SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de Ciudad Madero

INSTITUTO TECNOLÓGICO DE CIUDAD MADERO

División de Estudios de Posgrado e Investigación

RECONOCIMIENTO DE HABLA EN PALABRAS AISLADAS EN LENGUAS INDÍGENAS DE SAN LUIS POTOSÍ

OPCIÓN I

Tesis Profesional

Para obtener el grado de:

Maestro en Ciencias de la Computación

Presenta:

I.E. Jesús Alviso Vargas

G06071568

Director de Tesis:

Dr. Arturo Hernández Ramírez

Cd. Madero, Tamaulipas, México Marzo 2015

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Instituto Tecnológico de Ciudad Madero

"2015, Año del Generalísimo José María Morelos y Pavón"

Cd. Madero, Tamps; a 20 de Febrero de 2015.

OFICIO No.: U5.041/14
AREA: DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN
ASUNTO: AUTORIZACIÓN DE IMPRESIÓN DE TESIS

ING. JESÚS ALVISO VARGAS
NO. DE CONTROL G06071568
P R E S E N T E

Me es grato comunicarle que después de la revisión realizada por el Jurado designado para su examen de grado de Maestría en Ciencias de la Computación, el cual está integrado por los siguientes catedráticos:

PRESIDENTE :	DR.	JOSÉ ANTONIO MARTÍNEZ FLORES
SECRETARIO :	M.C.	JOSÉ APOLINAR RAMÍREZ SALDÍVAR
VOCAL :	DR.	ARTURO HERNÁNDEZ RAMÍREZ
SUPLENTE	DR.	HÉCTOR JOAQUÍN FRAIRE HUACUJA

DIRECTOR DE TESIS : DR. ARTURO HERNÁNDEZ RAMÍREZ

Se acordó autorizar la impresión de su tesis titulada:

"RECONOCIMIENTO DE HABLA EN PALABRAS AISLADAS EN LENGUAS INDÍGENAS DE SAN LUIS POTOSÍ"

Es muy satisfactorio para la División de Estudios de Posgrado e Investigación compartir con Usted el logro de esta meta.

Espero que continúe con éxito su desarrollo profesional y dedique su experiencia e inteligencia en beneficio de México.

A T E N T A M E N T E
"POR MI PATRIA Y POR MI BIEN"®

M. P. MARÍA YOLANDA CHÁVEZ CINCO
JEFA DE LA DIVISIÓN

S.E.P.
DIVISIÓN DE ESTUDIOS
DE POSGRADO E
INVESTIGACIÓN
I.T.C.M

c.c.p.- Archivo
Minuta
MYCHC 'NLCO ' jar

Ave. 1° de Mayo y Sor Juana I. de la Cruz, Col. Los Mangos, CP. 89440 Cd. Madero, Tam.
Tel. (833) 357 48 20, Fax, Ext. 1002, e-mail: itcm@itcm.edu.mx
www.itcm.edu.mx

Declaración de Originalidad

Declaro y prometo que este documento de tesis es producto de mi trabajo original y que no infringe los derechos de terceros, tales como derechos de publicación, derechos de autos, patentes y similares.

Además, declaro que en las citas textuales que he incluido (las cuales aparecen entre comillas) y en los resúmenes que he realizado de publicaciones ajenas, indico explícitamente los datos de los autores y las publicaciones.

Además, en caso de infracción de los derechos de terceros derivados de este documento de tesis, acepto la responsabilidad de la infracción y relevo de ésta a mi director y codirectores de tesis, así como al Instituto Tecnológico de Cd. Madero y sus autoridades.

10 de Marzo de 2015, Cd. Madero, Tamps.

_____

Jesús Alviso Vargas

Agradecimientos

Realizar un proyecto como el presente presenta una labor enorme, y de hacerlo por cuenta propia hubiese sido una labor muy exhaustiva. Afortunadamente, ese no fue el caso, y es necesario agradecer a aquellas personas que me apoyaron de una u otra manera durante el desarrollo de este trabajo:

- Primeramente a Dios, que siempre me ha guiado y llevado por el camino correcto.
- A mi asesor, el Dr. Arturo Hernández Ramírez, que me apoyó y guió durante el desarrollo de todo el proyecto.
- Al M.C.C Carlos Arturo Hernández Zepeda, que me guió durante los primeros pasos del desarrollo del proyecto.
- A mis padres, Jesús Alviso Castro y María Victoria Vargas Ortega, por su apoyo y confianza incondicionales hacia mi durante todo este tiempo.
- A mis hermanas, Judith Alejandra Alviso Vargas y María Victoria Alviso Vargas por apoyarme también durante este tiempo de desarrollo.
- Y por último, a la comunidad de hablantes del idioma Tének de México, sin su apoyo este proyecto nunca hubiera podido completarse.

A todos ustedes, gracias y sepan que sin su apoyo, este proyecto no hubiese podido ser completado.

Resumen

Hoy en día el desarrollo tecnológico crece dando grandes pasos muy rápidamente, obteniendo aplicaciones nuevas para el uso de la misma y facilitando la vida del ser humano. Sin embargo, aún existen ciertas áreas de oportunidad donde es posible aprovechar este desarrollo acelerado para solucionar problemas que no han considerado a las nuevas tecnologías como una herramienta de respuesta.

El presente trabajo busca aplicar el desarrollo tecnológico del campo de reconocimiento automático del habla (RAH) en un área muy poco investigada: las lenguas autóctonas del país. Como primer paso, se pretende acercar los métodos que han resultado efectivos en los idiomas más utilizados, generando un modelo, y usarse como punto de partida para la generación de un sistema útil, que permita -entre otras cosas- acercar el uso de la tecnología a aquellas personas que no pertenecen a la media de la población (lingüísticamente hablando).

El sistema de RAH que se implementará será uno de los más básicos: un sistema de reconocimiento de palabras aisladas, libres de ruido, y de vocabulario corto. El fin principal de este sistema es el de contar con un punto de partida para después desarrollar sistemas de habla continua, de vocabulario amplio.

El sistema de RAH que se presentará considerará un análisis fonético, utilizando los Coeficientes Cepstrales en Frecuencia de Mel como característica observable de la señal de audio, y utilizando como modelo una aplicación de los Modelos Ocultos de Markov.

El proceso de evaluación se realizará comparando las estrategias de clasificación junto con los parámetros de las mismas para encontrar aquella combinación con mejores resultados. En este trabajo se obtuvieron resultados de hasta un 80% de efectividad en la detección de palabras aisladas.

Finalmente, el propósito general de este proyecto es el de continuar con una escasa serie de trabajos anteriores que han comenzado con la aplicación de la tecnología hacia el sector indígena de México, desarrollando herramientas que permitan en un futuro apoyarlos en muchas de las complicaciones que presentan hoy en día.

Abstract

Nowadays, the technological development grows greatly at incredible speed, getting new applications to its use and making the life of the human being very easy. However, there is still some development areas where is possible to use this accelerated development to solve problems that have not considered the use of these new technologies as a tool to solve them.

This work tries to apply the technological development of the automatic speech recognition field in a very few researched fiel: the native languages of Mexico. As a first step, it looks for using those methods that have been effectives on the most used languages, developing a model, and using it as a starting point for the generation of a useful system, that allows -among other purposes- to approach the use of technology to those people that does not belong to the population means (speaking of language).

The speech recognition system to be implemented would be one of the most basic: a isolated words recognition system, noise-free and of a short dictionary of words. The main purpose of this system is to have a starting point to develop later systems of continuous speech, with a big dictionary of words.

The speech recognition system to be presented will consider a phonetic analysis, using the Mel-Frecuency Cepstral Coefficients (MFCC) as a observable feature of the audio signal, and using as a model an application of the Hidden Markov Models (HMM).

The evaluation process will be made comparing different classification strategies along with their parameters to find which combination would give better results. On this project the system got results up to a 80% of effectiveness on the detection of isolated words.

Finally, the main purpose of this project is to continue with a scarce series of previous projects that has begun with the application of the technology to the native population of Mexico, developing tools that would allow in a future to help them in many of the complications that they have to face nowadays.

Índice

Índice de Figuras y Tablas

1
Introducción

Es una realidad el hecho de que cada vez más el uso de la tecnología se vuelve un elemento cotidiano. Desde el uso de una computadora, hasta los teléfonos móviles inteligentes, o los sistemas automáticos de manejo y de diagnóstico de un vehículo.

Gracias a la explosión tecnológica, se ha logrado resolver problemas que antes no se podían solucionar, o eran muy complejos para llegar a una solución aceptable. Estos problemas no se limitan a aspectos técnicos, sino que también incluyen problemas sociales, económicos y culturales, donde el uso de la tecnología es aplicado en la toma de decisiones, y en la conservación y difusión de las estrategias de conservación y difusión de la cultura.

