simples que alargadas o aspiradas.

2.3.1. Características lingüísticas

La decisión de tomar las vocales orales y nasales simples (ver Tabla 1) dentro del conjunto está dada porque son la base para que un hablante pueda aprender más fácil las demás si primero se aprenden estas [39].

2.3.2. Características socio-lingüísticas de los hablantes

El corpus elaborado se basa en un conjunto de pronunciaciones de frases que se tomaron a 5 hablantes nativos (3 hombres y 2 mujeres) cumpliendo con tener un corpus balanceado en género, también se tuvo en cuenta escoger hablantes de distintos resguardos indígenas ya que esta lengua presenta variantes. Las características más relevantes de ellos son las siguientes:

Nombre	Edad	Nivel Educativo	Localización
Roxana Chocue	35	Licenciada en Etnoeducación	López Adentro Caloto
BenildaTrochez	47	Licenciada en Etnoeducación	López Adentro Caloto
Adonias Perdomo	50	Universitario	resguardo Pitayó
Abelardo Ramos	50	Universitario	Resguardo Togoima (tierra adentro)
José Fidel Secue	51	Bachiller Pedagógico	López Adentro Caloto

Tabla 2. Características socio-lingüísticas de los hablantes

2.3.3. Características Técnicas

a) *Recolección de las muestras de audio.* En esta actividad se contó con la colaboración del Grupo de Estudios Lingüísticos Pedagógicos y Socio Culturales del Sur-occidente Colombiano, el cual ya tiene los contactos para la toma de las muestras o grabaciones de audio con hablantes nativos de los resguardo del Cauca; igualmente dentro de los trabajos que realizan, se encuentra la identificación y clasificación de las palabras más comunes en el contexto de estas comunidades, lo que permitió agilizar el proceso de selección de las palabras que contienen el subconjunto de vocales de interés para este proyecto.

b) *Preparación de las muestras de audio.* Una vez se obtuvieron las muestras de audio se procedió a realizar la división y enumeración de las palabras pronunciadas por los integrantes de la comunidad Nasa. Como resultado se obtuvieron las repeticiones de cada una de las palabras, generalmente se trata de palabras monosílabas y bisílabas, que contienen el subconjunto de vocales seleccionadas para este proyecto, de estas repeticiones, se seleccionaron un porcentaje para el proceso de entrenamiento del sistema y un conjunto más pequeño para el proceso de prueba de reconocimiento de las vocales.

Para la grabación de estas muestras acústicas se realizó en una habitación en un ambiente normal (con sonido y ruido ambiente) por lo que siempre existirá presencia de ruido, para controlar que exista una buena recolección del audio se hizo el ejercicio en momentos donde no se encontraba muchas personas en el edificio para evitar interrupciones y sonidos extraños, las grabaciones se realizaron a puerta cerrada con una grabadora con micrófono unidireccional de 90 grados de captura (graba sonido que encuentra en un radio de 90 grados) con un protector que inhibe el ruido generado por acercarse demasiado al micrófono o jadeo de la persona de la cual se extraen las repeticiones de las palabras.

Otros datos importantes a configurar son:

- Frecuencia de grabación: 44100 Hz en Estéreo
- Formato de grabación: WAV
- Formato de muestra 16 bits
- Grabación a un canal (Mono)
- Tamaño de la ventana FFT 256
- Tipo de ventana de Hamming

2.4. Estadísticas

En la elaboración del corpus se realizaron un total de 2.824 grabaciones distribuidas en 362 carpetas, donde se relacionan las vocales nasales y orales de la lengua Nasa Yuwe, éstas se encuentran ordenadas por palabras y vocales dentro de las palabras elegidas, se tomaron un conjunto de 54 palabras orales simples y 32 palabras nasales simples, para cada palabra se realizaron entre 3 a 4 repeticiones. Finalmente al contabilizar las vocales (ver Anexo C) de interés de este proyecto se tuvieron las siguientes cantidades (Figuras 20, 21, 22):

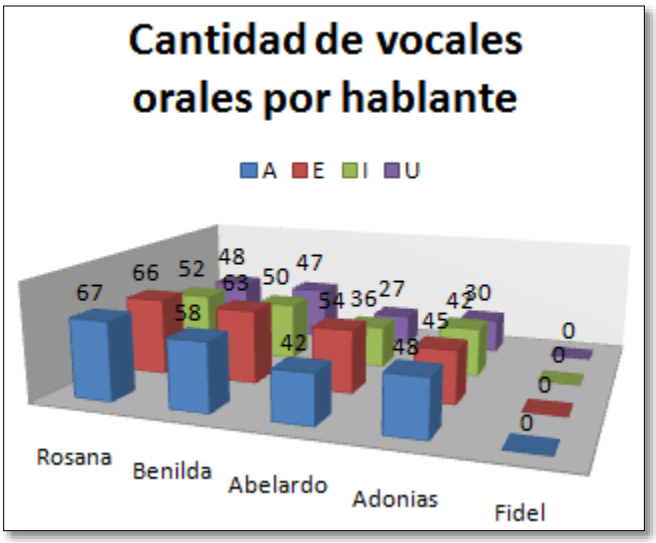


Figura 20. Cantidad de vocales orales por hablante

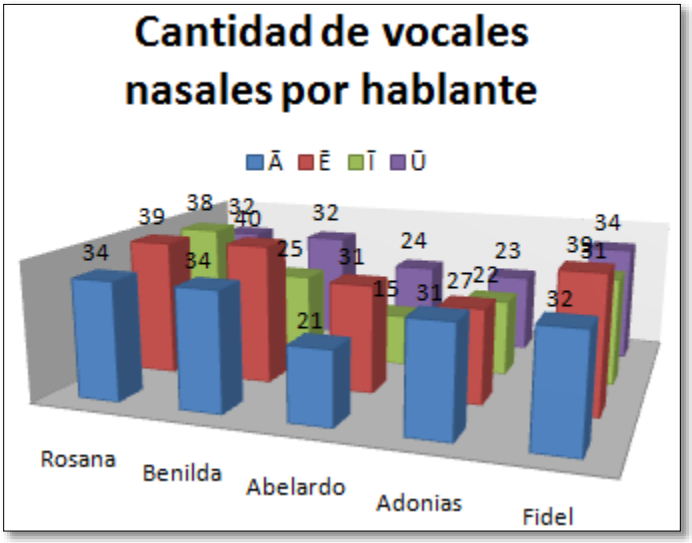


Figura 21. Cantidad de vocales nasales por hablante

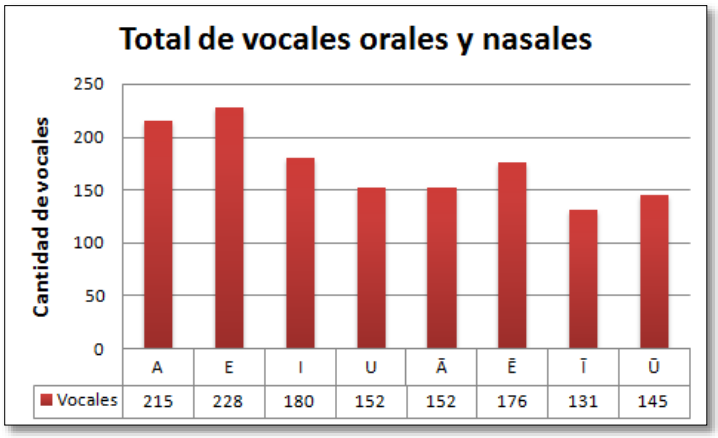


Figura 22. Total de vocales orales y nasales

2.5. Extracción de características

Los coeficientes Cepstrales en escala de frecuencias Mel (MFCC) sirven para representar el habla, basados en la percepción auditiva humana. Los MFCC permiten modelar la respuesta auditiva humana más apropiadamente que las bandas espaciadas linealmente de la transformada de Fourier (FT) lo que permite un procesado de datos más eficiente [44]. Teniendo en cuenta que la unidad básica del habla a reconocer son vocales, se procede a aplicar métodos de extracción de características sobre esas señales de voz a este proceso se le conoce como ventaneo de la señal y se ilustra en la Figura 23:

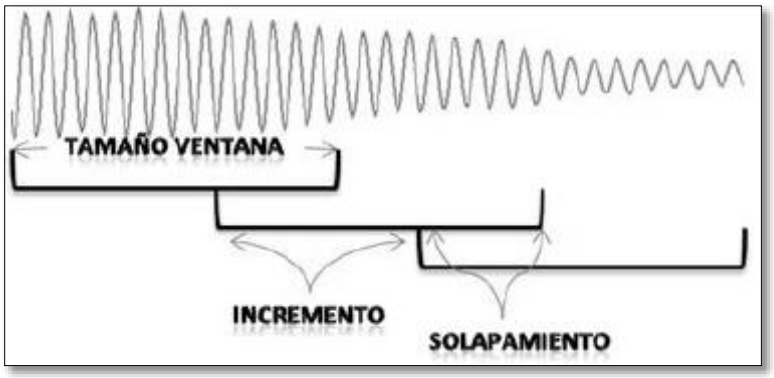


Figura 23. Ventaneo de la señal [11]

La elección del tamaño del incremento se realiza teniendo en cuenta el tamaño de la ventana de proceso de reconocimiento y se deben realizar pruebas con estos parámetros para ajustarlos de manera adecuada, para este proyecto se utiliza un muestreo de 44100 Hz en Mono. Posteriormente se debe realizar el ventaneo de las muestras. Para este fin se emplean los tipos de funciones de ventaneo, los más utilizados son: el tipo Hann y el tipo Hamming [44, 45].

El resultado final de esta fase es la construcción del corpus que consta de 8 directorios de las 8 vocales (4 nasales y 4 orales), extraídas de un conjunto de palabras monosílabas y bisílabas del diccionario de la lengua Nasa Yuwe(Ver Figura 24).

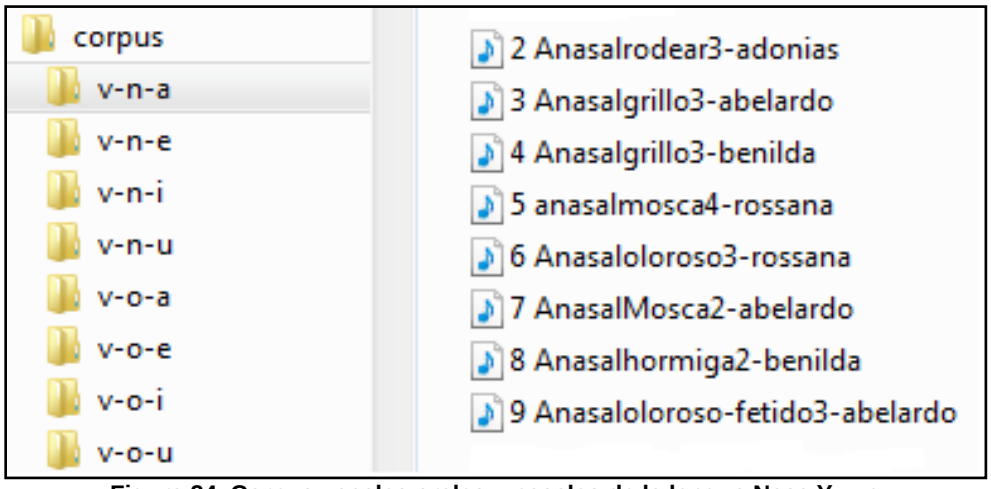


Figura 24. Corpus vocales orales y nasales de la lengua Nasa Yuwe

De las vocales del corpus finalmente se pudo obtener de 12 coeficientes LPC (ver Figura 25) por cada onda y en el caso de MFCC (ver Figura 26) se extrae una matriz de características, en ninguno de los casos se hizo reducción de la dimensión.

```
-0.597661 -0.320712 0.007360 0.167468 -0.176416 -0.037621 -0.074575 0.067131 0.017345 -0.212521 -0.121822 0.301479
```

Figura 25. Coeficientes LPC

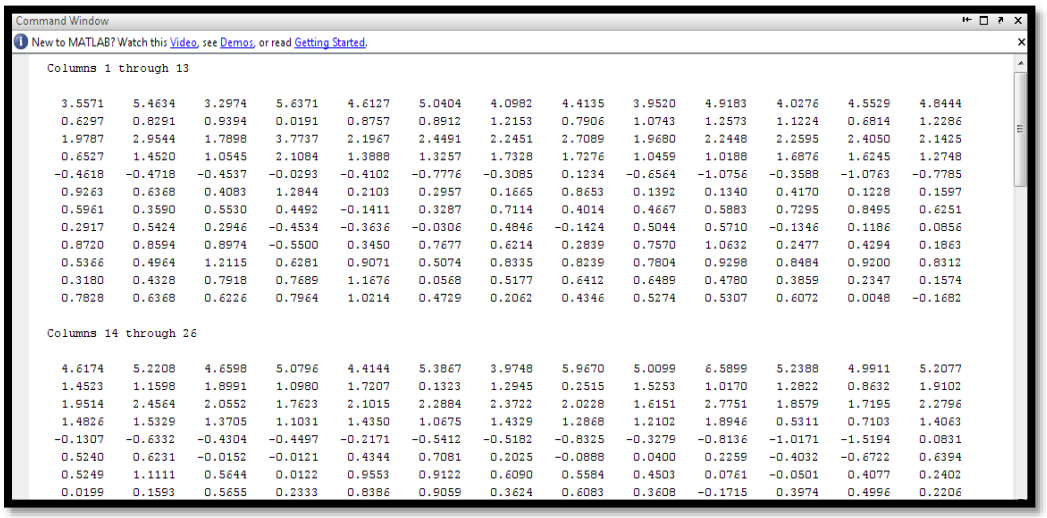


Figura 26. Matriz de Coeficientes MFCC

3. ANÁLISIS Y EXPERIMENTACIÓN DE LAS TÉCNICAS DE RECONOCIMIENTO DEL HABLA

Para realizar las pruebas con el corpus tomado de los distintos hablantes de la lengua Nasa Yuwe, se definió varias distribuciones del conjunto de vocales para los experimentos a los cuales se les aplicó las tres técnicas seleccionadas para este proyecto. El primero experimento compuesto por 4 directorios y cada directorio con 130 ondas de vocales orales simples; el segundo experimento compuesto por 4 directorios con 130 ondas de vocales nasales simples en cada uno; estos conjuntos son una combinación de ondas de los hablantes tanto de mujeres como de hombres; y el tercer experimento compuesto por 8 directorios con 130 ondas de vocales orales y nasales simples entre hombres y mujeres. El objetivo de estos experimentos es obtener resultados que permitan establecer con cuales características se comportan mejor las tres técnicas. Además se probaron las técnicas con hombres y mujeres por separado (ver Anexo D).

Para realizar las pruebas a las tres técnicas seleccionadas en este proyecto, se abordaron los 3 experimentos de la siguiente manera:

Experimento 1
Pruebas usando vocales a, e, i y u, orales simples de hablantes Nasa Yuwe. Este experimento está compuesto por 4 directorios (que representan las clases) y en cada uno con 130 vocales combinados entre pronunciaciones de hombres y mujeres.

Experimento 2
Pruebas usando vocales a, e, i y u nasales simples, de hablantes Nasa Yuwe. Este experimento está compuesto por 4 directorios (que representan las clases) y en cada uno con 130 vocales de hombres y mujeres.

Experimento 3
Pruebas usando vocales a, e, i y u orales y nasales simples, de hablantes Nasa Yuwe.
Este experimento está compuesto por 8 directorios (que representan las clases) y en cada uno con 130 vocales de hombres y mujeres.

Para la realización de los experimentos se tuvo en cuenta las siguientes condiciones:

- En el corpus, cada directorio que guarda las grabaciones de una vocal, debe tener grabaciones de mujeres y hombres.
- Las grabaciones de audio deben estar almacenadas de forma aleatoria, es decir, que en lo posible dos vocales de una misma clase (ejemplo clase vocal a), extraídas de una misma palabra y pronunciada por un mismo hablante, no pueden estar seguidas. De esta forma se obtiene un mejor entrenamiento.
- Tener en cuenta que las últimas grabaciones son las que se tomarán para clasificar en cada directorio. De este modo es conveniente que las grabaciones para clasificación, contengan vocales que pertenezcan a las diferentes palabras de la lengua Nasa Yuwe.
- Tomar las vocales que presenten mayor claridad, (sin efecto coarticulatorio) sin embargo, al tener tan pocas grabaciones no se tiene la posibilidad de eliminar las vocales que no sean las ideales, especialmente en el caso de las vocales nasales simples, cuyo número es muy reducido.

Las muestras tomadas se realizaron en presencia de ruido (la presencia de ruido acerca más a la realidad, es decir, en donde siempre existe ruido ambiente), con 5 hablantes (2 mujeres y 3 hombres).

Para los experimentos se realizaron en un equipo de cómputo con las siguientes características:
-Portátil HP Pavilion dv4-1222nr
-AMD Turion X2 Dual-Core Mobile Processor RM-72 (2.1 GHz, 1MB L2 Cache)
-250 GB de disco duro
-4 GB de memoria RAM
-Tarjeta de video ATI Radeon HD 3200

Para la grabación de las señales de audio se requiere como mínimo una grabadora con las siguientes características, las cuales se usaron en este proyecto:
-Una grabadora digital que capture a mínimo 44100Hz.
-16 bits de tasa de cuantización.
- Un supresor de ruido

Condiciones ideales de la grabadora [42]
-Micrófonos omnidireccionales.
-Condensador de electro con fuente de poder estable.
-Batería de corriente continua y no de fuente con corriente alterna.

La configuración para la extracción de los coeficientes se realizó de la siguiente forma: 1) Cuando se utilizó MFCC la frecuencia fue de 44100 Hz, para obtener 12 coeficientes se usó una longitud de fotograma en el muestreo de 0.03 de la frecuencia y ventana de Hamming. 2) Para la obtención de coeficientes con Predicción Lineal para obtener 12 coeficientes se utilizó una longitud de fotograma igual a la longitud de la señal.

Para calcular el porcentaje de error de reconocimiento, se tuvo en cuenta el número de vocales que hayan sido correctamente clasificadas (CC),dividido por el número de ondas de pruebas a clasificar (N) para así obtener el porcentaje de reconocimiento de vocales correctamente reconocidas (PC) (ver Fórmula 19) el cual es restado del 100 por ciento de vocales a clasificar obteniendo así el porcentaje de error de reconocimiento (PE) (ver Fórmula 20).

$$PC = \frac{CC * 100}{N}$$ **Fórmula 19**

$$PE = 100 - PC$$ **Fórmula 20**

3.1. Resultados mediante la técnica Dynamic Time Warping

Experimento 1

% Entrenamiento	70		80		90	
Numero de ondas	entrenamiento	pruebas	Entrenamiento	pruebas	entrenamiento	pruebas
	364	156	416	104	468	52
% Correcto	83.97		86.54		92.31	
% Error	16.03		13.46		7.69	
Ondas correctamente clasificadas	131		90		48	

Tabla 3. Experimento 1 con la técnica Dynamic Time Warping

Experimento 2

% Entrenamiento	70		80		90	
Numero de ondas	entrenamiento	pruebas	entrenamiento	pruebas	entrenamiento	pruebas
	364	156	416	104	468	52
% Correcto	67.31		69.23		65.38	
% Error	32.69		30.77		34.62	
Ondas correctamente clasificadas	105		72		34	

Tabla 4. Experimento 2 con la técnica Dynamic Time Warping

Experimento 3

% Entrenamiento	70		80		90	
Numero de ondas	entrenamiento	pruebas	entrenamiento	pruebas	entrenamiento	pruebas
	728	312	832	208	936	104
% Correcto	48.40		48.08		54.81	
% Error	51.60		51.92		45.19	
Ondas correctamente clasificadas	151		100		57	

Tabla 5. Experimento 3 con la técnica Dynamic Time Warping

Basados en los resultados se puede observar que a mayor porcentaje de ondas en el entrenamiento generalmente el porcentaje de error de reconocimiento disminuye. Esta técnica reconoce mayor número de ondas de vocales orales que nasales sin importar el porcentaje tomado para entrenamiento. Finalmente, cuando el corpus evaluado está compuesto de vocales orales y nasales, se nota un aumento del porcentaje del error de reconocimiento a medida que se incrementa el porcentaje utilizado para entrenamiento (ver Figura 26).

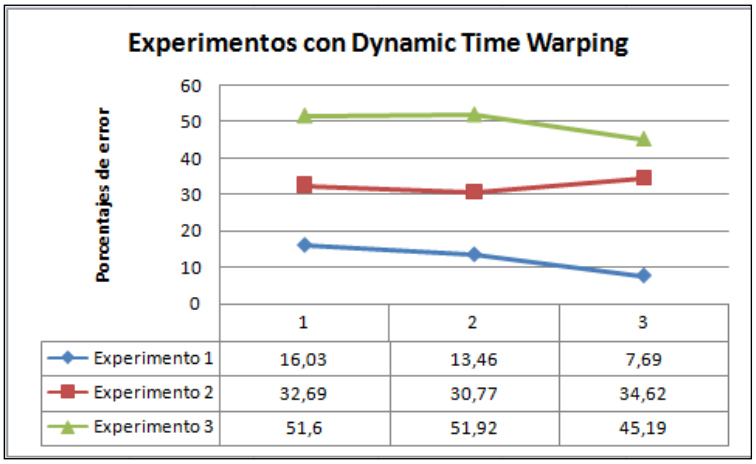


Figura 26. Resumen de experimentos 1,2 y 3 con DTW

3.2. Resultados mediante la técnica Redes Neuronales Artificiales

Propagación mínima de 0.0001 hasta un máximo de 0.1 y el número de capas usadas es 2. La primera capa es usada para calcular entradas y su entrada de red. La segunda capa calcula su entrada y sus entradas de red.

Experimento 1

En esta ejecución se utilizaron el 70% de las ondas de vocales para entrenamiento por cada clase, lo que equivale a 364 ondas para entrenamiento y 156 ondas para pruebas					
Propagación	**0.1**	**0.01**	**0.001**	**0.0001**	**0.00001**
% Correcto	45.51	80.13	92.31	37.82	25
% Error	54.49	19.87	7.69	62.18	75
Ondas reconocidas correctamente	71	125	144	59	39
En esta ejecución se utilizaron el 80% de las ondas de vocales para entrenamiento por cada clase, lo que equivale a 412 ondas para entrenamiento y 108 ondas para pruebas					
% Correcto	45.37	79.63	93.52	43.52	25
% Error	54.63	20.37	6.48	56.48	75
Ondas reconocidas correctamente	49	86	101	47	27
En esta ejecución se utilizaron el 90% de las ondas de vocales para entrenamiento por cada clase, lo que equivale a 464 ondas para entrenamiento y 56 ondas para pruebas					
% Correcto	55.36	85.71	94.64	39.29	25
% Error	44.64	14.29	5.36	60.71	75
Ondas reconocidas correctamente	31	48	53	22	14

Tabla 6. Experimento 1 con la técnica Redes Neuronales Artificiales

En la Tabla 6, el porcentaje de error de reconocimiento es menor siempre en la tercera columna lo que implica que para el corpus formado por hombres y mujeres de vocales orales es indiferente del porcentaje usado para entrenamiento y la propagación de la red neuronal.

Experimento 2

En esta ejecución se utilizaron el 70% de las ondas de vocales para entrenamiento por cada clase, lo que equivale a 364 ondas para entrenamiento y 156 ondas para pruebas					
Propagación	**0.1**	**0.01**	**0.001**	**0.0001**	**0.00001**
% Correcto	53.85	87.82	89.74	37.18	25
% Error	46.15	12.18	10.26	62.82	75
Ondas reconocidas correctamente	84	137	140	58	39
En esta ejecución se utilizaron el 80% de las ondas de vocales para entrenamiento por cada clase, lo que equivale a 412 ondas para entrenamiento y 108 ondas para pruebas					
% Correcto	57.41	88.89	90.74	43.52	25
% Error	42.59	11.11	9.26	56.48	75
Ondas reconocidas correctamente	62	96	98	47	27
En esta ejecución se utilizaron el 90% de las ondas de vocales para entrenamiento por cada clase, lo que equivale a 464 ondas para entrenamiento y 56 ondas para pruebas					
% Correcto	57.14	89.29	91.07	41.07	25
% Error	42.86	10.71	8.93	58.93	75
Ondas reconocidas correctamente	32	50	51	23	14

Tabla 7. Experimento 2 con la técnica Redes Neuronales Artificiales

Como se puede notar en la Tabla 7, el porcentaje de error de reconocimiento es menor en la tercera columna lo que implica que para el corpus formado por hombres y mujeres de vocales nasales es indiferente del porcentaje usado para entrenamiento y la propagación de la red neuronal.

Experimento 3

En esta ejecución se utilizaron el 70% de las ondas de vocales para entrenamiento por cada clase, lo que equivale a 728 ondas para entrenamiento y 312 ondas para pruebas					
Propagación	**0.1**	**0.01**	**0.001**	**0.0001**	**0.00001**
% Correcto	55.13	90.06	78.53	12.82	12.50
% Error	44.87	9.94	21.47	87.18	87.50
Ondas reconocidas correctamente	172	281	245	40	39
En esta ejecución se utilizaron el 80% de las ondas de vocales para entrenamiento por cada clase, lo que equivale a 824 ondas para entrenamiento y 216 ondas para pruebas					
% Correcto	54.63	86.57	82.41	12.96	12.50
% Error	45.37	13.43	17.59	87.04	87.50
Ondas reconocidas correctamente	118	187	178	28	27
En esta ejecución se utilizaron el 90% de las ondas de vocales para entrenamiento por cada clase, lo que equivale a 928 ondas para entrenamiento y 112 ondas para pruebas					
% Correcto	56.25	85.71	86.61	12.50	12.50
% Error	43.75	14.29	13.39	87.50	87.50
Ondas reconocidas correctamente	63	96	97	14	14

Tabla 8. Experimento 3 con la técnica Redes Neuronales Artificiales

Cuando el corpus está compuesto por vocales orales y nasales, de hombres y mujeres, se nota que el decremento de la propagación de la red deteriora el porcentaje de reconocimiento y en este sentido el número de las vocales reconocidas es aceptable

cuando la propagación oscila entre 0.001 y 0.01, aunque el porcentaje de ondas del corpus para entrenamiento aumente.

En los dos primeros experimentos se refleja una indiferencia de la propagación y porcentaje usado para entrenamiento. Según los resultados obtenidos la propagación que permite los mejores resultados cuando se evalúa sólo vocales orales o sólo vocales nasales para hombres y mujeres es 0.001.

Finalmente, en el tercer experimento, cuando se tiene un corpus conformado por vocales y orales pronunciadas por hombres y mujeres se observa un contraste de acuerdo con los anteriores resultados obtenidos; ya que la propagación debe ser 0.01 al utilizar un 70 y 80 por ciento para entrenamiento, pero para el 90 por ciento en el entrenamiento se debería tener una propagación de 0.001.

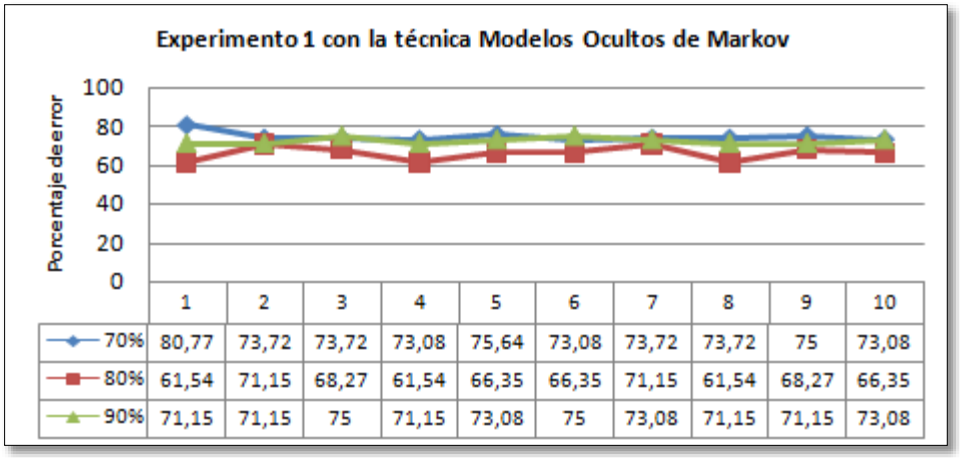
3.3. Resultados mediante la técnica Modelos Ocultos de Markov

Experimento 1

Cantidad de clases:	4	Cantidad de muestras de cada clase:	130
Total ondas de vocales de Nasa Yuwe:	520	Número de estados:	3

El porcentaje utilizado es 70% lo que equivale a 364 ondas para entrenamiento y 156 ondas para pruebas										
Ciclos Baum-Welch	**30**	**60**	**90**	**120**	**150**	**180**	**210**	**240**	**270**	**300**
% Correcto	19.23	26.28	26.28	26.92	24.36	26.92	26.28	26.28	25	26.92
% De error	80.77	73.72	73.72	73.08	75.64	73.08	73.72	73.72	75	73.08
Ondas reconocidas correctamente	30	41	41	42	38	42	41	41	39	42
El porcentaje utilizado es 80% lo que equivale a 416 ondas para entrenamiento y 104 ondas para pruebas										
% Correcto	38.46	28.85	31.73	38.46	33.65	33.65	28.85	38.46	31.73	33.65
% De error	61.54	71.15	68.27	61.54	66.35	66.35	71.15	61.54	68.27	66.35
Ondas reconocidas correctamente	40	30	33	40	35	35	30	40	33	35
El porcentaje utilizado es 90% lo que equivale a 468 ondas para entrenamiento y 52 ondas para pruebas										
% Correcto	28.85	28.85	25	28.85	26.92	25	26.92	28.85	28.85	26.92
% De error	71.15	71.15	75	71.15	73.08	75	73.08	71.15	71.15	73.08
Ondas reconocidas correctamente	15	15	13	15	14	13	14	15	15	14

Tabla 9. Experimento 1 con la técnica Modelos Ocultos de Markov



	1	2	3	4	5	6	7	8	9	10
70%	80,77	73,72	73,72	73,08	75,64	73,08	73,72	73,72	75	73,08
80%	61,54	71,15	68,27	61,54	66,35	66,35	71,15	61,54	68,27	66,35
90%	71,15	71,15	75	71,15	73,08	75	73,08	71,15	71,15	73,08

Figura 27. Experimento 1 con la técnica Modelos Ocultos de Markov

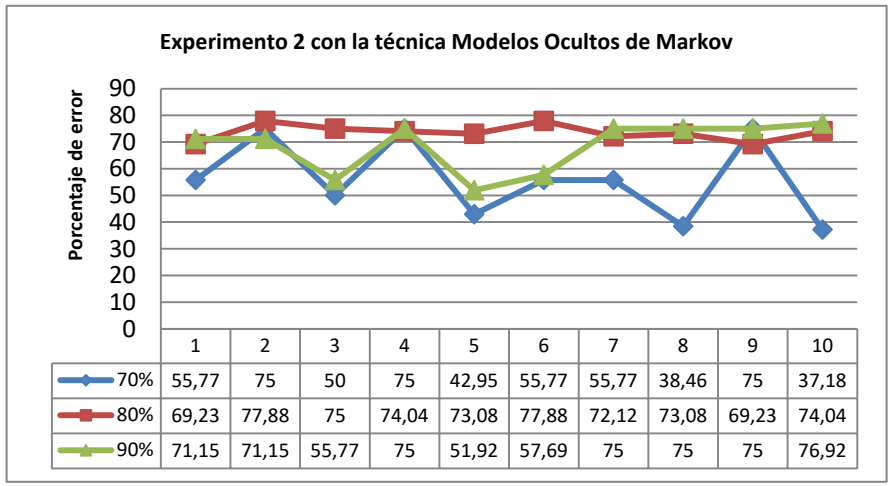
De acuerdo a la Tabla 9, los menores porcentajes de error de reconocimiento se presentan cuando se utiliza el 80% de ondas para el entrenamiento, en cuanto al número de ciclos de Baum–Welch se puede concluir que es variable, ya que el mayor número de ondas reconocidas se presentan en distintas cantidades de ciclos inclusive con el mismo porcentaje de entrenamiento. Para este primer experimento se destaca que el número de ciclos en común es 120 en las pruebas con los diferentes porcentajes de entrenamiento.

Experimento 2

Cantidad de clases: 4 Cantidad de muestras de cada clase: 130
Total ondas de vocales de Nasa Yuwe: 520 Número de estados: 3

El porcentaje utilizado es 70% lo que equivale a 364 ondas para entrenamiento y 156 ondas para pruebas										
Ciclos Baum-Welch	**30**	**60**	**90**	**120**	**150**	**180**	**210**	**240**	**270**	**300**
% Correcto	44.23	25	50	25	57.05	44.23	44.23	61.54	25	62.82
% De error	55.77	75	50	75	42.95	55.77	55.77	38.46	75	37.18
Ondas reconocidas correctamente	69	39	78	39	89	69	69	96	39	98
El porcentaje utilizado es 80% lo que equivale a 416 ondas para entrenamiento y 104 ondas para pruebas										
% Correcto	30.77	22.12	25	25.96	26.92	22.12	27.88	26.92	30.77	25.96
% De error	69.23	77.88	75	74.04	73.08	77.88	72.12	73.08	69.23	74.04
Ondas reconocidas correctamente	32	23	26	27	28	23	29	28	32	27
El porcentaje utilizado es 90% lo que equivale a 468 ondas para entrenamiento y 52 ondas para pruebas										
% Correcto	28.85	28.85	44.23	25	48.08	42.31	25	25	25	23.08
% De error	71.15	71.15	55.77	75	51.92	57.69	75	75	75	76.92
Ondas reconocidas correctamente	15	15	23	13	25	22	13	13	13	12

Tabla 10. Experimento 2 con la técnica Modelos Ocultos de Markov



	1	2	3	4	5	6	7	8	9	10
70%	55,77	75	50	75	42,95	55,77	55,77	38,46	75	37,18
80%	69,23	77,88	75	74,04	73,08	77,88	72,12	73,08	69,23	74,04
90%	71,15	71,15	55,77	75	51,92	57,69	75	75	75	76,92

Figura 28. Experimento 2 con la técnica Modelos Ocultos de Markov

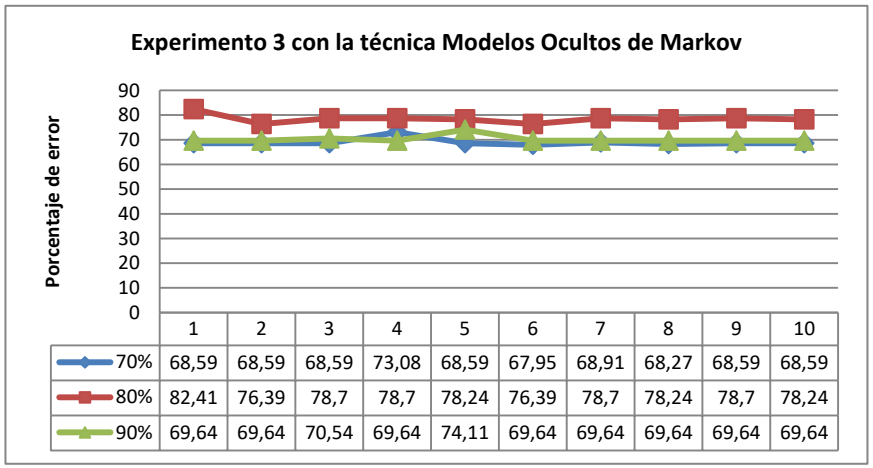
De acuerdo a la Tabla 10, los menores porcentajes de error de reconocimiento se presentan cuando se utiliza el 70% de ondas para el entrenamiento, en cuanto al número de ciclos de Baum–Welch se puede concluir que es variable, ya que el mayor número de ondas reconocidas se presentan en distintas cantidades de ciclos, inclusive con el mismo porcentaje de entrenamiento.