Otro aspecto importante de la actualidad es la diversidad cultural. Existe una mezcla cultural tanto nacional como internacional que lleva a la sociedad a adaptarse a nuevas tradiciones, o a modificar las ya existentes. Como consecuencia de esto, la "esencia cultural" de las naciones aparenta una lenta desaparición de la vida actual de sociedad.

Debido a este fenómeno, una cantidad cada vez menor de habitantes de una región continúan celebrando sus propias costumbres y tradiciones, o viviendo aspectos de su propia historia y cultura.

En México, así como en el resto del mundo, uno de los aspectos que se empiezan a perder es el habla de lenguas autóctonas del país [31]. Una cantidad cada vez menor de la población entiende y/o habla una de estas lenguas, y una cantidad aún menor es multilingüe (es decir, de las personas que hablan una lengua autóctona, pocas hablan otra lengua, dígase, el español). De esta situación de la pérdida del habla de lenguas autóctonas se derivan muchas problemáticas, generalmente de índole social.

De estos problemas, surge la idea del presente proyecto, el cual consiste en el desarrollo de una herramienta que permita representar –en cierto grado- la

función de una persona que hable tanto una lengua indígena como el español, que permita actuar como una manera tanto de comunicar a dos personas que anteriormente tenían gran dificultad para hacerlo, tanto como una manera de lograr preservar una lengua antes de que la explosión de diversidad cultural logre erradicarla. De esta manera, se podrían desarrollar diferentes sistemas que permitan resolver los problemas mencionados anteriormente.

1.1 Planteamiento del problema

Se requiere desarrollar un sistema de RAH (Reconocimiento Automático del Habla) que permita identificar con precisión diferentes palabras pronunciadas en una lengua autóctona de México. La problemática consiste en definir la estructura, los métodos y las diferentes características de un sistema de RAH (donde se involucra tanto el proceso de representación de la voz como el modelo de reconocimiento) de buena calidad para que, al realizarle las pruebas pertinentes, entregue buenos resultados.

1.2 Justificación

Actualmente existe alrededor del mundo una tendencia de migración, en la cual los pueblos indígenas buscan mejorar sus oportunidades reubicándose hacia regiones más urbanas e industriales [25]. México no es la excepción.

Sin embargo, muchas de estas personas en el país no conocen o no quisieron aprender el idioma español, por lo que al migrar a estas regiones el primer problema que surge es la incapacidad de comunicarse con los habitantes de las zonas urbanas, ya que la mayoría de los ciudadanos no son capaces de hablar las diferentes lenguas que se hablan en México.

De esta falta de comunicación, surgen diferentes problemas relacionados:

- Incapacidad de obtener una mejor calidad de vida, puesto que los empleadores o encargados de instituciones de apoyo social son incapaces de comprenderlos.
- Falta de apoyo por parte de los ciudadanos al no comprender las necesidades de estos migrantes [25].
- Injusticia, puesto que si llegan a ser inculpados de algún crimen, son incapaces de defenderse al no poderse comunicar efectivamente [30].

- Pérdida de la lengua, al haber cada vez menos hablantes en estos pueblos indígenas [29].

Una alternativa que permite enfrentar los problemas mencionados anteriormente de manera efectiva consiste en la creación de un sistema de identificación del habla, el cual interprete efectivamente un mensaje en una lengua autóctona. De esta manera, se pueden desarrollar herramientas tales como diccionarios y traductores enfocados a solucionar estos problemas de comunicación y conservación.

Este proyecto pretende dar el primer paso hacia la creación del sistema mencionado anteriormente, al realizar un sistema de reconocimiento de palabras de manera aislada.

1.3 Preguntas de investigación

Basado en el problema definido, surge una serie de preguntas que buscarán ser respondidas durante el desarrollo del proyecto. Estas son:

- ¿Cuáles son las características que permiten que el sistema entregue buenos resultados?
- ¿Qué resultados se obtendrán al aplicar las estrategias conocidas en una lengua no investigada anteriormente?
- ¿Qué tan grande debe ser la base de datos de voz para realizar un sistema como el que se propone?
- ¿Cómo saber que el sistema es efectivo?
- ¿Qué complicaciones se presentarán en cuanto a la variación en pronunciación de las personas?

1.4 Objetivos

1.4.1 Objetivo general

Desarrollar un sistema de Reconocimiento Automático del Habla que permita identificar correctamente un conjunto de palabras de una lengua autóctona de México de manera efectiva.

1.4.2 Objetivos específicos

- Definir la estrategia que se utilizarán para desarrollar el sistema RAH.
- Obtener las muestras de la lengua indígena.
- Entrenar correctamente el sistema de RAH.
- Someter el sistema de RAH a un conjunto de pruebas.
- Someter el sistema RAH a los estándares globales de reconocimiento del habla.
- Realizar pruebas del sistema RAH para encontrar puntos de mejora o en su defecto, comprobar su efectividad.

1.5 Alcances

- Realizar un sistema de RAH de una lengua autóctona de México.
- Entrenar efectivamente un conjunto de al menos 8 palabras diferentes, que el sistema pueda reconocer correctamente.
- Diseñar el sistema de tal manera que permita un entrenamiento futuro de más palabras.

1.6 Limitaciones

Para fines del desarrollo de esta tesis, el sistema se limitará a:

- Se desarrollará la identificación de palabras de una sola lengua autóctona.
- El sistema se limitará al reconocimiento de palabras aisladas.

1.7 Estructura de la tesis

Capítulo 2. Estado del Arte. En este capítulo se abordarán los conocimientos previos, investigaciones y desarrollos que se han llevado a cabo y que representan cierta importancia al desarrollo del proyecto, es decir, *aquello que se ha realizado y que es importante para el proyecto*.

Capítulo 3. Marco Teórico. Se abordarán los fundamentos, conceptos y estrategias teóricas que se utilizarán en el proyecto, es decir, *toda aquella base teórica fundamental para el desarrollo del proyecto*.

Capítulo 4. Metodología. En este capítulo se indicará de manera clara los diferentes procedimientos y estrategias realizadas para el desarrollo del proyecto, es decir, *cómo se ha realizado el proyecto*.

Capítulo 5. Evaluación. Una vez determinada la metodología, es necesario establecer criterios para analizar la efectividad de la misma. Este capítulo definirá los criterios utilizados para la misma, así como los resultados de la aplicación de la evaluación, es decir, *los resultados de aplicar la metodología propuesta*.

Capítulo 6. Conclusiones. Por último, es importante analizar los resultados obtenidos, obtener conclusiones respecto a ellos referente al desarrollo del proyecto y enunciar recomendaciones para su mejora o trabajo futuro, es decir, *enlistar lo aprendido del desarrollo del proyecto*.

2
Estado del Arte

2.1 Historia del reconocimiento de voz

A continuación se presenta un breve recuento de la historia del reconocimiento de voz [4]:

El reconocimiento de voz puede rastrearse desde el año de 1870, donde Alexander Graham Bell buscaba construir un dispositivo que facilitara el habla a las personas con problemas auditivos. De estos trabajos surgió el teléfono, dispositivo que se utiliza incluso en la actualidad.

En 1880, Tihamir Nemes solicita permiso para una patente para desarrollar un sistema de transcripción automática que identificara secuencias de sonidos y los imprimiera como texto. Pero fue rechazado debido a ser un "proyecto no realista".

30 años después (1910), AT&T Bell Laboratories construye la primera máquina capaz de reconocer voz (basada en plantillas) de los 10 dígitos del Inglés. Requería un extenso ajuste a la voz de una persona, pero una vez logrado tenía un 99% de certeza. Por lo tanto surge la esperanza de que el reconocimiento de voz fuese un proceso simple y directo.

Sin embargo, para mediados de la década de los 60's, se reconoce que el reconocimiento de voz era un proceso mucho más complejo de lo que habían anticipado. Por lo tanto empiezan a reducir los alcances y se enfocan a sistemas más específicos, con las siguientes características:

- Dependientes del Locutor.
- Flujo discreto de habla (con espacios / pausas entre palabras)
- Vocabulario pequeño (menor o igual a 50 palabras)

Estos sistemas empiezan a incorporar técnicas de normalización del tiempo (minimizar diferencia en velocidad del habla). Además, ya no buscaban una exactitud perfecta en el reconocimiento.

Para estos momentos, IBM y CMV se encontraban trabajando en reconocimiento de voz continuo, el HAL 9000, pero no se ven resultados hasta los años 70's.