Experimento 3

Cantidad de clases:	8	Cantidad de muestras de cada clase:	130
Total ondas de vocales de Nasa Yuwe:	1040	Número de estados:	3

El porcentaje utilizado es 70% lo que equivale a 364 ondas para entrenamiento y 156 ondas para pruebas										
Ciclos Baum-Welch	**30**	**60**	**90**	**120**	**150**	**180**	**210**	**240**	**270**	**300**
% Correcto	31.41	31.41	31.41	26.92	31.41	32.05	31.09	31.73	31.41	31.41
% De error	68.59	68.59	68.59	73.08	68.59	67.95	68.91	68.27	68.59	68.59
Ondas reconocidas correctamente	98	98	98	84	98	100	97	99	98	98
El porcentaje utilizado es 80% lo que equivale a 416 ondas para entrenamiento y 104 ondas para pruebas										
% Correcto	17.59	23.61	21.30	21.30	21.76	23.61	21.30	21.76	21.30	21.76
% De error	82.41	76.39	78.70	78.70	78.24	76.39	78.70	78.24	78.70	78.24
Ondas reconocidas correctamente	38	51	46	46	47	51	46	47	46	47
El porcentaje utilizado es 90% lo que equivale a 468 ondas para entrenamiento y 52 ondas para pruebas										
% Correcto	30.36	30.36	29.46	30.36	25.89	30.36	30.36	30.36	30.36	30.36
% De error	69.64	69.64	70.54	69.64	74.11	69.64	69.64	69.64	69.64	69.64
Ondas reconocidas correctamente	34	34	33	34	29	34	34	34	34	34

Tabla 11. Experimento 3 con la técnica Modelos Ocultos de Markov



	1	2	3	4	5	6	7	8	9	10
70%	68,59	68,59	68,59	73,08	68,59	67,95	68,91	68,27	68,59	68,59
80%	82,41	76,39	78,7	78,7	78,24	76,39	78,7	78,24	78,7	78,24
90%	69,64	69,64	70,54	69,64	74,11	69,64	69,64	69,64	69,64	69,64

Figura 29. Experimento 3 con la técnica Modelos Ocultos de Markov

Los resultados obtenidos al ejecutar la técnica de Modelos Ocultos de Markov con el 70, 80 y 90 por ciento de las ondas para entrenamiento, muestran que con el 80 por ciento, el error de reconocimiento es mayor. Los errores de reconocimiento son muy altos puesto que son superiores a 67.95 usando el 70%, de 76.39 usando 80% y de 69.64% al evaluar el corpus usando el 90% para entrenamiento, esto permite concluir que esta técnica no es apropiada para el reconocimiento de vocales orales y nasales teniendo en cuenta el corpus que se creó para este proyecto

.

Tabla de resultados con la comparación entre las técnicas: Dynamic Time Warping, Redes Neuronales y Modelos Ocultos De Markov

Resumen con los mejores resultados obtenidos de cada una de las técnicas

TÉCNICA	Vocales orales	Vocales nasales	Porcentaje de entrenamiento					
			70%		80%		90%	
			Reconocimiento		Reconocimiento		Reconocimiento	
			% Correcto	% Error	% Correcto	% Error	% Correcto	% Error
Dynamic Time Warping (DTW)	x		83.97	16.03	86.54	13.46	92.31	7.69
		x	67.31	32.69	69.23	30.77	65.38	34.62
	x	x	48.40	51.60	48.08	51.92	54.81	45.19
Redes Neuronales Artificiales (RNA)	x		92.31	7.69	93.52	6.48	94.64	5.36
		x	89.74	10.26	90.74	9.26	91.07	8.93
	x	x	90.06	9.94	86.57	13.43	86.61	13.39
Modelos Ocultos de Markov (HMM)	x		26.92	73.08	38.46	61.54	28.85	71.15
		x	62.82	37.18	30.77	69.23	48.08	51.92
	x	x	32.05	67.95	23.61	76.39	30.36	69.64

Tabla 12. Resumen con las técnicas DTW, RNA y HMM

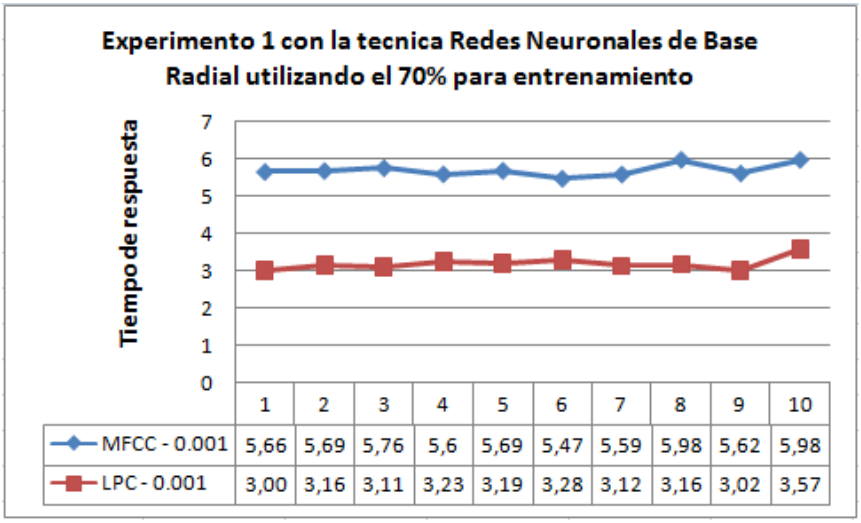
En la Tabla 12se resume las pruebas hechas a cada una de las técnicas seleccionadas para este proyecto, se observa que:

- Al comparar los porcentajes de error de reconocimiento de vocales orales simples por parte de la técnica RNA, éstos son inferiores a los errores de reconocimiento por parte de DTW y HMM.
- Los porcentajes de error de reconocimiento de vocales nasales simples de las técnicas DTW y HMM son superiores a los obtenidos con la técnica RNA.
- Cuando se combinan las vocales orales y nasales simples se mantiene la técnica RNA con los mejores porcentajes de reconocimiento, obteniendo hasta un 90.06% al utilizar el 70% del corpus para entrenamiento.

3.4. Resultados

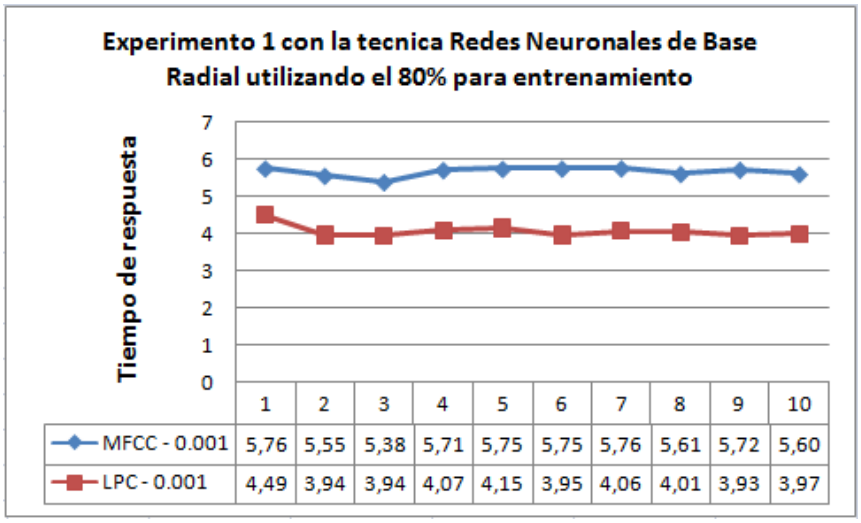
La técnica que mejor reconoce el subconjunto de vocales orales y nasales simples de la lengua Nasa Yuwe es Redes Neuronales Artificiales. La técnica demuestra que sin importar los diferentes porcentajes para entrenamiento y pruebas, esta técnica mantiene un reconocimiento superior a un 80%.La técnica DTW reconoció un promedio inferior al 50% con un número 130 ondas por cada clase y por último se encuentra la técnica de modelos ocultos de Markov con un promedio inferior al 30%. Por lo tanto la técnica a implementar para desarrollar el sistema de reconocimiento automático del habla de un subconjunto de vocales de la lengua Nasa Yuwe es Redes Neuronales de Base Radial.

Teniendo en cuenta que RNA de base radial es la técnica que mejores resultados de reconocimiento obtuvo, también se realizó pruebas donde se compara los tiempos de respuesta, de las técnicas de pre procesamiento MFCC y LPC utilizadas para este proyecto, con el fin de hacer la mejor elección en cuanto a menor tiempo de respuesta. Además se tiene en cuenta los mejores resultados de los experimentos, es decir, aquellos que aparecen marcados de color anaranjado. Para esto, se realizaron las siguientes pruebas de tiempo de respuesta, haciendo 10 tomas de tiempo y así determinar con cual técnica implementar el prototipo software.



	1	2	3	4	5	6	7	8	9	10
MFCC - 0.001	5,66	5,69	5,76	5,6	5,69	5,47	5,59	5,98	5,62	5,98
LPC - 0.001	3,00	3,16	3,11	3,23	3,19	3,28	3,12	3,16	3,02	3,57

Figura 30. Experimento 1 con la técnica redes de base radial utilizando el 70% para entrenamiento



	1	2	3	4	5	6	7	8	9	10
MFCC - 0.001	5,76	5,55	5,38	5,71	5,75	5,75	5,76	5,61	5,72	5,60
LPC - 0.001	4,49	3,94	3,94	4,07	4,15	3,95	4,06	4,01	3,93	3,97

Figura 31. Experimento 1 con la técnica redes de base radial utilizando el 80% para entrenamiento

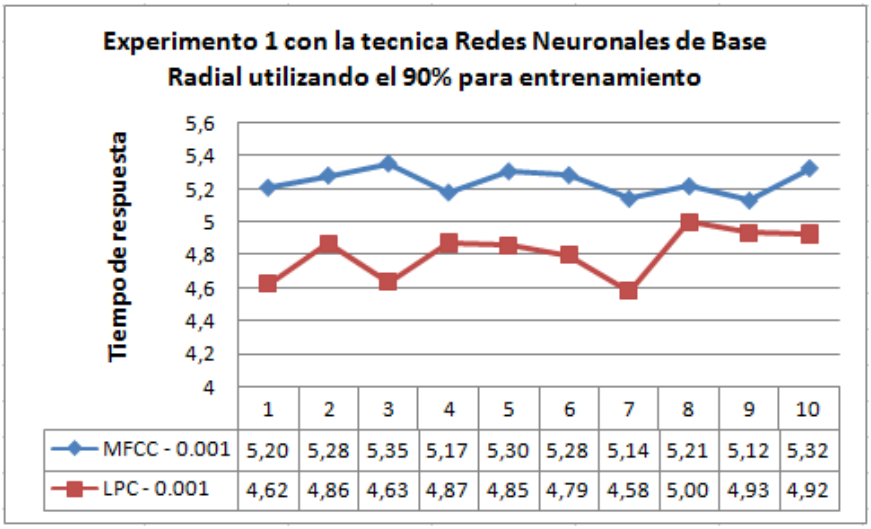


Figura 32. Experimento 1 con la técnica redes de base radial utilizando el 90% para entrenamiento

Las figuras 30, 31 y 32 nos indican claramente que el tiempo de respuesta al realizar las pruebas de reconocimiento de vocales orales simples utilizando diferentes porcentajes de entrenamiento, la técnica de extracción de características LPC es promedio 1.79 veces más rápido al tiempo utilizado por la técnica MFCC.

Experimento 2

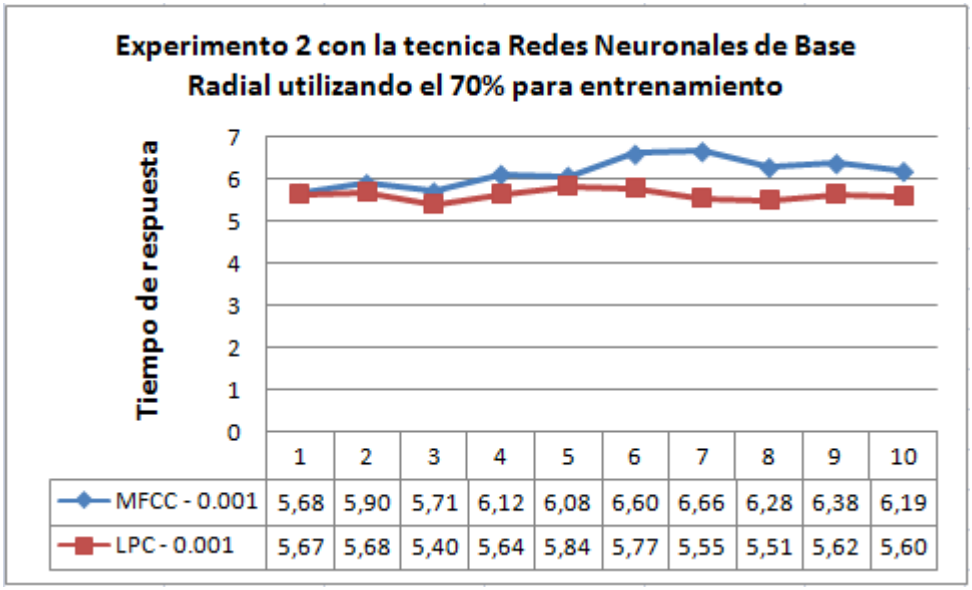
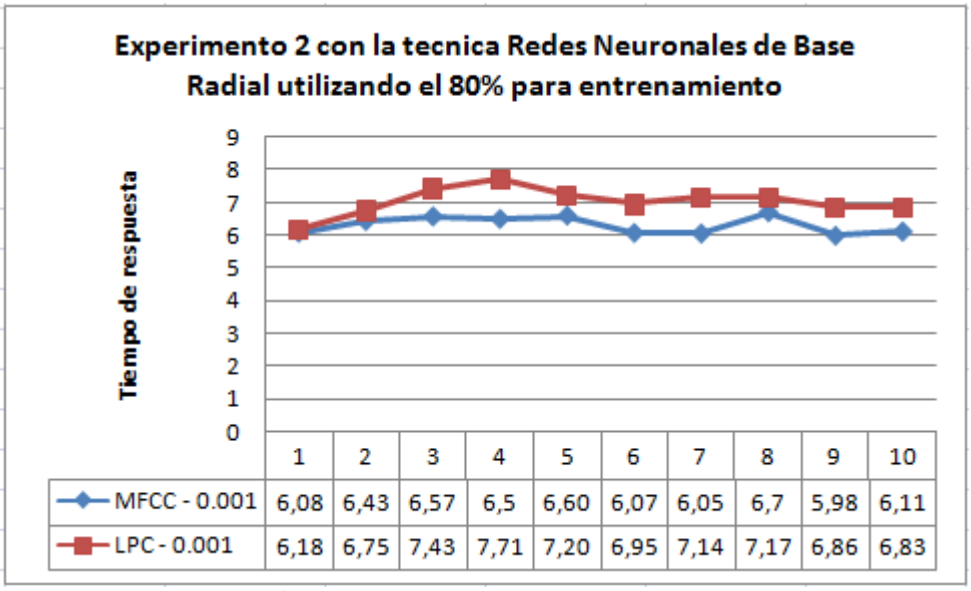
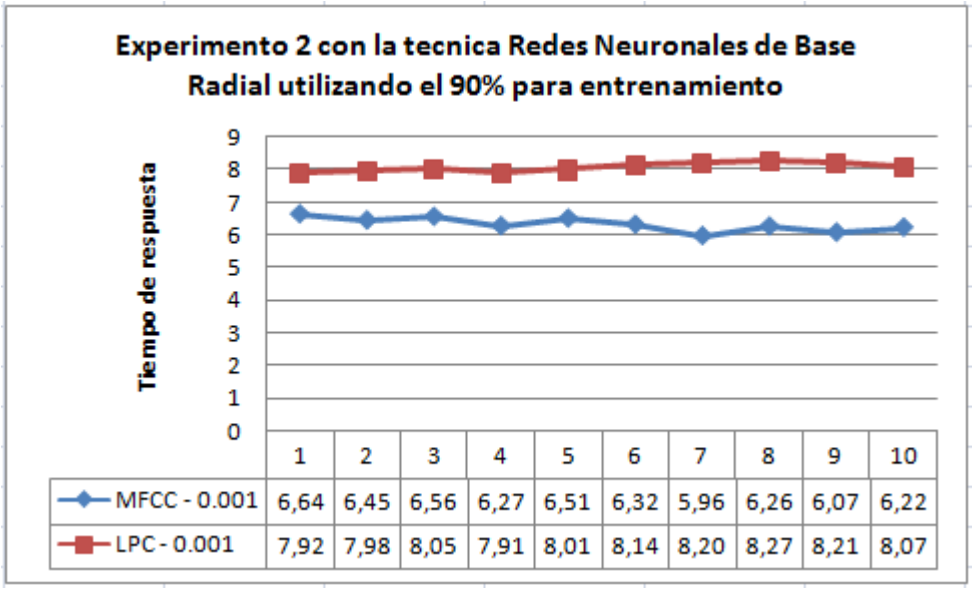


Figura 33. Experimento 2 con la técnica redes de base radial utilizando el 70% para entrenamiento



	1	2	3	4	5	6	7	8	9	10
MFCC - 0.001	6,08	6,43	6,57	6,5	6,60	6,07	6,05	6,7	5,98	6,11
LPC - 0.001	6,18	6,75	7,43	7,71	7,20	6,95	7,14	7,17	6,86	6,83

Figura 34. Experimento 2 con la técnica redes de base radial utilizando el 80% para entrenamiento

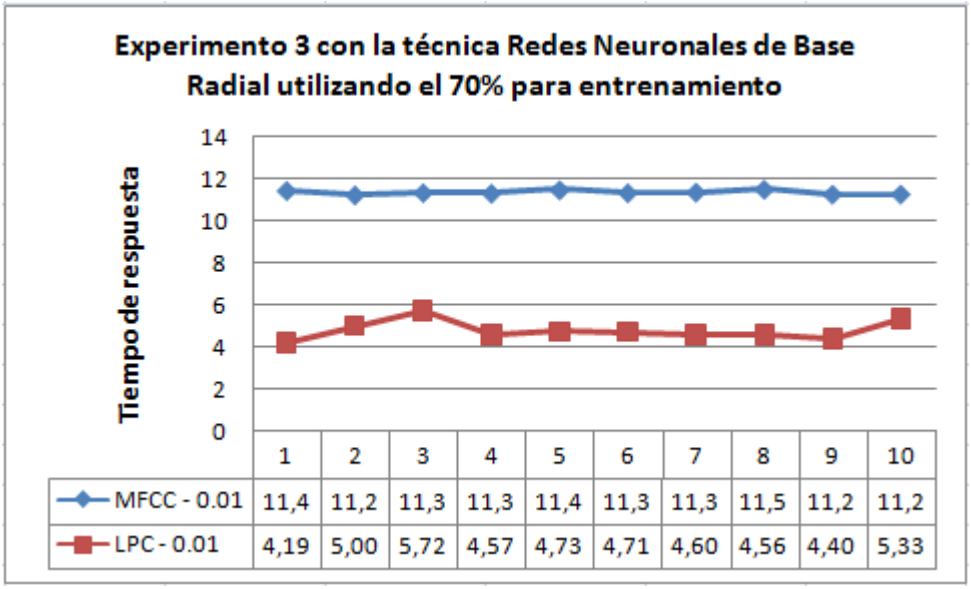


	1	2	3	4	5	6	7	8	9	10
MFCC - 0.001	6,64	6,45	6,56	6,27	6,51	6,32	5,96	6,26	6,07	6,22
LPC - 0.001	7,92	7,98	8,05	7,91	8,01	8,14	8,20	8,27	8,21	8,07

Figura 35. Experimento 2 con la técnica redes de base radial utilizando el 90% para entrenamiento

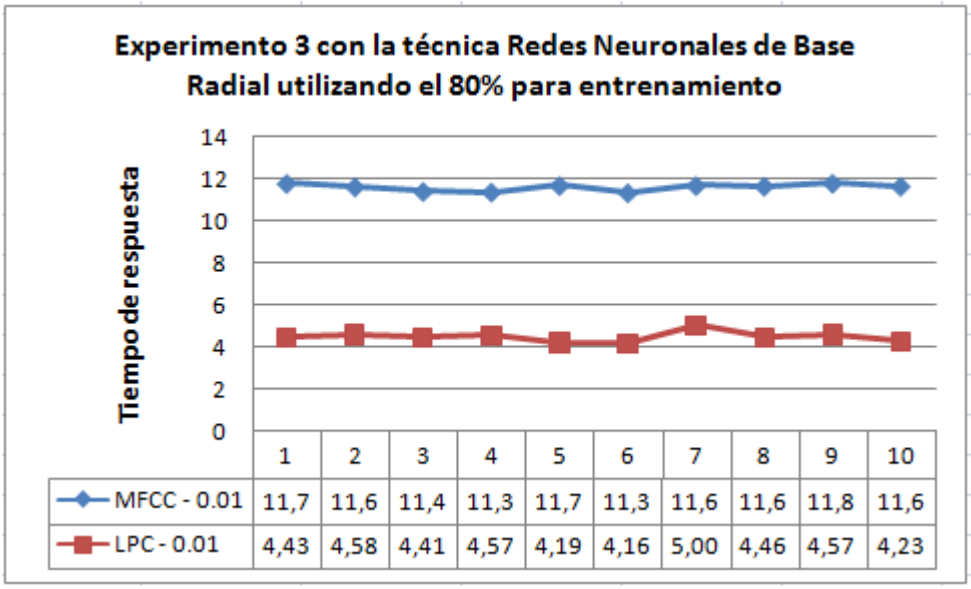
Las figuras 33, 34 y 35 nos muestran que el tiempo de respuesta al realizar las pruebas de reconocimiento de vocales nasales simples utilizando LPC es mayor al tiempo empleado por MFCC, sin embargo los tiempos de reconocimiento al utilizar el 70% para entrenamiento son muy similares entre LPC y MFCC, también se puede apreciar que MFCC es más rápido que LPC al reconocer solo vocales nasales simples utilizando un 80% de entrenamiento (ver figura 34).

Experimento 3



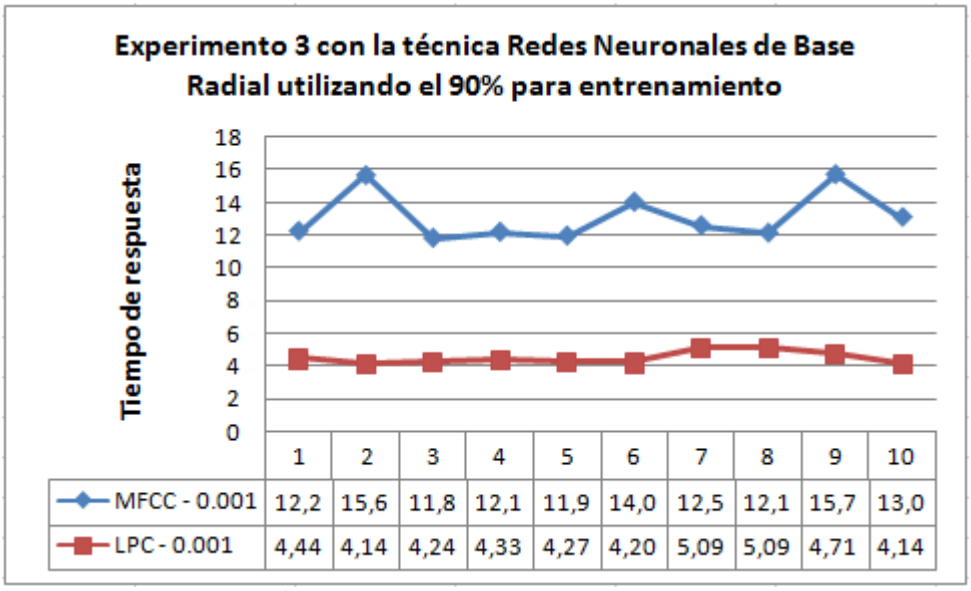
	1	2	3	4	5	6	7	8	9	10
MFCC - 0.01	11,4	11,2	11,3	11,3	11,4	11,3	11,3	11,5	11,2	11,2
LPC - 0.01	4,19	5,00	5,72	4,57	4,73	4,71	4,60	4,56	4,40	5,33

Figura 36. Experimento 3 con la técnica redes de base radial utilizando el 70% para entrenamiento



	1	2	3	4	5	6	7	8	9	10
MFCC - 0.01	11,7	11,6	11,4	11,3	11,7	11,3	11,6	11,6	11,8	11,6
LPC - 0.01	4,43	4,58	4,41	4,57	4,19	4,16	5,00	4,46	4,57	4,23

Figura 37. Experimento 3 con la técnica redes de base radial utilizando el 80% para entrenamiento



	1	2	3	4	5	6	7	8	9	10
MFCC - 0.001	12,2	15,6	11,8	12,1	11,9	14,0	12,5	12,1	15,7	13,0
LPC - 0.001	4,44	4,14	4,24	4,33	4,27	4,20	5,09	5,09	4,71	4,14

Figura 38. Experimento 3 con la técnica redes de base radial utilizando el 90% para entrenamiento

Entonces como conclusión se puede ver que las pruebas efectuadas teniendo en cuenta como medida de comparación el tiempo de respuesta, indican que la técnica LPC presenta los menores tiempos con el corpus conformado por vocales orales y nasales simples.

4. ADAPTACIÓN DEL MODELO Y CONSTRUCCIÓN DEL PROTOTIPO

Este capítulo contiene la descripción del proceso que se llevó a cabo para la construcción del prototipo software de reconocimiento del habla del subconjunto de vocales de la lengua Nasa Yuwe. Se consideró necesario hacer una precisión de los procesos pertinentes al prototipo de reconocimiento, por consiguiente, se presentan en las secciones siguientes la descripción de dichos procesos.

4.1. Modelado del sistema de reconocimiento del habla

A continuación se expone la fase de análisis y desarrollo del prototipo del sistema de reconocimiento del habla, dando cumplimiento con uno de los objetivos, para este se tiene en cuenta el ciclo de desarrollo de AUP [46]. En el proceso de desarrollo se tendrán en cuenta las 4 fases del ciclo de desarrollo como se presenta a continuación.

4.1.1. Inicio

El objetivo de esta fase es obtener una comprensión del alcance del prototipo y definir la arquitectura candidata. Para esto se llevo a cabo una recolección de requerimientos que se presentan a continuación.

Número de requisito	R 1
Nombre de requisito	Crear Modulo de Captura de Señal de voz
Fuente del requisito	Características de un Sistema de reconocimiento del habla
Prioridad del requisito	☒ Alta/Esencial ☐ Media/Deseado ☐ Baja/ Opcional
Descripción	El sistema debe permitir al usuario capturar la señal de voz tomada desde micrófono
Prerrequisito	

Tabla 13. Requisito Crear Modulo Captura de Señal de voz

Número de requisito	R 2
Nombre de requisito	Crear Modulo de Preénfasis
Fuente del requisito	Características de un Sistema de reconocimiento del habla
Prioridad del requisito	☒ Alta/Esencial ☐ Media/Deseado ☐ Baja/ Opcional

Descripción	El sistema debe permitir tomar las señales de voz grabadas y procesarlas con una técnica de pre procesamiento de señales como LPC o MFCC
Prerrequisito	Funcionalidad de los módulos anteriores.

Tabla 14. Requisito Crear Modulo de Preénfasis

Número de requisito	R 3
Nombre de requisito	Crear Modulo de Clasificación
Fuente del requisito	Características de un Sistema de reconocimiento del habla
Prioridad del requisito	☒ Alta/Esencial ☐ Media/Deseado ☐ Baja/ Opcional
Descripción	El sistema debe permitir tomar las características extraídas en el modulo de preénfasis para realizar un proceso de clasificación por medio técnicas de reconocimiento de patrones.
Prerrequisito	Vectores de características extraídos

Tabla 15. Requisito Crear Modulo de Clasificación

Número de requisito	R 4
Nombre de requisito	Crear interfaz grafica de ASR
Fuente del requisito	Características de un Sistema de reconocimiento del habla
Prioridad del requisito	☒ Alta/Esencial ☐ Media/Deseado ☐ Baja/ Opcional
Descripción	El sistema debe permitir al usuario seleccionar opciones de grabar, reproducir y verificar su pronunciación por medio de una interfaz donde podrá observar los resultados.
Prerrequisito	Módulos de captura de señal, conversión A/D, preénfasis y clasificación funcionando.

Tabla 16. Requisito Crear interfaz grafica de ASR

a) Identificación de Riesgos

Teniendo en cuenta el modelo dado por Pressman [47] se define el proceso de identificación de riesgos de la siguiente manera:

Riesgos de Requisitos

ID Requisito	Tipo de Riesgo	Riesgos
R1	De Personal	El personal no cuenta con los conocimientos requeridos para enfrentar la complejidad del requisito.
R2		Miembros del equipo no disponibles en momentos críticos, por calamidad domestica o enfermedad
R3	De estimación	El tiempo destinado para realizar el proceso de desarrollo del requisito está por debajo de lo necesario.
R4	De requisitos	Modificación de los requisitos que precisen modificaciones de diseño y rendimiento.

Tabla 17. Riesgos por Requisito

Tipo de Riesgo	Posibles Riesgos
Personal	Es difícil encontrar expertos en el área de reconocimiento del habla locales, para asesorías.
Organizativos	Posibles contratiempos por factores climáticos, orden público y tiempo de los hablantes y tesistas para reuniones y grabaciones de ondas.
Herramientas	Las herramientas de desarrollo son difíciles de configurar, su estudio y preparación tomo más tiempo del estimado. Las herramientas no son las adecuadas para realizar el desarrollo, se deben adecuar.
Requerimientos	Cambios de requerimientos que estipulen cambios en el diseño del prototipo.
Estimación	La complejidad del sistema esta subestimada.

Tabla 18. Riesgos por tipos

b) Plan de mitigación de riesgos

Para disminuir o sortear los distintos riesgos se tuvo en cuenta lo siguiente:

- *Investigación del estado del arte:* Para la falta de conocimientos requeridos por parte del personal que trabajó en este proyecto se hizo necesario realizar investigaciones de la bibliografía existente y relacionada con el reconocimiento automático del habla, de esto surgieron los antecedentes mencionados en esta monografía y también los tipos de sistemas de reconocimiento mencionados en el Anexo A, los cuales permitieron al personal entender el funcionamiento general de este tipo de sistemas. El tiempo destinado inicialmente a realizar estas investigaciones corresponde al primer trimestre de inicio de la ejecución del proyecto, sin embargo este tipo de investigaciones se realizó durante toda la duración del mismo, ya que muchas veces el lenguaje técnico usado en los distintos textos implicaba más lectura.

- *Reuniones constantes:* Con el fin de generar el corpus que es el componente básico para el prototipo de reconocimiento, se acudió a la colaboración del Grupo de estudios Lingüísticos para obtener información de las fuentes primarias acerca de la lengua Nasa Yuwe y además por medio de este grupo establecer el contacto y constante con los hablantes de la lengua Nasa Yuwe. El tiempo destinado a estas reuniones generalmente duraron un día, donde se tomó las muestras de audio necesarias para el posterior tratamiento de la señal.

- *Tiempo de desarrollo del prototipo:* Para evitar gastar demasiado tiempo de desarrollo del prototipo se utilizaron herramientas y códigos que ya tenían implementadas las técnicas a emplear en este proyecto. El tiempo destinado para este riesgo corresponde a la fase de desarrollo del proyecto, donde ya se tenía una base de conocimiento de los sistemas de reconocimiento. Como ya es bien conocida la ventaja y rapidez que tienen los lenguajes de Scripting y el entorno de trabajo de MATLAB se optó por trabajar con esta herramienta ya que es extensible mediante cajas de herramientas (toolbox), además cuenta con un entorno de desarrollo de interfaces que permitió un ágil desarrollo del prototipo.

- *Respaldo de expertos en la temática:* Para esto se contó con capacitaciones por parte de Roberto Perry quien es fonetista y Director del Departamento de fonética de la Universidad Nacional, igualmente se asistió a una capacitación en la Universidad Javeriana con sede en Cali, sobre herramientas de reconocimiento del habla y otra visita para un seminario de patrones. También fue necesario contar con la colaboración constante del director de tesis mediante reuniones.

c) Definición de requisitos funcionales y no funcionales

1) Esquema del sistema de Reconocimiento del Habla

- Funcionalidades del proyecto.
  - Creación del Módulo de Captura de la señal
  - Procesamiento de la señal análoga a digital

- Creación del Módulo de Extracción de Características de las señales del habla
- Almacenamiento de los vectores de características extraídos
- Creación del Módulo de Clasificación y Comparación de Patrones
- Creación de interfaz del sistema

2) Funcionalidades extras del sistema de reconocimiento del habla
   - Creación de ayudas para el sistema de reconocimiento del habla.

3) Requisitos no funcionales del prototipo de sistema de reconocimiento del habla

- Desempeño: el tiempo de respuesta debe ser menor a 4 segundos.

- Consistencia:
  - Los datos de audio deben ser consistentes.
  - Las características extraídas de las grabaciones deben corresponder con las entradas de nuevas grabaciones a clasificar.

d) Análisis y diseño

1) Modelo Conceptual

ACTOR	**DEFINICIÓN**
Usuario (Hablante)	Representa el concepto de la persona que interactúa con el Sistema de Reconocimiento del Habla para realizar acciones como grabar señal de audio, escuchar y verificar su pronunciación.

Tabla 19. Usuarios del Sistema

2) Vista lógica

Esta vista presenta la estructura conceptual de los componentes necesarios para dar solución a los requerimientos de la aplicación. La capa de presentación corresponde a una interfaz gráfica donde el usuario (Hablante) interactúa con la aplicación que está provista de controles como botones y vistas de los que ocurre al utiliza la aplicación.

3) Casos De Uso de Alto Nivel

A continuación se presenta el comportamiento del sistema de reconocimiento junto con el corpus de vocales de la lengua Nasa Yuwe y los actores externos con es en este caso el hablante.