A principios de los 70's se produce el primer producto de reconocimiento de voz, el VIP100 de Threshold Technology Inc. Éste utilizaba un vocabulario pequeño, era dependiente del locutor, y reconocía palabras aisladas. Sin embargo, ganó el U.S. National Award en 1972.

A raíz de esto nace el interés del U.S. Department of Defense, en especial del DARPA (Defense Advanced Research Projects Agency) y gracias al lanzamiento de grandes proyectos de investigación y de financiamiento por parte del gobierno estadounidense se impulsa la era de la inteligencia artificial.

Los proyectos financiados por DARPA buscaban el reconocimiento de habla continua, de vocabulario grande. Esto impulsa que los investigadores se enfoquen al entendimiento del habla.

Debido a las necesidades de los proyectos, los sistemas empiezan a incorporar módulos de:

- análisis léxico (conocimiento léxico)
- análisis sintáctico (estructura de palabras)
- análisis semántico (significado)
- análisis pragmático (intención)

De estos proyectos surge el sistema HARPY, desarrollado por CMU para DARPA SUR (Speech Understanding Research), considerado como el primer sistema exitoso de los proyectos de DARPA.

Entre los años 80's y 90's surgen los sistemas de vocabulario amplio, que ahora son la norma. (Más de1000 palabras). Adicionalmente bajan los precios de estos sistemas.

Para el siglo XXI, los sistemas de reconocimiento del habla aparecen en el ámbito casero y comercial, con aplicaciones que facilitan el uso de un sistema o como traductores.

2.2 Ejemplos actuales de sistemas de reconocimiento del habla

Actualmente se ha vuelto popular el uso de sistemas de reconocimiento automático del habla en software de uso cotidiano. Hoy en día se pueden ver diferentes aplicaciones que hacen uso de estos sistemas, a continuación se muestran algunos de los más populares

2.2.1 Sistema de reconocimiento de voz de Windows

El sistema de reconocimiento de voz de Windows, denominado "motor de reconocimiento de voz" [35] es una característica de los sistemas de Windows que permite la interacción con los diferentes programas por medio del habla continua. Estos motores, sin embargo, son específicos del idioma y la región donde serán utilizados los programas, así como requieren de un entrenamiento del usuario final del motor, lo que lo convierte en un sistema semi-dependiente del locutor. También especifica recomendaciones sobre la manera de pronunciar el dictado y las condiciones de ruido presentes.

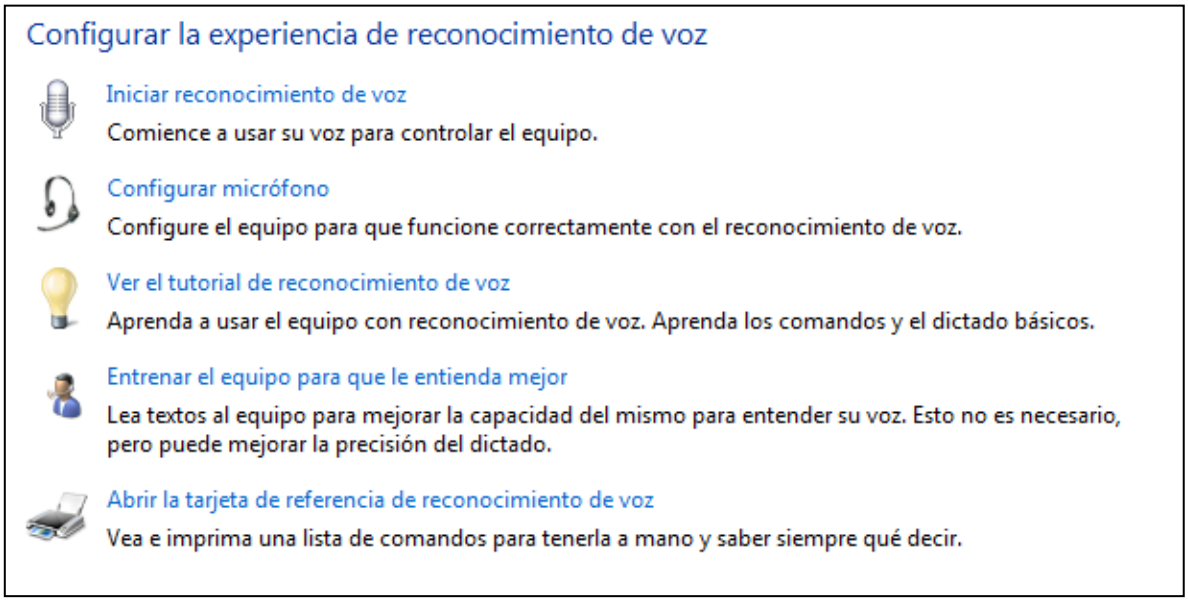


Fig. 2.1: Opciones de configuración y activación del sistema de reconocimiento de voz de Windows.

2.2.2 Siri

Siri [32] es un asistente personal instalado en los dispositivos iPhone e iPad fabricados por la compañía Apple, la cual es capaz de realizar acciones tales como enviar mensajes, actualizar la agenda, realizar llamadas, realizar búsquedas online, entre otras. Esto lo logra al combinar las estrategias de reconocimiento automático del habla con un sistema de lenguaje natural. Sin embargo, para utilizar Siri es necesario contar con una conexión activa a Internet, esto debido a que el proceso de reconocimiento del habla se realiza por medio de los servidores de Apple, el cual recibe la información del habla comprimida, la interpreta y envía el reconocimiento de vuelta al dispositivo.

Fig. 2.2: Ejemplo de la interfaz de Siri.

2.2.3 Ok Google

Google ha implementado en sus búsquedas un motor de reconocimiento automático del habla [34], el cual recibe por medio de un dispositivo de grabación (micrófono) el mensaje y lo envía a sus servidores, los cuales interpretan la señal, identifican el mensaje y realizan la búsqueda relacionada. Junto con este sistema de reconocimiento, se incorpora un modelo de lenguaje natural que permite que un usuario sin experiencia en consultas pueda realizar búsquedas sobre temas específicos. Sin embargo, al igual que Siri, es necesario contar con conexión activa a Internet.

2.3 Investigaciones actuales sobre RAH

2.3.1 Representación de la voz

El hecho de trabajar con vectores de características, sin duda, constituye un elemento clave en un sistema de RAH (Reconocimiento Automático del Habla). El concepto mismo de vector de rasgos o de características es lo que dota al reconocimiento automático de su enorme potencial práctico al reducir la complejidad de la voz, a la muy manejable información condensada en un vector de datos numéricos.

La etapa de elección de características es crítica y la bondad del sistema final estará completamente determinada por las características escogidas.

Los métodos para extraer características se pueden dividir en [8]:

Métodos paramétricos de extracción de características. Estos métodos están basados principalmente en la técnica de predicción lineal. A partir de los coeficientes de predicción se realizan unas transformaciones que dan lugar, por ejemplo, a los coeficientes de reflexión o a los pares de líneas espectrales. Básicamente, el análisis predictivo lineal aproxima la envolvente del espectro de voz prediciendo muestras de la señal a partir de una combinación lineal de las muestras precedentes. Para ello, se minimiza el cuadrado de las diferencias entre la muestra actual y la predice linealmente sobre un intervalo finito, determinando así un conjunto de coeficientes de predicción.

Métodos no paramétricos de extracción de características. Dentro de los métodos no paramétricos, se aplican técnicas no paramétricas de análisis de la señal de voz. Entre ellas se incluyen las técnicas de análisis por bancos de filtros mediante herramientas basadas en las Transformadas de Fourier, las técnicas de procesado homomórfico y el análisis cepstral.

Métodos híbridos. Éstos representan las técnicas resultantes de la combinación de algunos de los métodos anteriormente mencionados. Ejemplo de estas métodos son la técnica de análisis fruto de la combinación de la predicción lineal y el análisis cepstral, LPC-Cepstrum, y una de las técnicas con la que mejores resultados se obtienen en aplicaciones de reconocimiento, el Mel-Cepstrum, fruto de la combinación del análisis cepstral de la señal de voz y una

transformación de la escala lineal de frecuencias en función de la influencia que poseen las bandas críticas en la sensibilidad del sistema auditivo humano.