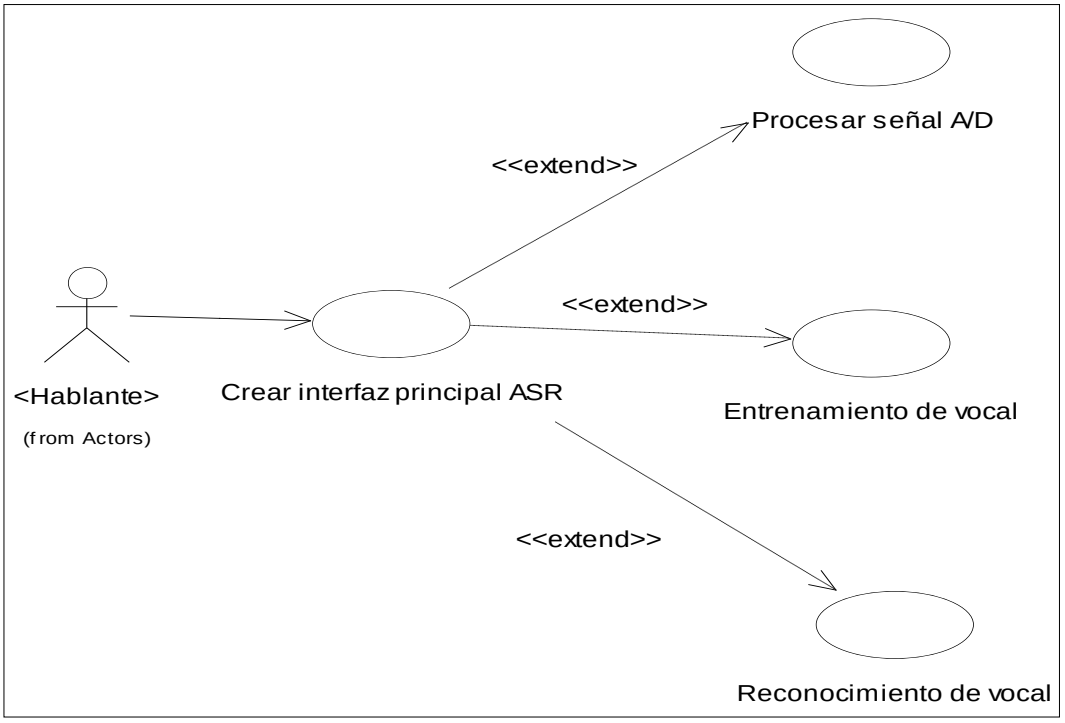


Figura 39. Vista global del sistema

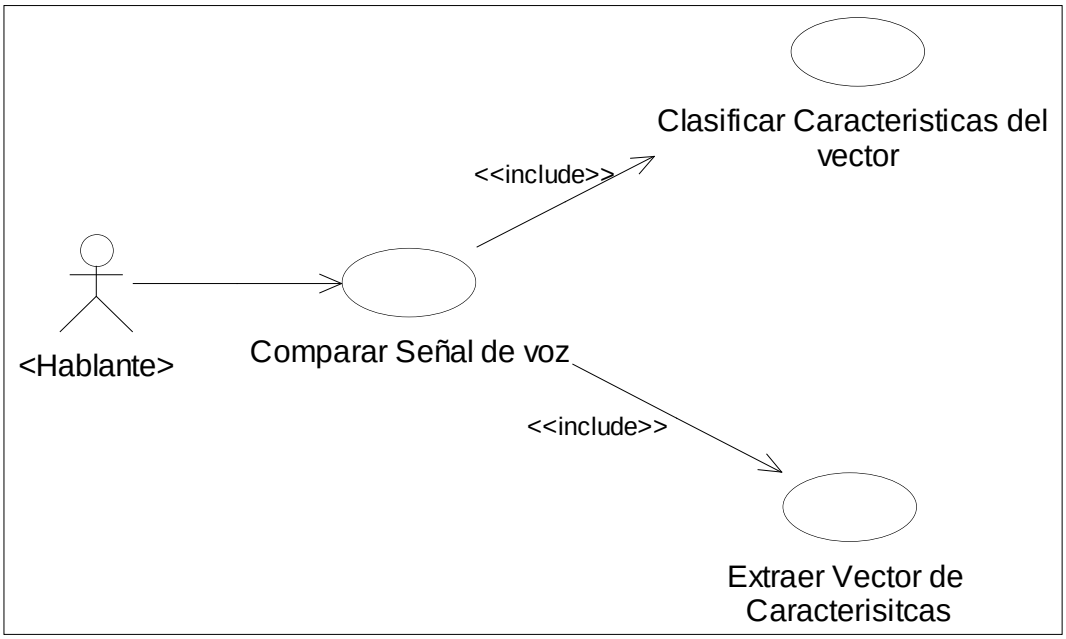


Figura 40. Caso de uso de entrenamiento del prototipo

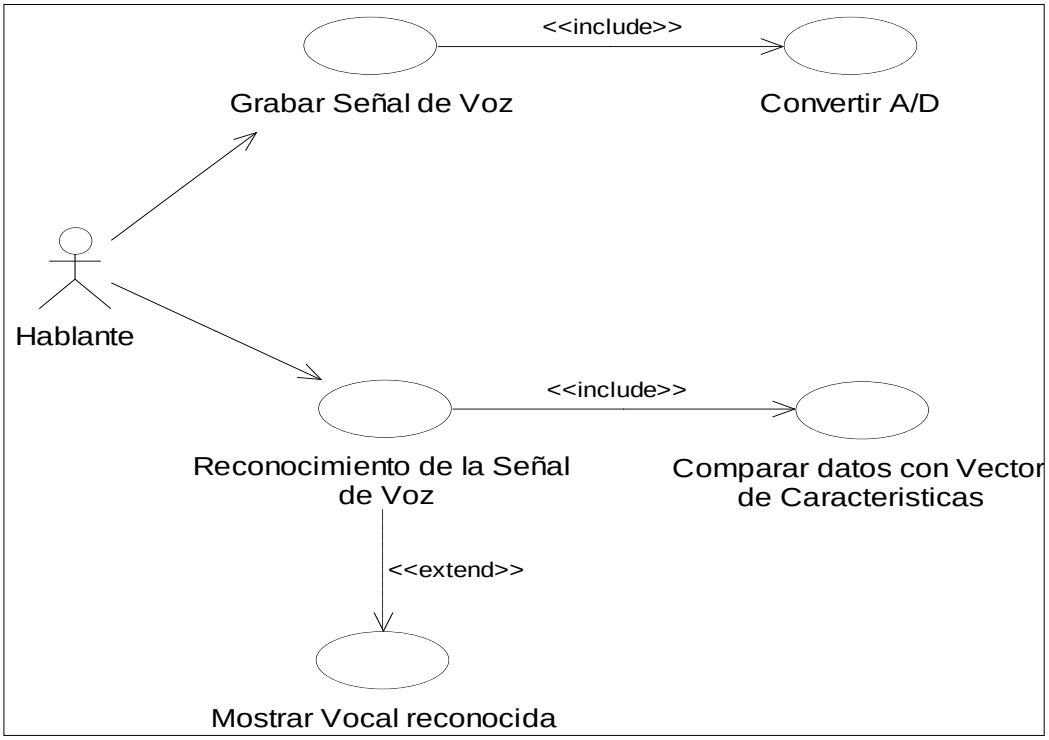


Figura 41. Caso de uso de pruebas del prototipo

4) Casos de uso reales

Esta sección contiene los casos de uso más críticos de la aplicación.

CASO DE USO REAL: Grabar Señal de Voz	
Actores: Hablante	
Propósito: Este Caso de Uso permite grabar la señal de voz desde un micrófono por parte de un hablante.	
Resumen: El Hablante realiza la grabación de voz presionando el botón de grabar donde pronunciara en un mínimo espacio de tiempo una vocal de la lengua Nasa Yuwe como por ejemplo: a, e, i, u.	
Prioridad: Alta	
	
CURSO NORMAL DE LOS EVENTOS	
Acción del actor	**Respuesta del sistema**
1. El usuario selecciona el botón grabar ubicado en la parte inferior izquierda [A].	**2.** El sistema permite capturar por un tiempo mínimo la señal de voz desde micrófono [B].

Tabla 20. Caso de Uso Real Grabar Señal de Voz

CASO DE USO REAL: Reconocimiento de la Señal de voz	
Actores: Hablante	
Propósito: Este Caso de Uso permite verificar la señal de voz recibida por el micrófono y emitida por parte de un hablante.	
Resumen: El Hablante realiza la verificación de la señal de voz, presionando el botón de verificar	
Prioridad: Alta	
	
CURSO NORMAL DE LOS EVENTOS	
Acción del actor	**Respuesta del sistema**
1. El usuario selecciona el botón verificar ubicado en la parte inferior derecha[A].	**2.** El sistema procede a cargar los patrones de las vocales previamente almacenados en el proceso de entrenamiento. Realiza una comparación de cada patrón con la señal de onda recibida y grabada (de acuerdo al anterior caso de uso real: Grabar señal de voz) y posteriormente despliega el tipo de vocal reconocida [B].

Tabla 21. Caso de Uso Real Reconocimiento de la Señal de voz

5) Diagrama de clases del prototipo de reconocimiento

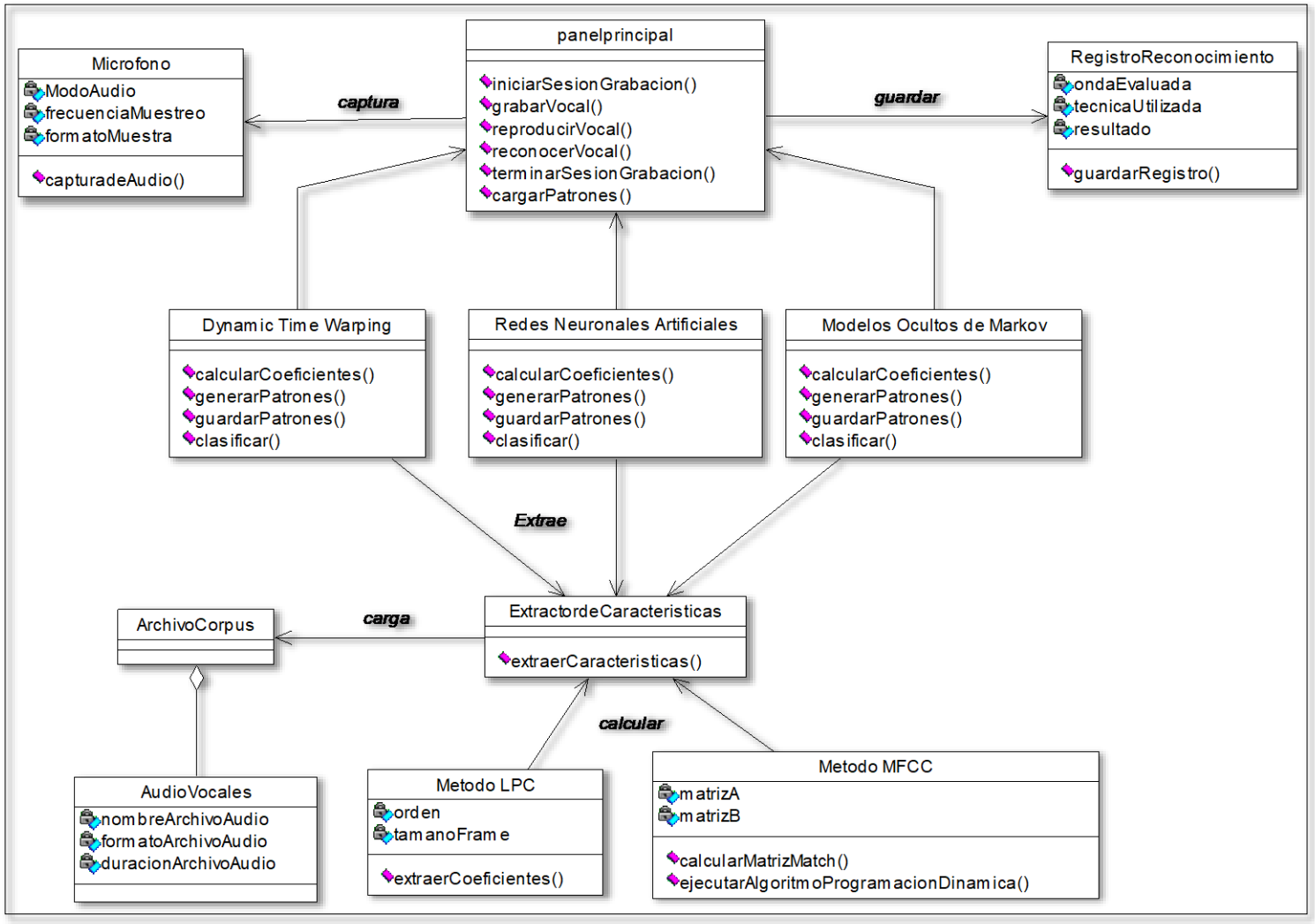


Figura 42. Diagrama de clases del prototipo de reconocimiento de la lengua Nasa Yuwe

Descripción de las clases del modelo de prototipo de reconocimiento

- **Dynamic Time Warping:** Clase de interfaz que se encarga de implementar la técnica Dynamic Time Warping. Entre sus operaciones se encuentra la de calcular los coeficientes (mediante MFCC), generar los patrones de cada vocal (matrices de coeficientes), guardar los patrones en disco, para que estos sean utilizados en el proceso de reconocimiento de la nueva pronunciación de la vocal emitida por el hablante.

- **Redes Neuronales Artificiales:** Clase de interfaz que se encarga de implementar la técnica de Redes Neuronales Artificiales. Entre sus operaciones se encuentra la de calcular los coeficientes (mediante LPC) generar los patrones de cada vocal (matrices de coeficientes) guardar los patrones en disco, para que estos sean utilizados en el proceso de reconocimiento de la nueva pronunciación de la vocal emitida por el hablante.

- **Modelos Ocultos de Markov:** Clase de interfaz que se encarga de implementar la técnica Modelos Ocultos de Markov. Entre sus operaciones se encuentra la de calcular los coeficientes (mediante LPC), generar los patrones de cada vocal (matrices de coeficientes), guardar los patrones en disco, para que estos sean utilizados en el proceso de reconocimiento de la nueva pronunciación de la vocal emitida por el hablante.

- **Micrófono:** Sirve para capturar la onda emitida por el hablante. Para este proyecto se configura el micrófono a una frecuencia de muestreo de 44100 Hz, tipo de archivo WAV y en modo MONO.

- **Método LPC:** Clase que implementa el método de extracción de características mediante Predicción Lineal.

- **Método MFCC:** Clase que implementa el método de extracción de características mediante los Coeficientes Cepstrales en las escala de la Frecuencia de Mel.

- **AudioVocales:** Clase que representa a las ondas que se encuentran almacenados en el corpus. Esta clase permite conocer las características de dichas ondas, como por ejemplo, el formato en este caso WAV.

- **RegistroReconocimiento:** Clase encargada de almacenar en disco la onda evaluada *(ondaEvaluada)* la técnica utilizada *(tecnicaUtilizada)* para el reconocimiento y el resultado*(resultado)* o vocal que el prototipo reconoció de la onda pronunciada por el hablante.

- **panelprincipal:** Esta clase de interfaz sirve para controlar el prototipo del sistema de reconocimiento. Será la encargada de controlar el inicio de sesión, cargar los patrones generados en el proceso de entrenamiento, grabación de la onda a evaluar, reproducir la señal de la vocal emitida por el hablante, luego con estos pasos previos enviará los mensajes a las clases encargadas de la clasificación (las clases relacionadas con las técnicas) mostrar los resultados o tipo de vocal reconocida y finalizar la sesión de grabación de un hablante determinado.

6) Diagramas de secuencia del entrenamiento y pruebas del prototipo de reconocimiento

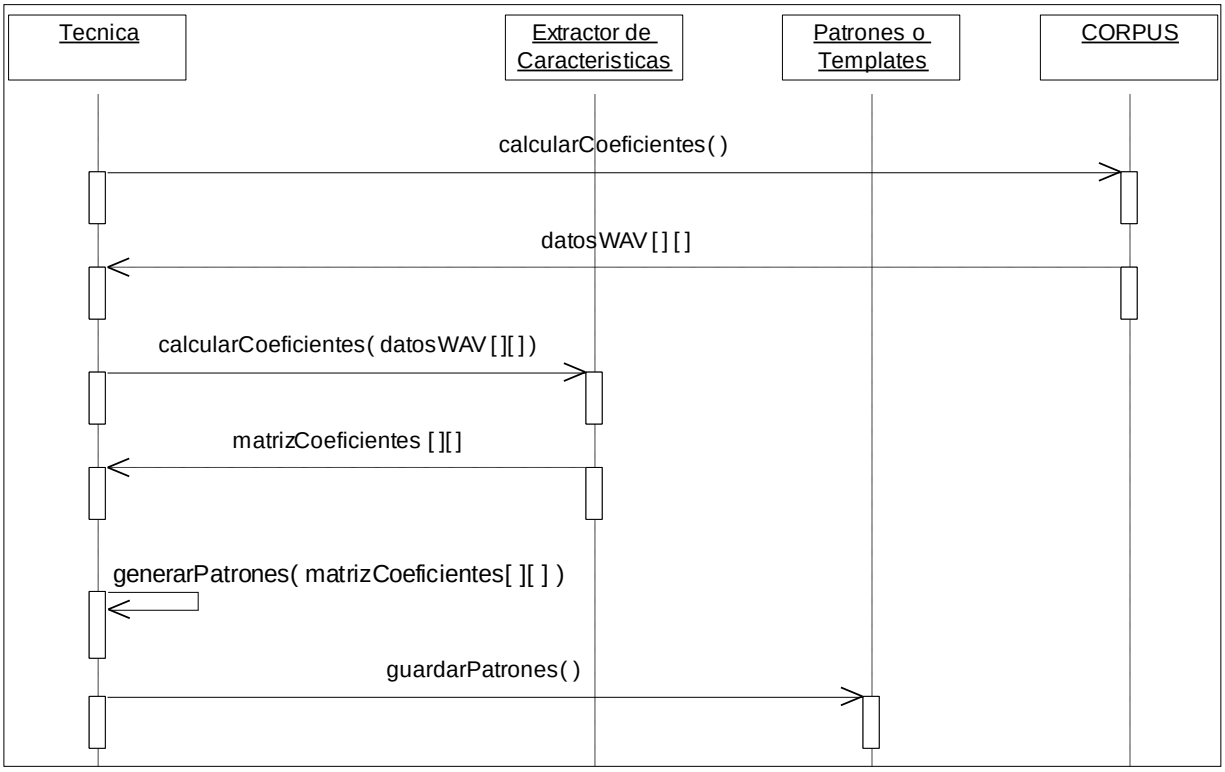


Figura 43. Diagrama de secuencia del entrenamiento del prototipo de reconocimiento

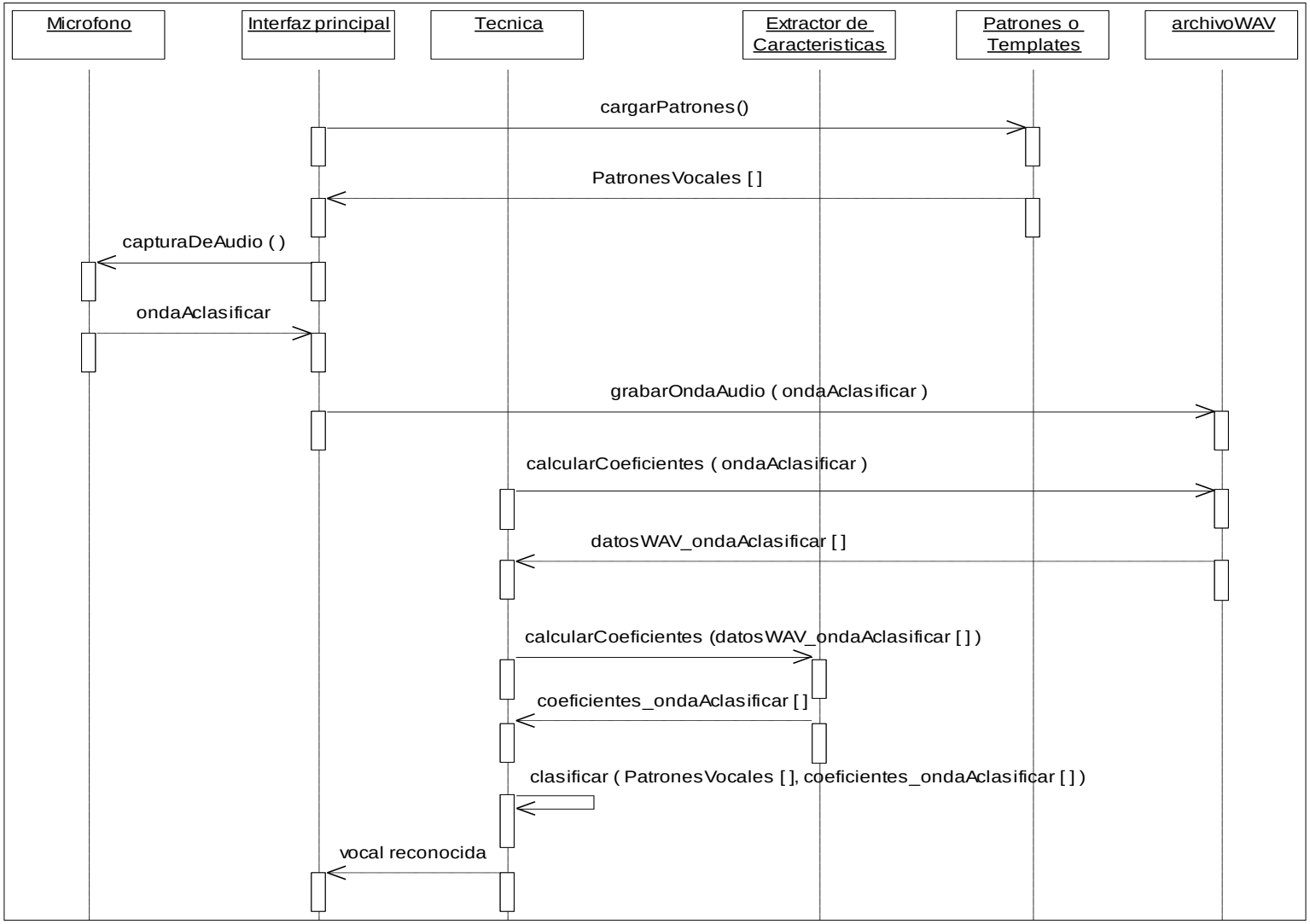


Figura 44. Diagrama de secuencia de pruebas del prototipo de reconocimiento

8) Diagramas de flujo general del prototipo de reconocimiento

Los pasos a seguir por las diferentes técnicas trabajadas en este proyecto son similares, por lo tanto para efectos de muestra del flujo de actividades para el reconocimiento de las vocales orales y nasales simples, se presenta a continuación los diagramas de flujo de los pasos fundamentales:

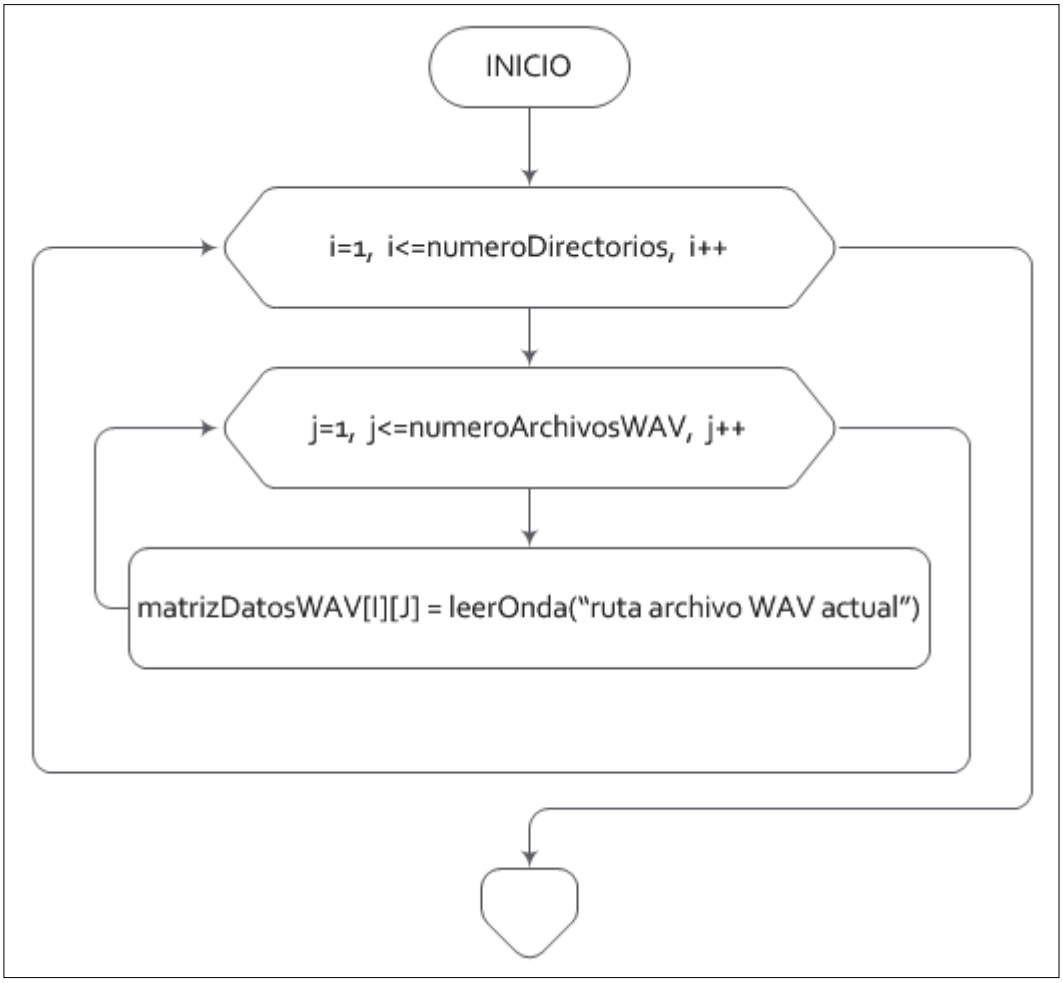


Figura 45. Diagrama de flujo para cargar datos

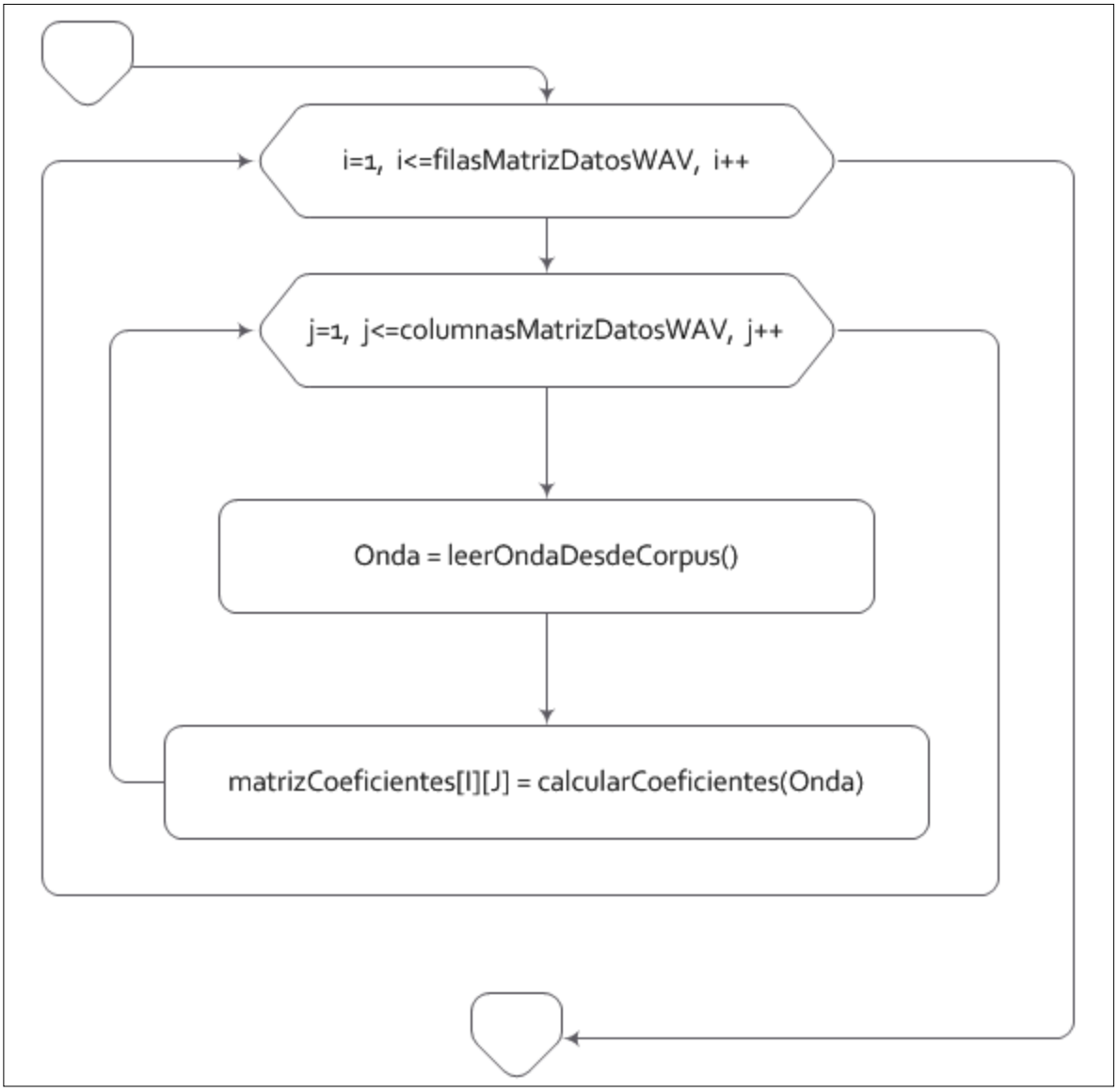


Figura 46. Diagrama de flujo para calcular las características de la vocal

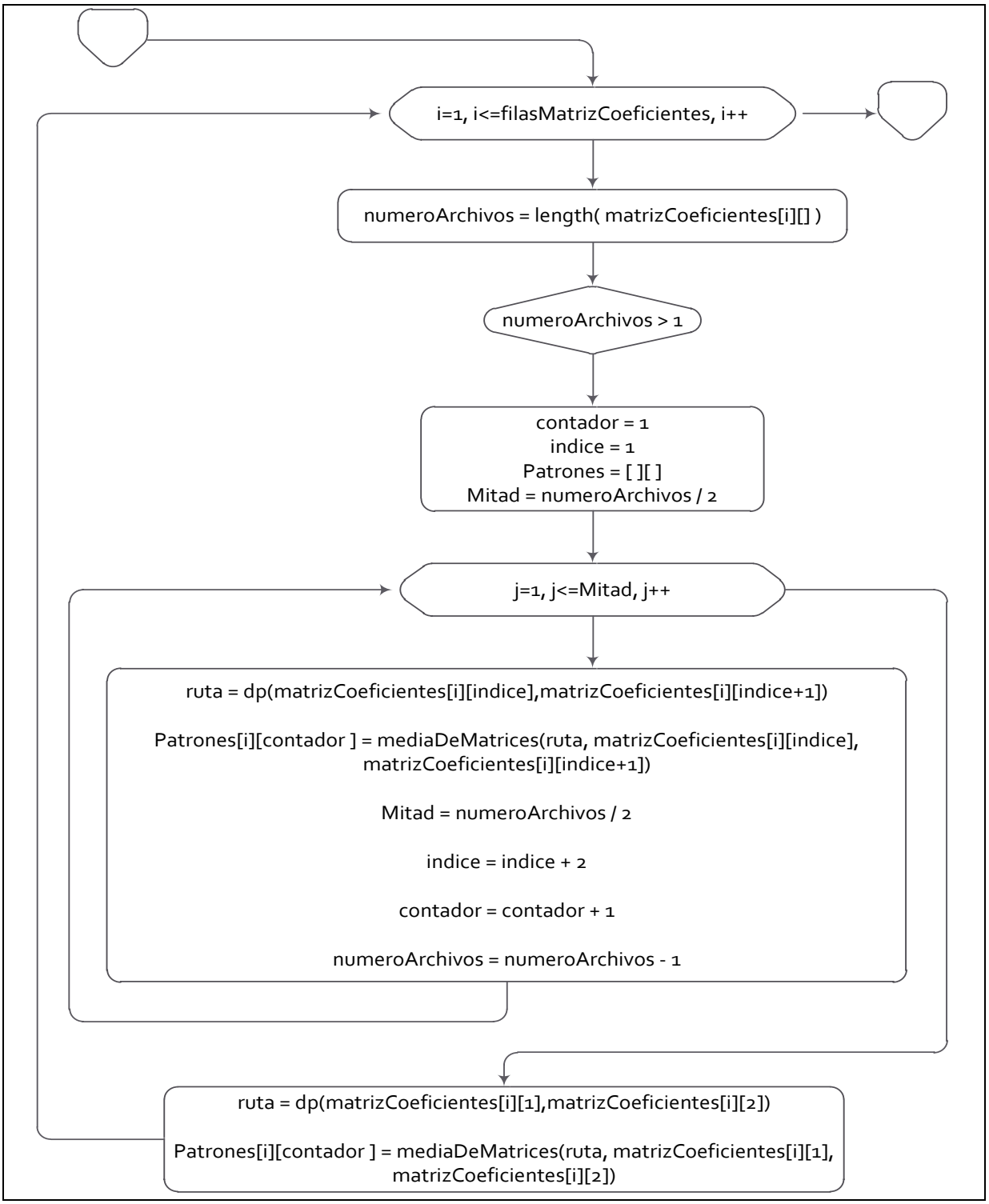


Figura 47. Diagrama de flujo para calcular los patrones

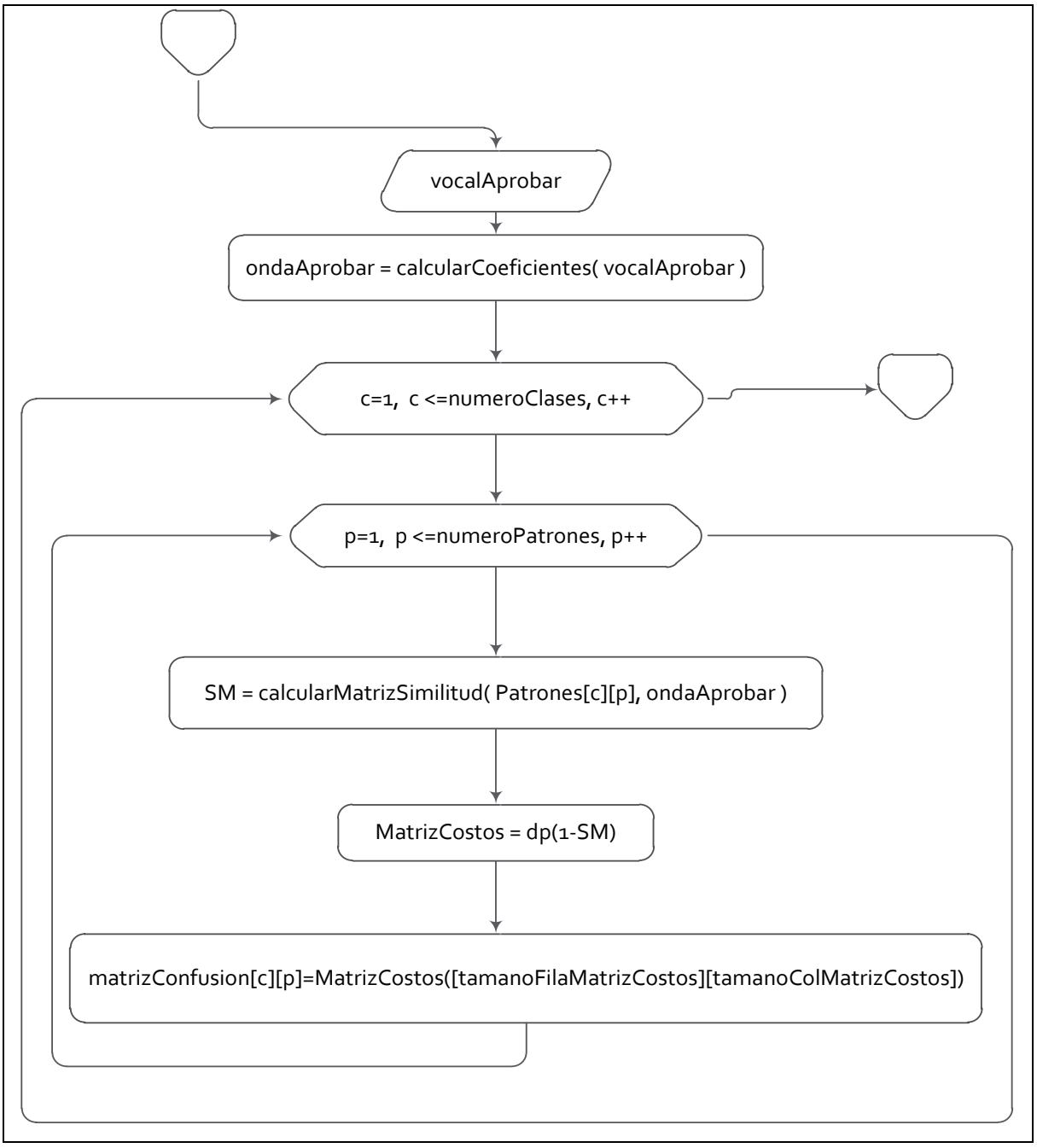


Figura 48. Diagrama de flujo de la clasificación de las ondas de prueba

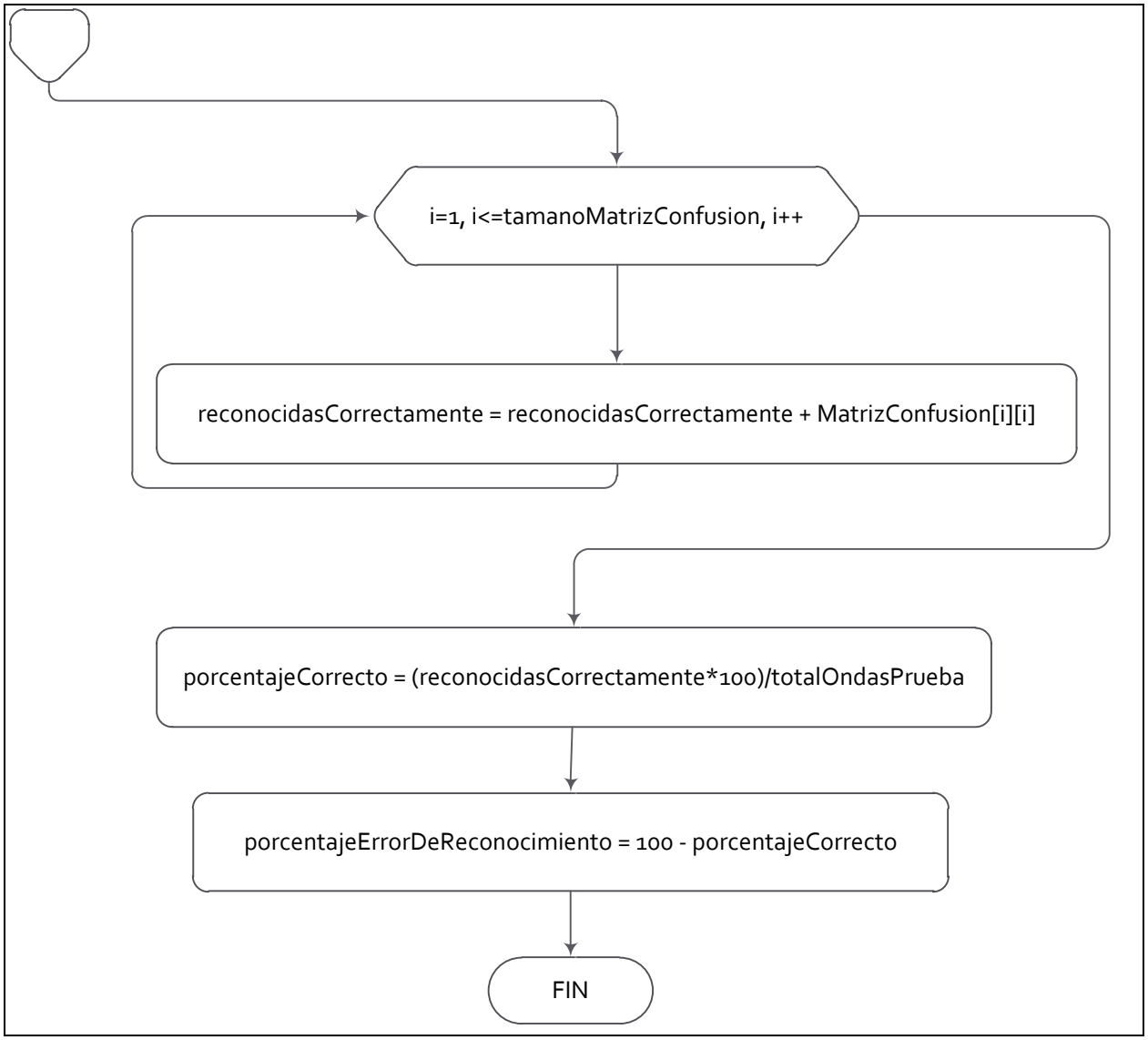


Figura 49. Diagrama de flujo para calcular el error de reconocimiento

9) Adaptación de la arquitectura de filtros y tuberías para el prototipo de reconocimiento

La estructura general del software a construir tiene un conjunto de entradas que se van transformando al tiempo que pasan por un conjunto de módulos que actúan como filtros, estos se ejecutan completamente antes de pasar al próximo, el flujo de datos procesado en un filtro se convierte en la entrada del siguiente, este flujo pasa a otro modulo por medio de una tubería [48], a continuación se observa el conjunto de módulos y el flujo de datos.