2.3.2 Modelos

Flores Paulín menciona en [4] los modelos más efectivos para identificar el habla, como se muestra a continuación:

Redes Neuronales. Mencionada en 1988 por Lippman, las redes neuronales son estructuras de procesamiento de información, formadas por nodos simples conectados entre sí mediante pesos y agrupados en capas, de las que se distinguen las capas de entrada y salida. Las capas de entrada y salida pueden tener uno o más nodos y una o más entradas/salidas, mientras que entre la entrada y la salida hay una o mas capas ocultas. En el RAH las redes neuronales han conseguido resultados prometedores, aunque su mayor problema es el entrenamiento que se tienen que realizar a las mismas

Modelos Ocultos de Markov. Este modelado surgió como una alternativa a un método anterior de comparación por plantillas, buscando solucionar los problemas que presentaban las plantillas. Al utilizar modelos estocásticos se incorporó la filosofía de comparación de patrones, difiriendo en la forma en la que se obtienen los patrones, el tipo de patrón, la medida de distancia y la forma de realizar el alineamiento temporal utilizando estos últimos. Actualmente es el más utilizado, debido a su facilidad de entrenar y a sus buenos resultados.

Modelos de Mezclas Gaussianas. Este otro modelo ha presentado también resultados favorables al realizar el proceso de RAH por medio de funciones de densidad de probabilidad, aunque sus mayores aplicaciones han sido en la rama hermana de detección del hablante. Se suele utilizar en combinación con los Modelos Ocultos de Markov para mejorar su porcentaje de detección.

2.4 Lenguas autóctonas de México

Fig. 2.3: Instituto Nacional de Lenguas Indígenas

El Instituto Nacional de Lenguas Indígenas (INALI), elaboró el "catálogo de lenguas indígenas mexicanas" [26], bajo mandato federal. Este catálogo se dividió en dos etapas.

El INALI publicó como resultado de la primera etapa del proyecto, en el año 2005, el Catálogo de lenguas indígenas mexicanas: Cartografía contemporánea de sus asentamientos históricos. Esta obra consiste en una colección de 150 mapas elaborados a partir de la información censal levantada en el año 2000 por el Instituto Nacional de Estadística, Geografía e Informática (INEGI). En tales mapas se consignan, con respecto al territorio histórico de cada pueblo indígena del país, las localidades donde un determinado porcentaje de su población habla la respectiva lengua nacional originaria.

En la segunda etapa del proyecto, la atención se centró en la diversidad lingüística correspondiente al habla propia de los pueblos indígenas arraigados en el territorio nacional. De esta etapa, se llegó a la conclusión de que la realidad lingüística del país es mucho más compleja de lo que en términos generales se había creído, por lo que fue necesario considerar nuevos criterios de agrupación de las diferentes lenguas autóctonas habladas en el país.

2.4.1 Familias, agrupaciones y variantes

Según el "Catálogo de lenguas indígenas mexicanas" [26], las lenguas autóctonas nacionales se dividen en 3 niveles: familias, agrupaciones y variantes (en ese orden)

Una familia es la estructura de mayor nivel dentro de la clasificación. Se define como un conjunto de lenguas cuyas semejanzas estructurales y léxicas se deben a un origen histórico común. En el catálogo se identificaron 11 familias distintas.

Una agrupación representa el segundo nivel de clasificación, debajo de la familia. Se define como el conjunto de variantes lingüísticas comprendidas bajo el nombre dado históricamente a un pueblo indígena. En este nivel fueron identificadas 68 diferentes agrupaciones.

Una variante representa el último nivel de clasificación. Se define como una forma de habla que: a) presenta diferencias estructurales y léxicas en comparación con otras variantes de la misma agrupación lingüística; y b) implica para sus usuarios una determinada identidad sociolingüística, que se diferencia de la identidad sociolingüística de los usuarios de otras variantes. El catálogo

identifica, por último 364 diferentes variantes. Es importante mencionar que, de acuerdo a lo especificado por el INALI, cada variante lingüística debe ser considerada como una lengua independiente.

2.4.2 Proyectos de comunicación entre hablantes y no hablantes

A continuación se muestran algunos esfuerzos realizados para apoyar a la población no hablante del español en cuestiones legales y médicas.

La Comisión Nacional para el Desarrollo de los Pueblos Indígenas (CDI) cuenta con un proyecto de apoyo a aquellas personas que no pueden hablar español, por medio de un proyecto llamado "Atención a Indígenas en Materia Penal y Penitenciaria "[28], En este proyecto se prestan servicios como la traducción o la financiación de las cuotas de fianza de los sistemas legales para ayudar a las personas indígenas a salir de prisión, bajo ciertas condiciones.

En el año 2011 surgió en Internet un sitio denominado "Diccionario Educein Huasteco (Tének) - Español" [27], el cual se ha basado en una guía publicada en 1983. Este sitio tiene como propósito principal "desarrollar una herramienta útil para el personal de salud que no habla huasteco para aplicarse en la realización de un historial clínico, en la comunicación y educación en temas de salud y el entendimiento más adecuado del estado clínico de los pacientes".

2.4.3 Conservación de la lengua

Entre las acciones que desde el INALI se realizan para reivindicar las lenguas, está la creación de “nidos de lengua”, donde adultos enseñan a los niños las cuestiones lingüísticas en las zonas con más riesgo de perder su habla nativa. [31]

También se publicó la convocatoria del Premio de Literaturas Indígenas de América [31], que fue entregado durante la edición 2013 de la Feria Internacional del Libro en Guadalajara. Este premio también fue convocado para el año 2014.

2.5 Investigaciones y trabajos relacionados

2.5.1 Corpus de Lenguas Indígenas Mexicanas para la Identificación Automática del Lenguaje Hablado (2013)

Elaborada por Carlos Arturo Hernández Zepeda [5], esta tesis documenta el desarrollo de un corpus de tres lenguas indígenas, siguiendo estándares internacionales, a fin de contar con una base para trabajos futuros. De igual manera, se presentó, como posibles trabajos futuros, la aplicación de identificación de lenguaje (LID) con los Shifted Delta Coefficient (SDC) como parámetro y las Máquinas de Soporte Vectorial (SVM) como técnica de modelado.

2.5.2 Técnicas para el reconocimiento de voz en palabras aisladas en la lengua náhuatl (2009)

Esta tesis, presentada por Juan Carlos Flores Paulín [4], muestra los resultados de la aplicación de diferentes técnicas y parámetros para identificar el habla en la lengua náhuatl, obteniendo los mejores resultados al combinar los Modelos Ocultos de Markov (HMM) con los Coeficientes Cepstrales en Escala de Mel (MFCC)

2.5.3 On the Development of Speech Resources for the Mixtec Language (2013)

Santiago-Omar Caballero-Morales [12] presenta en este artículo un corpus de habla para el mixteco ubicado en el estado de Oaxaca. Dentro de las aplicaciones que presenta con el uso del corpus es el desarrollo de un sistema de RAH adaptable al hablante.

3
Marco Teórico

3.1 La voz

3.1.1 ¿Qué es?

La voz es una forma de energía de naturaleza analógica. Ésta es producida por medio de variaciones de presión reflejadas en ondas sonoras, cuyas vibraciones son captadas por el sentido del oído (o en el caso electrónico, algún dispositivo como un micrófono), procesadas e interpretadas para poder entender un mensaje.

3.1.2 ¿Cómo se mide?

Según Contreras en [10], Debido a que la voz es un fenómeno relacionado con vibraciones y la presión, puede ser medida electrónicamente. Idealmente, una señal de voz es representada por medio de una gráfica, donde se pueden apreciar tres elementos fundamentales:

- La amplitud, que representa la intensidad de la voz, también conocido como volumen
- La frecuencia, que representa la variación en el tono de la voz. Esta es una de las características más importantes para la articulación de palabras, puesto que el habla se compone de variaciones de tonos ejercida por una persona.
- El tiempo, el cual es una de las formas más comunes de representar una gráfica como variable independiente (x)

3.2 Reconocimiento automático del habla

3.2.1 ¿Qué es?

La comunicación oral ha sido una de las formas más efectivas que ha tenido el ser humano para comunicarse. También es un hecho que el uso de la tecnología es constante y cada vez mayor. De esta manera, se ha buscado cambiar el método tradicional de comunicación con una máquina (tecleando), por uno más efectivo (hablando). El reconocimiento automático del habla (RAH) consiste entonces en el procedimiento que necesita realizar una computadora para captar, procesar e identificar un conjunto de palabras pronunciadas por un hablante.

Es importante mencionar que el proceso de RAH es diferente a un sistema de comprensión del habla. El proceso de comprensión del habla incluye sistemas de RAH, pero además incluye reglas semánticas y gramaticales para comprender el mensaje dicho y entenderlo correctamente.