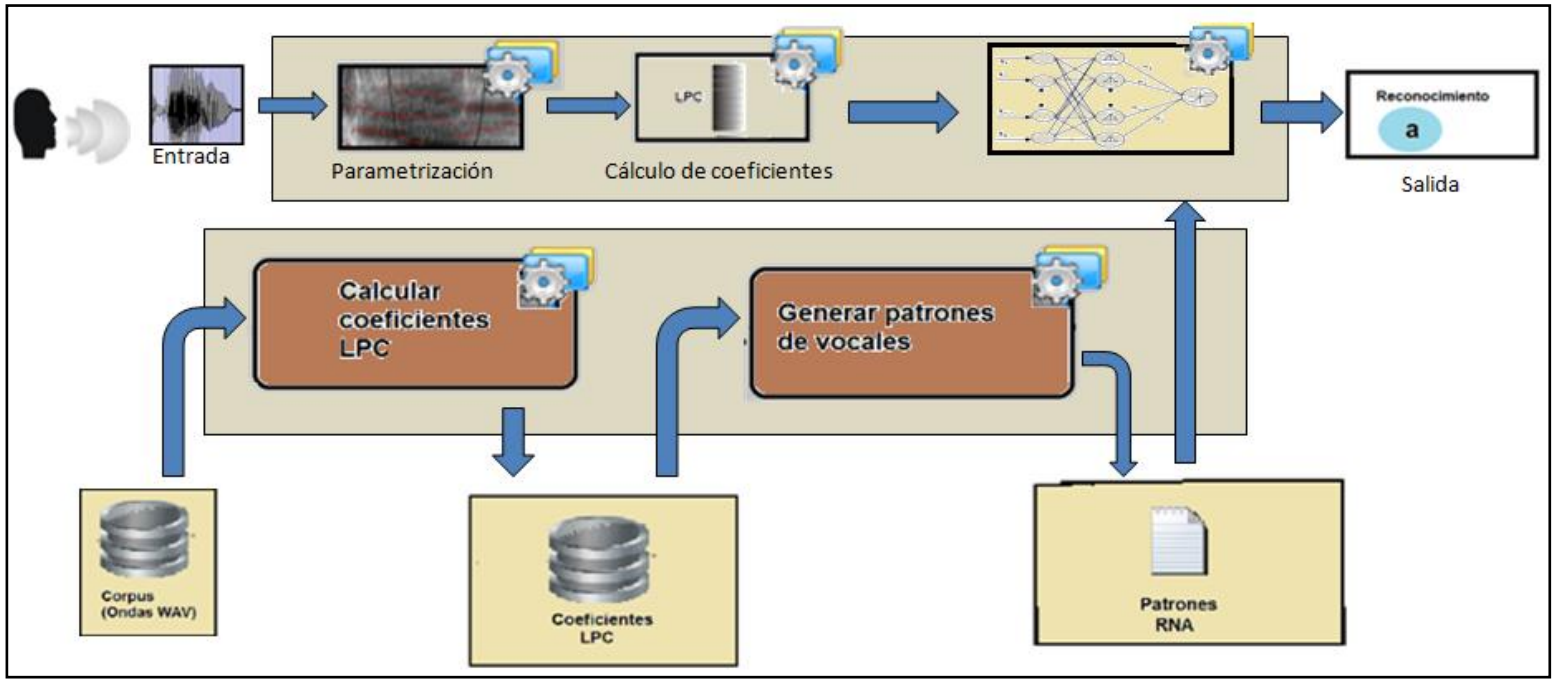


Figura 50. Adaptación de la arquitectura de filtros y tuberías para el prototipo de reconocimiento

4.1.2. Elaboración

La construcción del prototipo de reconocimiento del subconjunto de vocales orales y nasales simples de la lengua Nasa Yuwe consideró herramientas que fuesen de licencia libre, sin embargo por la complejidad en el manejo de las mismas se optó por la herramienta MATLAB cuya licencia es propietaria. Así que se recurrió a una licencia que hubiese sido adquirida por la Universidad del Cauca de uso académico para efectos de este proyecto. Por ello a continuación se hablará de esta herramienta:

MATLAB [49] (abreviatura de MATrix LABoratory, "laboratorio de matrices") es un lenguaje de computación técnica de alto nivel y un entorno interactivo para desarrollo de algoritmos, visualización de datos, análisis de datos y cálculo numérico. Con MATLAB, es posible resolver problemas de cálculo técnico más rápidamente que con lenguajes de programación tradicionales, tales como C, C++ y FORTRAN.

MATLAB es usado para aplicaciones que incluyen procesamiento de señales e imágenes, comunicaciones, diseño de sistemas de control, sistemas de prueba y medición, modelado y análisis financiero y biología computacional. Los conjuntos de herramientas complementarios (colecciones de funciones de MATLAB para propósitos especiales, que están disponibles por separado como *toolboxes*) que amplían el entorno de MATLAB permitiendo resolver problemas especiales en estas áreas de aplicación.

Las cajas de herramientas o *toolboxes* y código fuentes usadas en este proyecto se describen a continuación:

a) Dynamic Time Warping [50]

Este código es una implementación de esta técnica y además muestra ejemplo de uso de ésta, comparando dos ondas o grabaciones de la misma palabra, los códigos utilizados en este proyecto para llevar a cabo esta técnica se exponen a continuación:

- **simmx.m** Es una utilidad para calcular la matriz de similitudes, puede ser mediante el cálculo de distancias entre cada par de fotogramas.

- **dp.m** Es la implementación de un algoritmo simple de programación dinámica, que permite tres pasos con pesos iguales.

b) Neural Network Toolbox [51]

Conjunto de códigos para trabajar con los distintos tipos de redes neuronales (como por ejemplo: perceptron, backpropagation y base radial entre otras). A continuación se nombran algunas de las características que soporta este Toolbox:

- **Funciones de creación de redes:** Para este proyecto se crearon redes de base radial mediante ***newpnn,*** el cual crea una red de dos capas, las cuales calculan los pesos de entrada y sus entradas de red.

- **Funciones de entrenamiento:** simulan el modelo previamente establecido, por ejemplo, mediante ***sim.***

- **Funciones de transferencia:** para este proyecto se usó una función de transferencia de base radial, esto se lleva a cabo mediante ***radbas.***

Este Toolbox es mucho más amplio pero para efectos de este proyecto sólo se nombraron las anteriores para dar un esbozo de lo que puede hacerse con el mismo.

c) Modelos Ocultos de Markov

Para uso de esta técnica se utilizaron códigos para entrenamiento y pruebas,a continuación se mencionarán los de mayor relevancia:

- **hmm.m** [52]**:** Este código permite realizar el proceso de entrenamiento del modelo, teniendo en cuenta que para la obtención de los coeficientes se hizo con LPC, se configuró el numero de ciclos de Baum – Welch a usar, el número de estados, la matriz de coeficientes y su longitud. El resultado de este proceso es la matriz de promedios, matriz de covarianza, matriz de estados de transición, los estados iniciales y la curva de probabilidad logarítmica.
- **hmm_cl.m** [53]**:** Este código permite realizar el proceso de pruebas o el cálculo de reconocimiento; es decir, la comparación entre la onda(s) de prueba y los patrones generados en el proceso de entrenamiento. Como resultado obtendremos la probabilidad de la matriz de datos o coeficientes y el arreglo de las probabilidades de cada una de las secuencias.

4.1.3. Construcción

Finalmente el prototipo de software está distribuido en directorios de la siguiente forma:

Prototipo de reconocimiento Nasa Yuwe

- **Corpus:** En este se almacenan las grabaciones u ondas de las vocales, dividido en 8 directorios. Los primeros 4 directorios contienen las vocales nasales simples y los últimos 4 directorios las vocales orales simples.

- **Grabaciones:** Destinadas a almacenar cada una de las sesiones de pruebas hechas con cada hablante. Cada sesión es un directorio que se crea y guarda tanto las ondas a evaluar como un archivo que registra: nombre de la grabación, técnica utilizada para reconocimiento y el resultado o tipo de vocal reconocida.

- **Imágenes:** Contiene imágenes de íconos de los botones de la interfaz gráfica.

- **Patrones:** Guarda los archivos generados por MATLAB donde se encuentran los patrones generados por cada una de las técnicas durante el proceso de entrenamiento.

- **Técnicas:**

    - **Dynamic Time Warping:** Código fuente de esta técnica.

    - **Modelos Ocultos de Markov:** Código fuente de esta técnica.

    - **Redes Neuronales Base Radial:** Código fuente de esta técnica.

- **Utilidades:** Códigos de aquellas tareas que hacen posible el funcionamiento del prototipo, como por ejemplo, la creación del archivo de la sesión de grabación.

Para la creación del archivo ejecutable del prototipo software, se utiliza la siguiente sintaxis: ***mcc [-options] fun [fun2...].*** Para este proyecto se cuenta con una interfaz gráfica generada con la herramienta ***guide*** de MATLAB, el cual genera dos archivos, uno con extensión **.m** y otro con extensión **.fig**. Esta interfaz es la que controla el resto de funcionalidades del prototipo, y por ello para la creación del ejecutable se hizo el siguiente llamado:

mcc–m *<nombre_programa_principal.m><nombre_programa_principal.fig>*

Finalmente se obtiene el prototipo de reconocimiento como se muestra en la Figura 52:

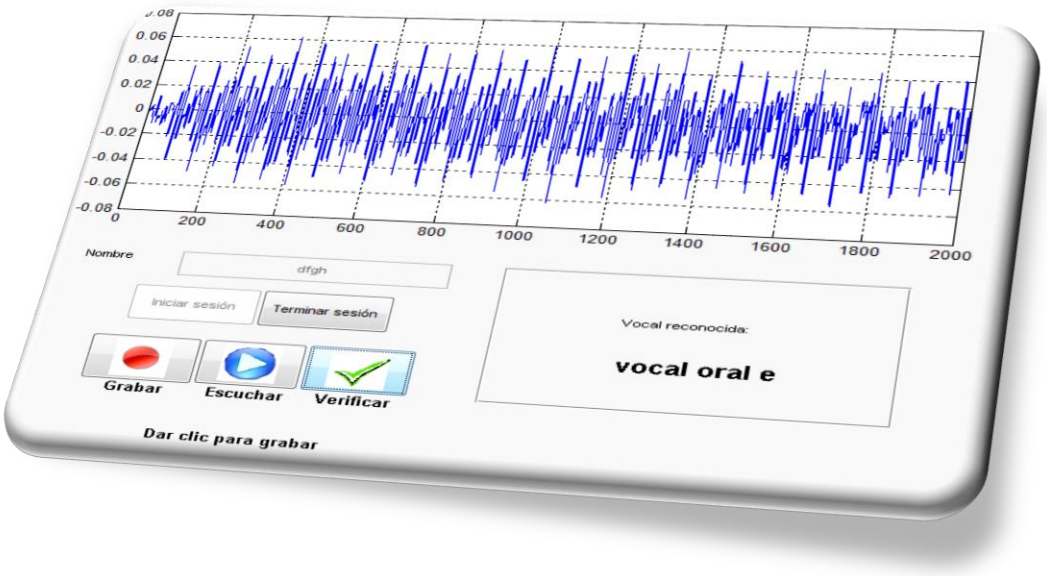
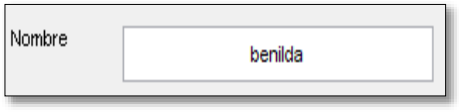
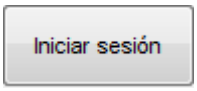


Figura 51. Interfaz del prototipo del sistema de reconocimiento

Descripción de los pasos para usar la interfaz del prototipo de reconocimiento:



Paso 1: Espacio para escribir el nombre de la persona de quién se va a tomar las muestras a evaluar por parte del prototipo.

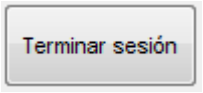


Paso 2: ***"Iniciar sesión"*** lo que implica crear un archivo de texto, sobre el cual se va a almacenar la técnica usada y lo que el prototipo reconoció.

Paso 3: Dar clic sobre este botón para grabar, y cuando en la interfaz muestre el mensaje ***"Grabando…"*** se debe pronunciar la vocal

Paso 4: Comprobar si se ha grabado la vocal pronunciada o solamente se grabó ruido, si se presenta el primer caso podemos seguir con el siguiente paso, de lo contrario volver al **Paso 3**

Paso 5: Una vez comprobada la vocal grabada, se da clic sobre este botón para verificar a cuál vocal pertenece la onda grabada en el **Paso 3.**



Paso 6: Finalmente se da clic sobre el botón ***"Terminar sesión"*** y de esta forma queda listo para iniciar la grabación de un nuevo hablante.

4.1.4. Transición

a) Diseño de la prueba

El diseño de la prueba (ver Anexo D) consta de una serie de repeticiones de forma sencilla en las cuales el usuario repite vocales orales y nasales de la lengua Nasa Yuwe de forma aleatoria, teniendo en cuenta que se repita la misma vocal el mismo número de veces, aunque el funcionamiento del prototipo es muy sencillo se realiza una explicación a los usuarios para que se familiaricen con la interfaz de usuario.Las pronunciaciones se realizaron como se observa en la Tabla 22, por medio de 5 pruebas donde las vocales se repitieron en forma aleatoria.

Pruebas	1	2	3	4
Prueba 1	a	e	i	u
Prueba 2	ã	ẽ	ĩ	ũ
Prueba 3	a	ẽ	i	ũ
Prueba 4	ã	ĩ	ũ	e
Prueba 5	a	ĩ	ẽ	u

Tabla 22. Diseño de pruebas en vivo para el prototipo de reconocimiento

El proceso de las pruebas al prototipo software de reconocimiento de habla del subconjunto de vocales orales y nasales simples de la lengua Nasa Yuwe se realizó teniendo en cuenta las siguientes características:

- Funcionalidad
- Interfaz de usuario
- Facilidad de Uso

Para poder verificar cada uno de las características se efectuaron pruebas de unidad, pruebas de usabilidad y pruebas de integración, su planteamiento y desarrollo se pueden apreciar en el Anexo E.

b) Conclusiones de las pruebas del software

- Debido a que el prototipo reconocimiento consta de pocas funcionalidades (grabar, reproducir y verificar) el riesgo a cometer errores es bajo, ya que la única entrada de datos por parte del usuario es cuando se va a iniciar sesión y esta entrada es un tipo primitivo como lo es la cadena de caracteres, donde su única razón de ser es servir de identificador de hablante. Además esta entrada es validada por el sistema para que no se repita el nombre de sesión.

- Las pruebas permitieron verificar que el mensaje de respuesta del prototipo a entradas no validas es breve, entendible y coherente.

- Las pruebas de usabilidad permitieron verificar que el sistema tiene una velocidad de respuesta adecuada al verificar la onda de la vocal de prueba.

- Las pruebas de usabilidad mostraron que algunas funcionalidades del sistema son un poco complicadas de aprender a manejar para algunos de los participantes, debido a la falta de experiencia, esto puede ser consecuencia de que ellos no están familiarizados con los íconos habituales para representar las acciones de grabar, reproducir y verificar.

- Al aplicar las pruebas de usabilidad se encontró que se hace necesario plantear ayudas, como es el caso de videos o manuales de usuario impreso,para el manejo de las funcionalidades del prototipo de reconocimiento. Para dar solución a este inconveniente se realizó un manual de usuario (ver Anexo F).

- Las pruebas de usabilidad aplicadas mostraron que los participantes estuvieron conformes con la presentación y manejo del sistema, las cuales cumplieron sus expectativas, a pesar de que en la prueba se tuvieron algunas dudas.

- Al aplicar las pruebas de usabilidad se notó que se requiere mayor tiempo de práctica con el sistema para mejorar los niveles de familiaridad de uso de las diferentes funcionalidades.

- Luego de haber realizado las pruebas de integración a cada una de las técnicas prototipo de reconocimiento, se ve que las diferentes técnicas del prototipo pasaron las pruebas realizadas; aunque se nota una disminución en el reconocimiento por parte de las tres técnicas y especialmente con la técnica de redes neuronales.