3.2.2 Niveles de comprensión de un sistema de RAH

Navarro menciona en [8] que en la comunicación oral existen varios niveles de percepción, que interactúan entre sí. Cada uno de estos niveles aplica cierto conocimiento al proceso de comprensión del habla y, de esa manera, también es aplicable a los sistemas de RAH. Desde este punto de vista, los niveles básicos son los siguientes:

- Acústico. Se analizan las características físicas de la señal vocal para extraer información relevante en el conocimiento.
- Fonético. Se determina un objeto sonoro elemental (fonemas, sílabas, palabras, letras, etcétera) que conforman los demás elementos.
- Léxico. En él se generan hipótesis de palabras en función de las hipótesis de unidades menores. Esto es, si del nivel acústico se obtienen fonemas, en el nivel léxico se determina la combinación de éstas para generar unidades mayores, palabras.

- Sintáctico. Este nivel considera las reglas gramaticales basadas en el uso y normalización del lenguaje. Esto es, actua sobre la forma de combinar palabras para generar frases.
- Semántico. Se generan hipótesis sobre el significado de las frases obtenidas, eliminando interpretaciones absurdas y comprobando la coherencia del mensaje
- Pragmático. Este nivel se puede considerar como por encima del semántico, y se considera como la relación entre los símbolos obtenidos y los usuarios que los producen.

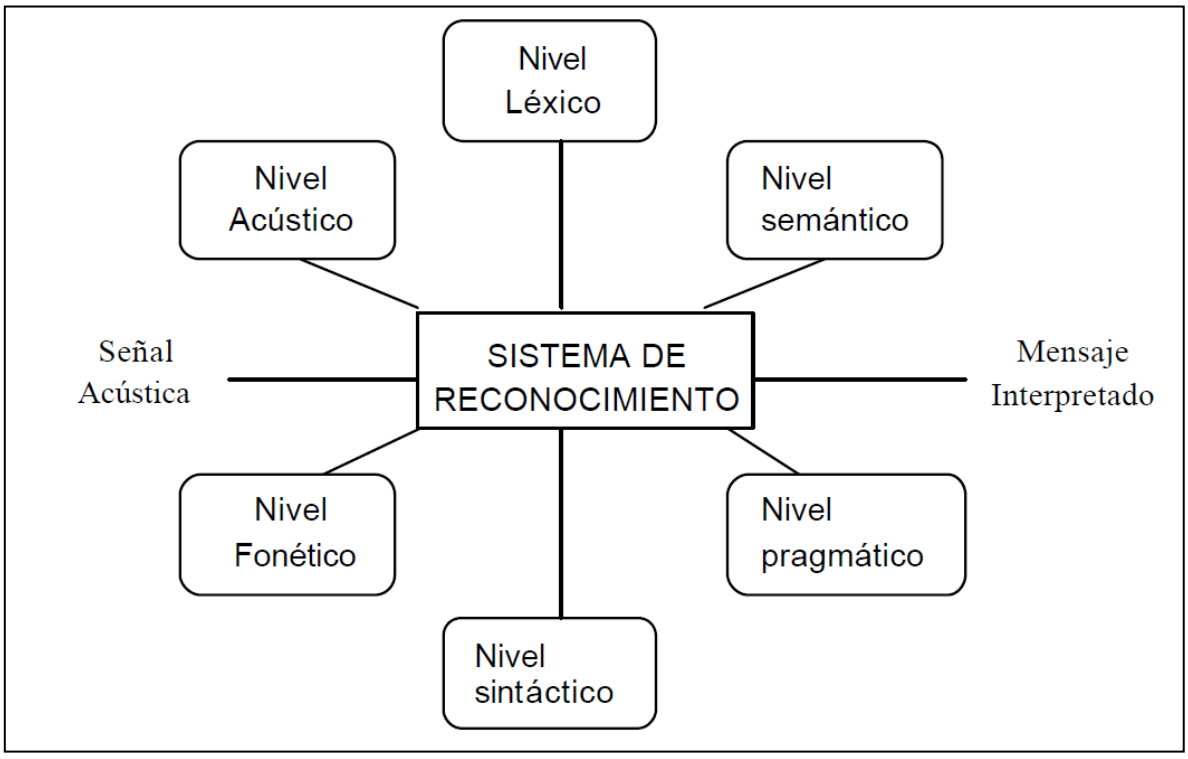


Fig. 3.1: Diferentes niveles presentes en el reconocimiento de mensajes. Tomado de [8].

3.2.3 Restricciones de sistemas de RAH

Debido a la complejidad de los sistemas de RAH, se ha buscado restringir los sistemas para facilitar el proceso, y eventualmente ir aumentando el nivel de complejidad de los sistemas. Navarro en [8] clasifica las diferentes restricciones de sistemas de RAH.

Dependencia del locutor

La diferencia del locutor se define de acuerdo a la necesidad de un entrenamiento previo de parte del locutor o usuario en el sistema. De esta dependencia se distinguen tres sistemas:

- Monolocutores, diseñados para el funcionamiento de un solo locutor.
- Multilocutores, los cuales se componen de un conjunto restringido de locutores
- independientes de locutor los cuales no requieren entrenamiento de parte del locutor para ser identificado en el sistema, es decir, no requiere de un entrenamiento previo.

De estos tipos, los sistmas monolocutores y multilocutores obtienen mejores tasas de reconocimiento, pero requieren de un entrenamiento mas intensivo, especialmente cuando se desea agregar un nuevo usuario.

Tipo de habla

Es posible clasificar los sistemas de RAH de acuerdo a la forma en que el locutor ha de pronunciar las palabras. Estos pueden ser clasificados como:

- Sistemas de palabras aisladas, en los que el locutor es condicionado a pronunciar las palabras con una separación mayor a los 300ms.
- Sistemas de palabras conectadas, donde el locutor puede pronunciar las palabras de manera mas fluida, pero cuidando que haya una pequeña diferencia notable entre el final de una palabra y el inicio de otra
- Sistemas de habla continua, donde al locutor se permite hablar sin restricciones y el sistema debe ser capaz de interpretar el mensaje tal y como es producido por una persona cuando se comunica con us semejantes. Estos sistemas incrementan notablemente la dificultad de los sistemas de RAH, especialmente en el segmentado de palabras.

Talla del Léxico

Los sistemas de reconocimiento, dependiendo del número de palabras del vocabulario, se pueden clasificar en pequeños, medianos y grandes, según tengan decenas, centenas o más de mil palabras, respectivamente.

El problema principal que aparece conforme crece el vocabulario es el de la confusión entre palabras, que incrementa las tasas de error del sistema. Por otro lado, en el caso de pequeños vocabularios cada palabra puede modelarse individualmente, ya que es razonable esperar suficientes datos para entrenar cada palabra, y es posible almacenar los parámetros de cada modelo de palabra separadamente.

3.3 Corpus

El término corpus de habla, según [5], hace referencia a colecciones de grabaciones de habla digitales junto con (aunque no forzosamente) anotaciones, metadatos y documentación. Los corpus de habla son la fuente principal de datos e información para la investigación, ya sea básica o aplicada, y desarrollo de tecnología en el área del Procesamiento digital de la voz..

Hernández define corpus de habla (Speech Corpus) como: "señales de tiempo físico, en la mayoría de los casos de presión de sonido u otras señales de tiempo medibles grabadas desde el acto de hablar, y a su vez asociadas con un conjunto de anotaciones, meta datos y/o documentación almacenados en un medio digital".

3.3.1 Ruido

¿Qué es?

El ruido puede ser descrito, según se menciona en [20], como toda señal no deseada que se superpone con la señal deseada e interfiere con el proceso de medida y análisis. El ruido puede provenir de diferentes fuentes, ya sea del instrumento de medición (micrófono), del medio de transmisión, o por interferencias externas, entre otras.

Técnicas de reducción de ruido

Si bien existen diversas maneras de reducir el ruido de una señal, se hablarán de las técnicas más frecuentemente utilizadas de reducción de ruido en una señal ya capturada [18].

Las técnicas más populares de reducción de ruido realizan una limpieza del habla contaminada luego de efectuar una transformación sobre la señal. Usualmente se recurre a técnicas de transformación tales como la Transformada Discreta de Fourier (DFT), la Transformada de Karhunen-Loeve, la Transformada de Coseno Discreta (DCT), y la Transformada de Paquetes de Wavelets.

Las estrategias que siguen las diferentes técnicas de reducción de ruido pueden caracterizarse de acuerdo al enfoque seguido para suprimir el mismo. Así, se pueden distinguir las técnicas de sustracción espectral (el cual substrae de la señal contaminada las componentes del ruido), las técnicas de filtrado óptimo y adaptativo (donde se estima una señal limpia y se compara con la obtenida a fín de minimizar la diferencia entre ambas) y las técnicas de estimación de señal limpia usando modelos estadísticos. Es importante mencionar que estas técnicas requieren conocer las características del ruido, y por tanto, esperan que se tenga una señal de “silencio”, donde se pueda apreciar el ruido y pueda ser analizado independientemente.

3.4 Caracterización de la señal

3.4.1 Coeficientes Cepstrales en Escala de Mel

Uno de los métodos híbridos de caracterización de la señal que mejores resultados ha mostrado son los Coeficientes Cepstrales en Escala de Mel (MFCC, por sus siglas en inglés). Este método es producto del análisis cepstral de la señal de voz (técnica no paramétrica) y una transformación lineal de los componentes de la señal de acuerdo a una escala basada en la sensibilidad del oído humano a ciertas frecuencias. El procedimiento general [38] para obtener los MFCC se muestran a continuación:

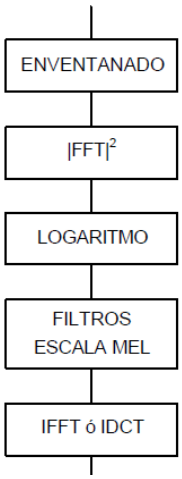


Fig. 3.2: Procedimiento general de cálculo de coeficientes MFCC. Tomado de [8]

Ventaneo

Debido a que muchas de las herramientas de transformación de una señal aplican para señales periódicas, y la voz es una señal no periódica, se aplica la estrategia de ventaneo [38], el cual consiste en segmentar la señal en diferentes "ventanas" y considerar cada una de ellas como parte de una señal periódica. De esta manera, se tendría que aplicar el resto del procedimiento tantas veces como ventanas se tengan.

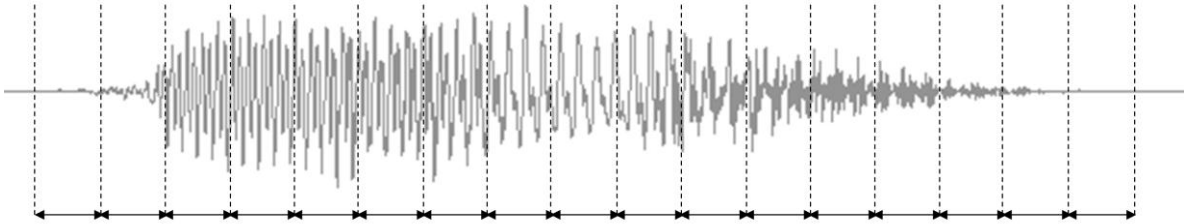


Fig. 3.3: Ejemplo de una señal segmentada en ventanas.

Segmentar la señal en ventanas genera un corte abrupto en los extremos de las mismas, lo que puede ocasionar problemas en el cálculo de las componentes de la señal, por lo que generalmente se recurren a técnicas de tratamiento de estas ventanas.

Una de estas técnicas consiste en la aplicación de un filtro Hamming. Un filtro Hamming consiste en una función cosenoidal el cual se multiplica a una ventana dada. El resultado final consiste en una ventana cuyos se han suavizado para hacer mas representativa la información ubicada al centro de la misma.

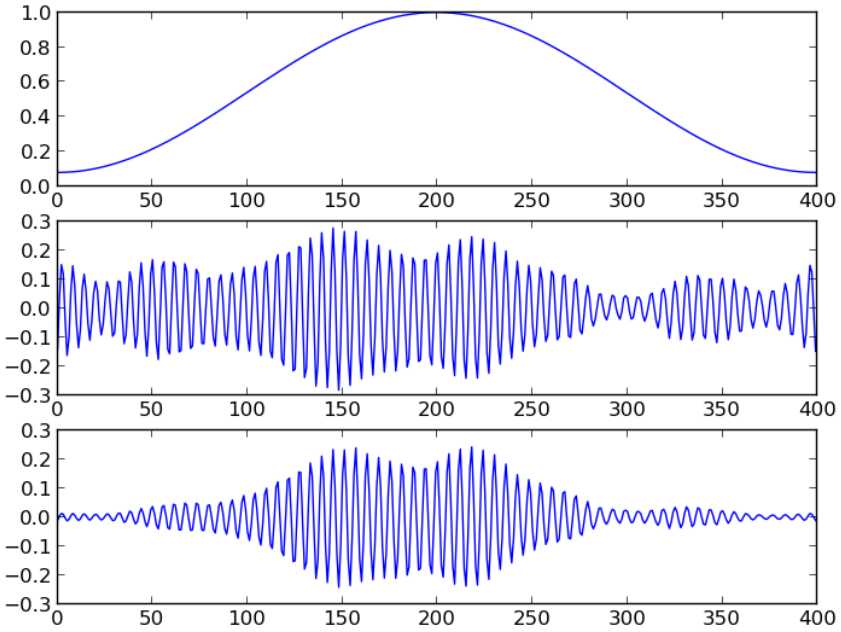


Fig. 3.4: Filtro Hamming, señal de entrada y señal filtrada. Tomado de [38].

Transformada de Fourier

Es una herramienta matemática que permite representar una señal periódica que está en función del tiempo a un conjunto de frecuencias (armónicas) que la componen. La transformada entrega los coeficientes de cada armónica en una parte real y una parte imaginaria. Lo importante para el cálculo de los coeficientes MFCC es la magnitud de cada armónica [38], por lo que se eleva al cuadrado tanto la parte real como la imaginaria y se suman. De esta manera se tiene un coeficiente real para cada armónica. Generalmente se utiliza la trasformada rápida de Fourier (FFT) para esta operación

Filtro Mel y escala logarítmica

Consiste en un conjunto de filtros de forma logarítmica [38] que representan la manera en la que el oído escucha las diferentes frecuencias (escucha más las de baja frecuencia y menos las de alta frecuencia). Los filtros son un conjunto de coeficientes que se multiplican a los valores de las armónicas para aumentar o reducir su importancia en el cálculo de los MFCC.

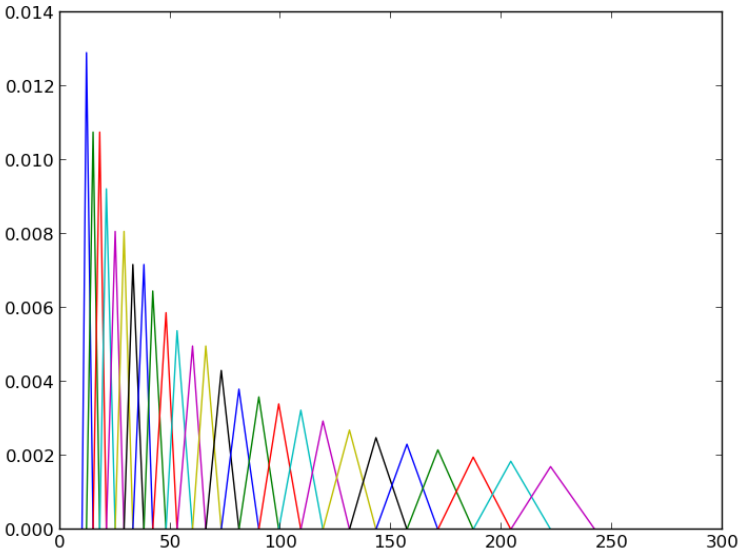


Fig. 3.5: Conjunto de filtros Mel. Tomado de [38].

Transformada de Coseno Discreta

Los coeficientes obtenidos del filtrado Mel pueden ser representados en una gráfica, y después se aplicará una segunda transformación, en este caso utilizando la transformada de coseno discreta. El propósito de esta segunda transformación es convertir los coeficientes en elementos aún más informativos. Los coeficientes obtenidos de esta segunda transformación son infinitos, pero generalmente se toman 12 coeficientes, del segundo al treceavo, que son los más representativos de la señal de habla [38].