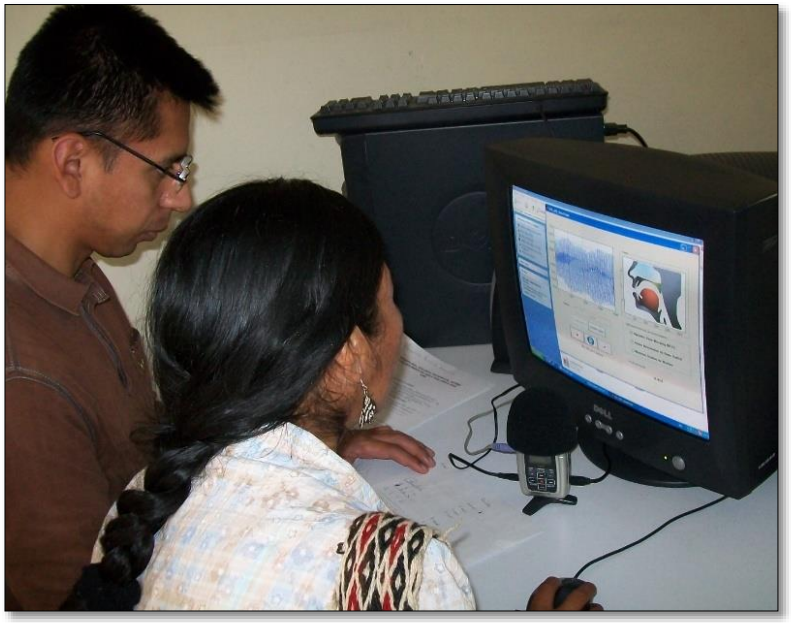


Figura 52. Aplicación de la Prueba (1)

Figura 53. Aplicación de la Prueba (2)

5. DIFICULTADES PRESENTADAS Y SOLUCIONES PLANTEADAS DURANTE EL DESARROLLO DEL PROYECTO

- El primer inconveniente para el desarrollo del proyecto fue la ubicación geográfica de la población objetivo, por ser una comunidad indígena en busca de preservar sus costumbres y autonomía, los nasa se encuentran ubicados en territorio rural del municipio de caloto, a 3 horas de la ciudad de Popayán, donde reside y labora el equipo de desarrollo del proyecto.

  Para abordar este inconveniente se optó por citarlos a reuniones con los miembros del proyecto, con el fin realizar las grabaciones de las palabras y realizar las pruebas del prototipo, siendo informados previamente de la agenda a realizarse.

- En el proceso de desarrollo se trabajaron con diferentes herramientas de desarrollo de software tanto en el ámbito del software libre (Sphinx 3, Java) como en herramientas propietarias (Matlab). Pero no se contaba con conocimientos previos de uso y configuración por parte de los miembros de desarrollo del proyecto, esto dificultó el proceso por la larga curva de aprendizaje. Se utilizaron herramientas que implementan las técnicas de reconocimiento antes nombradas, además de la utilización de un lenguaje de scripting que permitió un desarrollo ágil y rápido del prototipo de reconocimiento.

- El proceso de grabación de señales de audio de las vocales escogidas de la lengua Nasa Yuwe se realizaron en un ambiente de ruido no controlado, en diferentes formatos y con configuraciones diferentes de muestreo, problema que se resolvió usando Praat y Audacity.

  Para este caso se tomaron nuevas grabaciones con una configuración previamente estudiada y que se adecuara para los módulos de procesamiento de señal que se implementaron, además se contaron con micrófonos especiales para realizar este tipo de grabaciones (grabadora con micrófono configurable a escucha de 90 grados, con supresor de ruido).

6. CONCLUSIONES Y RECOMENDACIONES

6.1. CONCLUSIONES

- Si se observa el modelo computacional de un sistema de reconocimiento del habla, se ve que los módulos de éste no son independientes, esto implica que el flujo de datos del proceso que empieza desde la entrada de la señal, conversión de análogo a digital, el pre procesamiento, clasificación y finalmente reconocimiento se ve afectada por el filtro que sucede en cada uno de estos procesos, lo cual afecta directamente el resultado de reconocimiento.

- El estudio de las técnicas de reconocimiento de patrones permitió realizar un comparativo entre tres de las técnicas para el desarrollo de sistemas de reconocimiento del habla, los resultados fueron satisfactorios cuando se proboRedes Neuronales de Base Radial, ésta técnica muestra resultados de reconocimiento superiores al 80% con un corpus compuesto por 130 pronunciaciones para cada hombre y mujer utilizados en diferentes porcentajes para entrenamiento y pruebas, la técnica DTW reconoció un promedio inferior al 50% y por último se encuentra la técnica de modelos ocultos de Markov con un promedio inferior al 30%.

- La experiencia de investigar este tipo de tecnologías nos mostró que los sistemas de reconocimiento del habla no toman en cuenta el contexto cultural, no conocen el significado de lo que el hablante dice y sólo se limitan a clasificar las vocales desde un punto de vista acústica, es entonces recomendable para las personas que quieran trabajar en reconocimiento del habla inmiscuirse con la comunidad. Se debe considerar todo el contexto desde el hablante como ser de una comunidad, esto servirá también en futuros proyectos para identificar variantes de las lenguas.

- La experiencia de desarrollar una aplicación para una comunidad indígena, involucra conocer un poco de una cultura que aunque se encuentra geográficamente cerca, era totalmente ajena y desconocida, pudiendo entender que ellos no están tan familiarizados a trabajar con herramientas computacionales y también que éstas no están adaptadas a sus proceso de etnoeducación, por lo tanto se opto por una interfaz gráfica sencilla con pocas opciones para evitar confusión en uso y además se contaba con la experiencia de la Comunidad Virtual de Etnoeducación.

- El estudio del reconocimiento automático del habla actualmente es ampliamente estudiado y aplicado en diferentes formas y contextos, ya sea que las máquinas sean capaces de reconocer palabras aisladas o habla continua, pero no existe aún un reconocedor universal para todas las lenguas, y menos para lenguas minoritarias como la Nasa Yuwe, por esta razón el prototipo de reconocimiento del subconjunto de vocales para esta lengua se constituye en el primer paso para implementar herramientas que sean útiles a la comunidad Nasa y éstas sean comandadas mediante la voz.

- El reconocimiento automático del habla se ve frecuentemente degradado por el efecto coarticulatorio, que consiste en la influencia de un fonema sobre otro. En este proyecto dicho efecto también se presentó, influyendo así en el porcentaje de reconocimiento de cada una de las técnicas probadas. Esto se presenta debido a que las vocales evaluadas se encontraban en diferentes partes de las palabras (al inicio de la palabra, en el medio y al final) rodeada por consonantes u de otras vocales, es por esto que es aconsejable que el proceso de segmentación se realice de forma manual, ya que se puede hacer un corte más preciso y genera menos errores que si se realizara de manera automática.

- Aunque la complejidad de la lengua Nasa Yuwe y falta de conocimiento de ésta por parte de los tesistas se miraba como una dificultad no fue determinante para realizar una clara diferenciación de los rasgos entre vocales aquí tratadas, puesto que se conto con la ayuda del grupo de estudios lingüísticos de la universidad del Cauca, quienes ya tienen bien caracterizadas las vocales.

- La construcción del prototipo de reconocimiento automático del habla para un subconjunto de vocales de la lengua Nasa Yuwe, constituye un aporte a la comunidad indígena en su camino al acercamiento a las tecnologías de la computación sin que éstas atropellen su cosmovisión, sino que, por el contrario se conviertan en afianzadoras de la cultura Nasa. Es por esto que es importante continuar con la elaboración de este prototipo hasta convertirlo en un sistema capaz de enseñar la lengua Nasa Yuwe tanto a habitantes de la comunidad como a personas que busquen establecer contacto con esta cultura.

6.2. RECOMENDACIONES Y TRABAJO FUTURO

- Construir un corpus de la lengua Nasa Yuwe, que cuente con más hablantes y tipos de grabaciones que integren las vocales usadas y no usadas dentro de este proyecto, ya que el utilizado para las pruebas en este proyecto solamente consta de unos pocos minutos y el número de hablantes es muy limitado.

- Si se pretende continuar la construcción de una herramienta software de reconocimiento de habla de la lengua Nasa Yuwe, se deberá construir un corpus que contenga pronunciaciones que abarquen a todos los resguardos y comunidades de las diferentes localizaciones del departamento o del país, para así lograr que la herramienta no sea dependiente del locutor.

- Utilizar técnicas de pre procesamiento y de reconocimiento de patrones que no fueron utilizadas para comprobar su funcionamiento.

- El modelo Computacional para modelos ocultos de Markov necesita un modelo Acústico y del lenguaje, para realizar un óptimo reconocimiento, esta sería una de las consecuencias para que los resultados de reconocimiento de esta técnica fueran restringidos ya que no se cuenta con un modelo del lenguaje para la lengua Nasa Yuwe.

- Crear herramientas educativas que utilicen el motor de un sistema de reconocimiento para realizar actividades de aprendizaje de palabras y vocales de la lengua Nasa Yuwe.

BIBLIOGRAFIA

[1] L. Rabiner y B.H. Juang.*Fundamental Speech Recognition*, Prentice- Hall International Inc. 1993.

[2] L. Rabinery B.H. Juang,*Speech Recognition by Machine.The Digital Signal Processing Handbook*, CRC Press, IEEE Press.1998.

[3]C. García y D. Tapias. “La Frecuencia Fundamental de la Voz y sus Efectos en Reconocimiento de Habla Continua”, Telefónica Investigación y Desarrollo S.A. Madrid España.

[4] Z. Abul, “Connected Word Speech Recognition”, Faculty of Computing & Information Technology, King Abdul Aziz University, Kingdom of Saudi Arabia

[5] A. Peinado, y J. Segura, *Speech Recognition Over Digital Channels,* Universidad de Granada España, Jhon Wiley & Sons, Ltd. 2006.

[6] Holmes, J. y Holmes W. *Speech Synthesis and Recognition,* Editorial Taylor y Francis Group, 2003

[7] X. Huang, A. Acero, H. Wuen Hon, Spoken Language Processing, A Guide to Theory, Algorithm, and System Development, Editorial Pretince Hall Carnegie Mellon University, 2001

[8] J. Oropeza, “Algoritmos y Métodos para el Reconocimiento del habla en español mediante sílabas”, IPN México D.F. Computación y Sistemas Vol. 9 no. 3 p.p. 270-286, México, 2006 [Online] Disponible en: http://www.ejournal.unam.mx/cys/vol09-03/CYS09307.pdf [Visitada Agosto 2009]

[9] F.Casacuberta, R. Garcia, J. Llisterri, *“Desarrollo de Corpus para Investigación en Tecnologías del Habla (ALBAYZIN)”*, Universidad Politécnica de Valencia, Universidad Politécnica de Madrid, Universidad Autónoma de Barcelona.

[10] A. Martínez,F. Martínez, O. Vidal Cabreray J. GoddardClose, “Estudio del efecto Coarticulatorio en el habla”, Departamento de Ingeniería Eléctrica, Universidad AutónomaMetropolitana, Iztapala, México, vol. 25, pp. 67-77, Marzo 2004.

[11] H. Borrero, Y. Baquero y Z. Alezones, *“*Reconocimiento de Palabras Aisladas utilizando LPC Y DTW, para Control de Navegación de un Mini-Robot*”,* [Online] Disponible en: http://catic.unab.edu.co/2congresomecatronica/images/docum/13robotvoz.pdf

[12] V. Peña Reconocimiento De Palabras Clave En Conversaciones Espontáneas En Castellano, Universidad Autónoma de Madrid, 2008

[13] K.R. Aida–Zade, C. Ardil y S.S. Rustamov, Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems, World Academy of Science, Engineering and Technology N. 19, 2006

[14] “Praat: doing phonetics by computer” [Online]:http://www.fon.hum.uva.nl/praat/

[15] A. Betancourth, C. Córdoba. “Adecuación de señales de voz para el sistema de estimación de la frecuencia fundamental CGAWAVF”, Universidad del Cauca, Cauca, Colombia, 2008.

[16] P. Pérez, Construcción de un reconocedor de voz utilizando Sphinx y el corpus DIMEx100 , Facultad de Ingeniería, Universidad Autónoma de México, México D.F, 2006.

[17] R. Hasan, M. Jamil, G. Rabbani, Speaker identification using mel frequency cepstral coefficients, Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, 2004.

[18] D. Fandiño, Estado del arte en el reconocimiento automático de voz, Universidad Nacional de Colombia, 2005.

[19] L. Kuncheva, Combining pattern classifiers: Methods and algorithms. Editorial John Wiley & Sons 2004.

[20] B. Alvira, A. Sarria. “Estimación de la frecuencia fundamental de señales de voz del

suroccidente colombiano aplicando la técnica *Wavelet*", Universidad del Cauca, Cauca, Colombia, 2006.

[21] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, New Jersey, 1957.

[22] S. Haykin, *Neural Networks: A Comprehensive Foundation (2nd Edition),* Prentice Hall, July 1998.

[23]*N. Pecan, DOFF. Hams and OWE Neural Network for Continuous Speech Recognition 2001*

[24] G. Colmenares, *Redes Neuronales*, Universidad de Los Andes, Mérida-Venezuela, [Online] Disponible en: http://webdelprofesor.ula.ve/economia/gcolmen/programa/redes_neuronales/capitulo4_funciones_bases_radiales.pdf

[25] W. Chou, B. Hwang Juang, Reconigtion in Speech and Language Processing, Editorial CRC Press, 2003.

[26] C. San Martín y R. Carrillo, *"*Implementación de un Reconocedor de Palabras Aisladas Dependiente del Locutor*",* 2004.

[27] P. Escobar, Reconocimiento automático de fonemas usando redes neuronales, Tesis, Facultad de Ingeniería Electrónica, Universidad del Cauca, Popayán, 1999.

[28] I. Villamil, "Aplicaciones en Reconocimiento del habla utilizando HTK", Tesis de maestría, Pontificia Universidad Javeriana, Santa Fe de Bogotá, Mayo 2005 [Online] Disponible en: http://www.javeriana.edu.co/biblos/tesis/ingenieria/tesis95.pdf

[29] HTK Toolkit, Versión 3.0, 2000 [Online]. Disponible en: http://htk.eng.cam.ac.uk

[30] M. Jackson, "Automatic Speech Recognition: Human Computer Interface for Kinyarwanda

language” M.S. thesis, Makerere University Uganda, 2005 [Online] Disponible en: http://www.fon.hum.uva.nl/david/ba_shs/kinyarwanda_rcognizer_final_report_1.pdf

[31] F. Adriaans, M. Heukelom, M. Koolen, T. Lentz, O. d Rooij, D. Vreeswijk, R. Son, “Speech Technology Project 2004 Building an HMM speech recogniser for Dutch”, Holanda, July 2004. [Online] disponible en: http://www.fon.hum.uva.nl/david/ba_shs/educational_dutch_speechRecognizer_FinalReport2004.pdf

[32] J. Oropeda, “Algoritmos y métodos para el reconocimiento de voz en Español mediante silabas” en Computación y Sistemas, Vol. 9 Núm. 3, pp. 270-286, 2008.
Disponible en: http://www.ejournal.unam.mx/cys/vol09-03/CYS09307.pdf

[33] I. Kirschninig, N. Aguas, A. Ahuactzin, “Aplicación de Tecnología de Voz en La Enseñanza del Español”**.** Universidad de las Américas – Puebla- México, 2000.
Disponible en: http://ict.udlap.mx/people/ingrid/ingrid/HAVOL2000a.pdf

[34] CSLU Toolkit. Ultima versión 2004. [Online]. Disponible en: http://cslu.cse.ogi.edu/toolkit/

[35] G. Meneses, M. Castillo. “Compresión y descompresión de voz mediante técnicas de procesamiento digital de imágenes utilizando wavelets”, Universidad del Cauca, Cauca, Colombia, Marzo 2008.

[36] L. Villaseñor, M. Montes y Gómez, D. Vaufreydaz J-F. Serignat, “Elaboración de un Corpus Balanceado para el Cálculo de Modelos Acústicos usando la Web”, Laboratorio de Tecnologías del Lenguaje, Ciencias Computacionales, INAOE, México. Laboratoire CLIPS/IMAG, Francia, Diciembre, 2003. Disponible en: http://www-prima.inrialpes.fr/Vaufreydaz/Telechargement/Villasenor03b.pdf

[37] un corpus para el asturiano, las tecnologías lingüísticas en la consolidación de las lenguas minorizadas Roser Saurí Colomer, computer Science Departament ,Brandeis University.

[38] T. McEnery y A. Wilson, Corpus Linguistics, 2nd Edition, Editorial Edinburgh University Press Ltda, 2001.

[39] T. Rojas, M. Farfán, *Aprendamos desde nuestro fogón CARTILLA PARA EL APRENDIZAJE DE NASA YUWE como segunda lengua,* 2007.

[40] T. Rojas, “Por los caminos de la recuperación de la lengua Paéz (Nasa Yuwe)”, Universidad del Cauca, Popayán, 2006. Págs. 279 – 286. Editorial Letrarte editores.

[41] T. Rojas, “Desde arriba y por abajo construyendo el alfabeto nasa. La experiencia de la unificación del alfabeto de la lengua Páez (Nasa Yuwe) en el Departamento del Cauca – Colombia”, Abril 2002.

[42] R. Perry, Curso de Introducción al análisis de la señales del habla, Universidad Nacional de Colombia, 2009

[43] “El editor de audio libre y multiplataforma”[Online]: http://audacity.sourceforge.net/?lang=es

[44] C. Esteve, “Reconocimiento del Locutor Dependiente del Texto mediante la Adaptación de Modelos Ocultos de Markov fonéticos”, Área de tratamiento de voz yseñales, Universidad Autónoma de Madrid, 2007

[45] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech”, Journal of Acoustic Society of America, Vol. 87, Nº 4, April 1990, pp. 1738-1752.

[46] S. Ambler, The Agile Unified Process, version actualizada Mayo 2006 [Online]. Disponible en http://www.ambysoft.com/unifiedprocess/agileUP.html

[47] Roger S. Pressman. Ingeniería de Software. Un enfoque práctico. Mc Graw Hill, 2002.

[48] L. Bass, P. Clements, and R. Kazman, *Software Architecture in Practice*, Addison Wesley, 1998

[49] MathWoks accelerating the pace of engineering and science, [Online]

http://www.mathworks.es/products/matlab/description1.html

[50] Dynamic Time Warp (DTW) in Matlab, [Online] Disponible en:
http://labrosa.ee.columbia.edu/matlab/dtw/

[51] Neural Network Toolbox, [Online] Disponible en:
http://www.mathworks.com/help/toolbox/nnet/

[52] Gaussian Observation Hidden Markov Model, [Online] Disponible en:
http://read.pudn.com/downloads95/sourcecode/others/382708/hmm/hmm.m__.htm

[53] Calculate Likelihood for Hidden Markov Model, [Online] Disponible en:
http://read.pudn.com/downloads74/sourcecode/speech/264483/hmm/hmm_cl.m__.htm