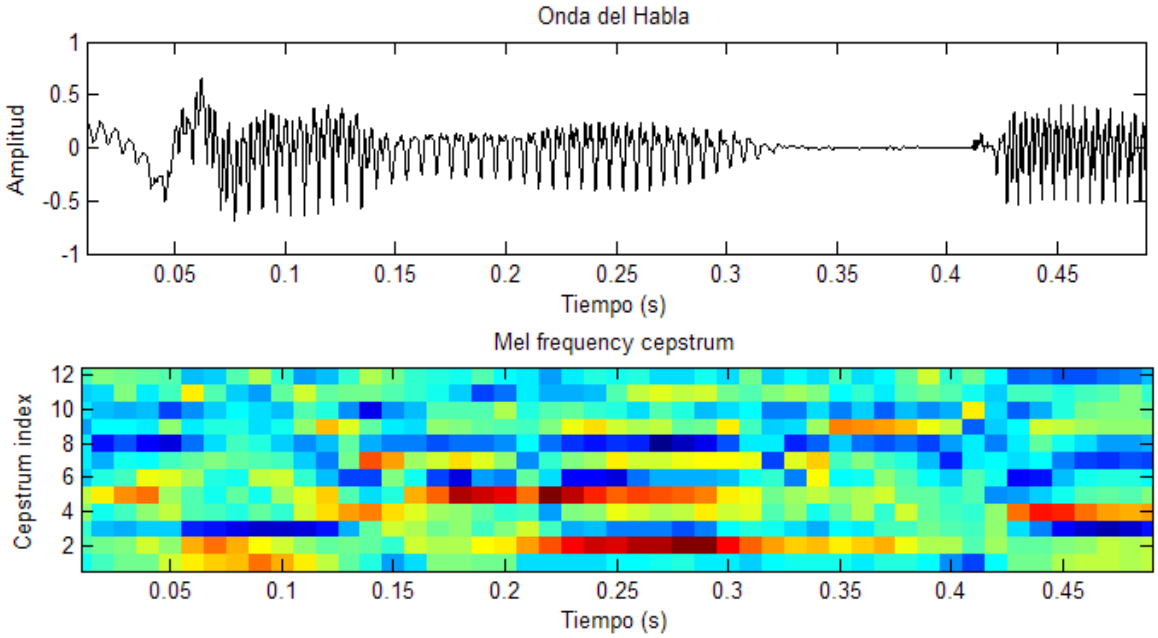


Fig. 3.6: Señal de voz de entrada y espectro de MFCC, del 1 al 12 de la señal.

3.5 Clasificación

3.5.1 Distancias euclidianas

Esta técnica de clasificación es una de las más sencillas de implementar [41]. Dado un conjunto de entrenamiento, se obtiene el centroide de cada clasificación al promediar cada una de las características de cada elemento perteneciente a tal clasificación. Al contar con el centroide de cada clasificación diferente, un elemento a clasificar será comparado por medio de distancias euclidianas con cada centroide. La distancia menor será la que indique la clasificación a la cual pertenece este elemento.

Para realizar una clasificación por distancias euclidianas, según [41] es necesario primero obtener el centroide de cada clasificación en el conjunto de entrenamiento. Sea $x_1, x_2, \ldots, x_n$ el conjunto de elementos que pertenecen a una misma clasificación, donde cada x_k contiene un conjunto de i características numéricas observables $a_1^k, a_2^k, \ldots, a_i^k$. El centroide $\bar{x}$ es definido por:

$$\bar{x_j} = \frac{1}{i}\sum_{m=1}^{n} a_j^m \qquad donde\ j = 1,2,\dots,i$$

La distancia euclidiana entre dos puntos $\mathrm{P}(\mathrm{p}_1, \mathrm{p}_2, \dots, \mathrm{p}_\mathrm{n})$ y $\mathrm{Q}(\mathrm{q}_1, \mathrm{q}_2, \dots, \mathrm{q}_\mathrm{n})$se define como:

$$(d_E(P,Q))^2 = (p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2$$

Para clasificar un nuevo elemento se define de la siguiente manera: Sea P un elemento a clasificar y sean n centroides $\bar{\mathrm{x}}^\mathrm{k}, \mathrm{donde\ k} = 1,2,\dots,\mathrm{n}.$ La clasificación es dada por:

$$class = \min_k\left(d_E(P, \bar{x}^k)\right) \qquad donde\ k = 1,2,\dots,n$$

3.5.2 Árboles de desición

Un árbol de decisión consiste en un conjunto de condiciones sobre las cuales un elemento es clasificado de acuerdo a condiciones verdadero/falso. La construcción de los árboles de decisión puede ser realizada por diferentes algoritmos conocidos (tal como J48 o ID3).

3.5.3 Naive Bayes

Las redes bayesianas [21], junto con los árboles de decisión y las redes neuronales artificiales, han sido los tres métodos más usados en aprendizaje automático durante estos últimos años en tareas como la clasificación de documentos o filtros de mensajes de correo electrónico.

El sistema de clasificación de Naive Bayes, como cualquier sistema de clasificación de patrones se basa en lo siguiente: dado un conjunto de datos (divididos en dos conjuntos de entrenamiento y de prueba) representados por pares (atributo, valor) el problema consiste en encontrar una función f(x) (llamada hipótesis) que clasifique dichos conjuntos.

Entre las características que poseen los métodos bayesianos en tareas de aprendizaje se pueden resaltar los siguientes:

- Cada ejemplo observado va a modificar la probabilidad de que la hipótesis formulada sea correcta (aumentándola o disminuyéndola). Es decir, una hipótesis que no concuerda con un conjunto de ejemplos más o menos grande no es desechada por completo sino que lo que harán será disminuir esa probabilidad estimada para la hipótesis
- Estos métodos son robustos al posible ruido presentes en los ejemplos de entrenamiento y a la posibilidad de tener entre esos ejemplos de entrenamiento datos incompletos o posiblemente erróneos.
- Los métodos bayesianos permiten tener en cuenta en la predicción de la hipótesis el conocimiento a prior o conocimiento del dominio en forma de probabilidades.

Procedimiento

Según [37], el procedimiento para realizar la clasificación por Naive Bayes se describe a continuación:

Dado un conjunto de atributos $\mathrm{F}_1, \mathrm{F}_2, \ldots, \mathrm{F}_\mathrm{n}$ y un conjunto de clasificaciones $\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_\mathrm{m}$, se define una fórmula que permita determinar el porcentaje de probabilidad que el conjunto de atributos pertenezca a una clasificación C_j , donde $1 \leq j \leq m$. Se tiene entonces el siguiente enunciado de probabilidad:

$$p\left(C = C_j \middle| F_1, F_2, \ldots, F_n\right)$$

Usando el teorema de Bayes, es posible descomponer el enunciado en:

$$p\left(C = C_j \middle| F_1, F_2, \ldots, F_n\right) = \frac{p\left(C = C_j\right) p(F_1, F_2, \ldots, F_n \,|C = C_j)}{p(F_1, F_2, \ldots, F_n)}$$

Lo posterior es aplicar la regla de la cadena al numerador. Como recordatorio, la regla de la cadena se describe a continuación:

$$p(A_1, A_2, \dots, A_n) = p(A_1)\, p(A_2, \dots, A_n | A_1)$$

Ahora, se aplica la regla de la cadena al numerador como se explica a continuación:

$$p\big(C = C_j\big)\, p(F_1, F_2, \dots, F_n \,|C = C_j)$$

$$= p\big(C = C_j\big)[\, p(F_1|C = C_j)\;\; p(F_2, \dots, F_n \,|C, F_1)\,]$$

$$= p\big(C = C_j\big)[\, p(F_1|C = C_j)\, p(F_2|C = C_j, F_1)\, p(F_3, \dots, F_n \,|C = C_j, F_1, F_2)\,]$$

…

$$= p\big(C = C_j\big)[\, p(F_1|C = C_j)\, p(F_2|C = C_j, F_1) \dots p(F_n \,|C = C_j, F_1, F_2, \dots, F_{n-1})\,]$$

A continuación se realiza el supuesto “ingenuo” de independencia de los atributos. Se asume que cada F_i es independiente de cualquier otra F_j para toda i≠j , es decir :

$$p(F_i \,|F_j\,) = p(F_i\,)$$

Entonces, aplicando criterio de independencia al numerador:

$$p\big(C = C_j\big)[\, p(F_1|C = C_j)\, p(F_2|C = C_j, F_1) \dots p(F_n \,|C = C_j, F_1, F_2, \dots, F_{n-1})\,]$$

$$= p\big(C = C_j\big)[\, p(F_1|C = C_j)\, p(F_2|C = C_j) \dots p(F_n \,|C = C_j)\,]$$

Entonces el teorema queda como:

$$p\big(C = C_j \big| F_1, F_2, \dots, F_n\big) = \frac{p\big(C = C_j\big)\, p(F_1|C = C_j)\, p(F_2|C = C_j) \dots p(F_n\,|C = C_j)}{p(F_1, F_2, \dots, F_n)}$$

Lo cual se puede simplificar como:

$$p\big(C = C_j \big| F_1, F_2, \dots, F_n\big) = \frac{p\big(C = C_j\big) \prod_{i=1}^{n} p(F_i|C = C_j)}{p(F_1, F_2, \dots, F_n)}$$

Ahora, este enunciado representa la probabilidad de que la clasificación sea C_j. Ahora, la función de clasificación debe de comparar todos los posibles valores de clasificación y determinar aquella con mayor probabilidad como la clasificación correspondiente. Es decir, la función de hipótesis del clasificador se define como:

$$f(x) = \underset{j}{argmax} \left(p\big(C = C_j \big| F_1, F_2, \dots, F_n\big) \right) \quad ; \quad 1 \leq j \leq m$$

Lo cual, al desarrollar todo el argumento, es igual a:

$$f(x) = \underset{j}{argmax} \left(\frac{p\big(C = C_j\big) \prod_{i=1}^{n} p(F_i|C = C_j)}{p(F_1, F_2, \dots, F_n)} \right) \quad ; \quad 1 \leq j \leq m$$

Es importante mencionar que $p(F_1, F_2, \dots, F_n)$ es un valor constante que depende del valor de los atributos y que será equivalente para todo C_j, por lo que es posible retirarlo de la función para simplificarla, lo que terminaría como:

$$f(x) = \underset{j}{argmax}\left(p(C = C_j)\prod_{i=1}^{n} p(F_i|C = C_j)\right) \quad ; \quad 1 \leq j \leq m$$

Naive Bayes con información continua

Si el conjunto de atributos $\mathrm{F}_1, \mathrm{F}_2, \ldots, \mathrm{F}_\mathrm{n}$ pertenecen a un conjunto de de datos contínuos y no discretos, se puede realizar el supuesto de que los atributos siguen una distribución gaussiana (normal) [21]. De esta manera se puede aplicar el siguiente criterio de probabilidad:

$$p(x = v|c) = \frac{1}{\sqrt{2\,\pi\,\sigma_c^2}} * e^{\left(-\frac{(v-\mu_c)^2}{2\,\pi\,\sigma_c^2}\right)}$$

3.6 Modelos Ocultos de Markov

Los modelos ocultos de Markov [14] (HMM) constituyen una de las técnicas que se ha utilizado con más éxito en el reconocimiento automático del habla (RAH). Principalmente, esta técnica ha permitido modelar adecuadamente la gran variabilidad en el tiempo de la señal de voz.

Una notación habitual de un HMM es la representación como una quintupla $(\mathrm{Q}, \mathrm{V}, \pi, \mathrm{A}, \mathrm{B})$:

- El conjunto de estados $Q = \{1,2, \ldots, N\}$. El estado definido en un instante dado se denota como q_t.
- El conjunto de posibles valores $V = \{v_1, v_2, \ldots, v_M\}$ observables en cada estado. M es el número de observacionesvistas y cada v_k hace referencia a una observación distinta.
- 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. $a_{ij} = P(q_t = j\,|q_{t-1} = i\,)$, 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 encontraba en el estado i .

- El conjunto de probabilidades $B = \{b_j(v_k)\}$ de las observaciones. Se define $b_j(v_k) = P(o_t = v_k \,|q_t = j\,)$, es decir, la probabilidad de observar v_k cuando se está en el estado j en el instante t .La secuencia de observaciones se denota como un conjunto $O = \{o_1, o_2, \dots, o_T\}$.

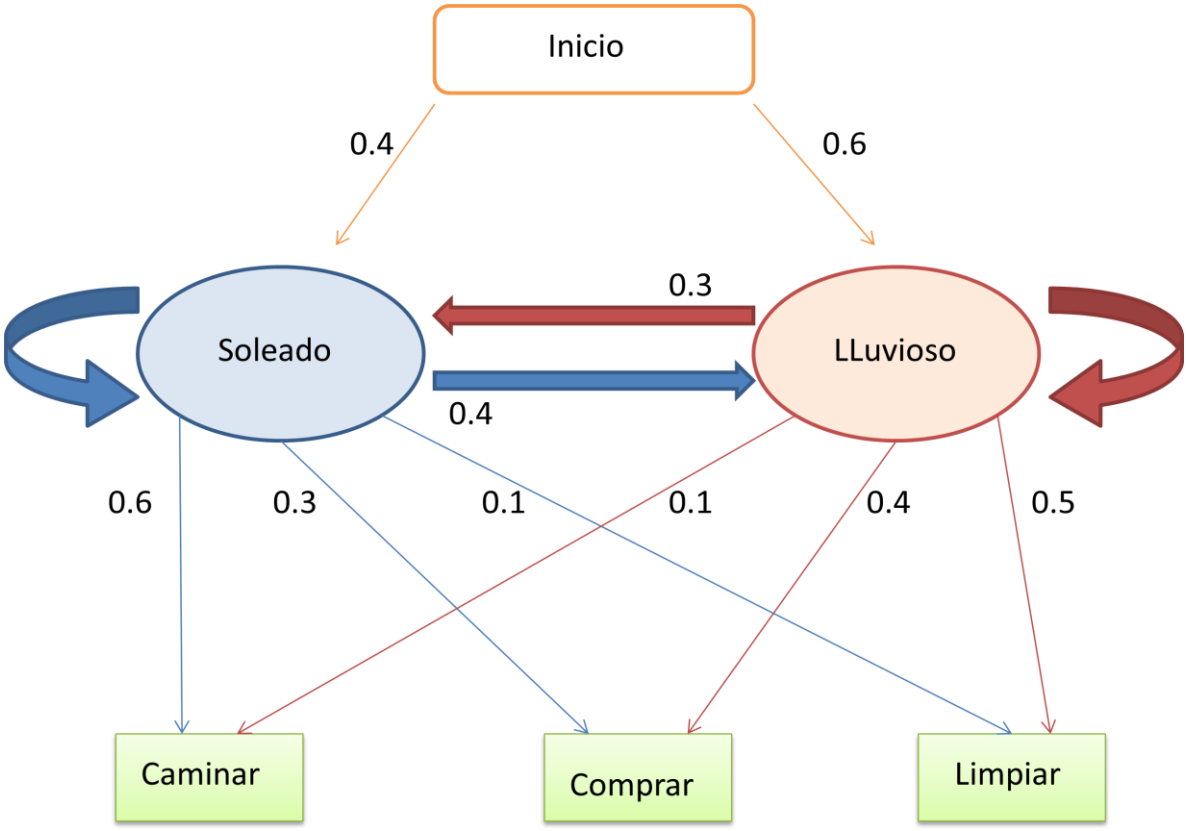


Fig. 3.7: Ejemplo de un HMM.

3.6.1 Procedimiento Backward-Foward

El problema a resolver para el reconocimiento del habla utilizando HMM es el siguiente: dada una secuencia de observaciones $O = (o_1, o_2, \dots, o_n)$, ¿Cuál es la probabilidad de que el modelo dado genere la secuencia O?

Rabiner menciona en [14] un procedimiento llamado “El procedimiento foward-backward”, el cual soluciona (entre otros) el problema a resolver para la identificación del habla.

Para resolver el problema mencionado anteriormente, se utilizará el concepto “foward” del procedimiento. Se tiene entonces, una variable foward descrita a continuación:

$$\alpha_t(i) = P(o_1 o_2 \dots o_t, q_t = S_i|\lambda)$$

Esto es, la probabilidad de que se genere la observación parcial $o_1 o_2 \dots o_t$ (hasta el tiempo t) y que se encuentre en el estado S_i en el tiempo t, dado el modelo λ. Se puede encontrar $\alpha_t(i)$ de manera inductiva, como se muestra a continuación (Es importante recordar que π, a y b están dados por el modelo λ):

1) Inicialización

$$\alpha_1(i) = \pi_i b_i(o_1) \qquad 1 \leq i \leq N$$

2) Inducción

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}) \qquad 1 \leq t \leq (T-1) \quad ; \quad 1 \leq j \leq N$$

3) Finalización

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

Para resolver el problema, se realiza el procedimiento siguiente:

1) Con la fórmula de inicialización, se obtiene la probabilidad $\alpha_1(i)$ de cada estado i posible de S
2) Para obtener $\alpha_{t+1}(i)$, se utiliza la fórmula de inducción. Se considera que ya se cuenta con el α_t de cada estado i posible de S dada la fórmula de inicialización, o una aplicación anterior de la fórmula de inducción. Otra vez, esta fórmula se aplica para cada estado i posible de S

3) Se repite la aplicación de la fórmula de inducción hasta encontrar $\alpha_T(i)$ de cada estado i posible de S
4) La probabilidad, entonces, de que se genere la secuencia O dado el modelo es la sumatoria de todos los $\alpha_T(i)$

3.7 Lenguas Autóctonas

Las poblaciones indígenas tienen usos y costumbres propias. Poseen formas particulares de comprender el mundo y de interactuar con él. Visten, comen, celebran sus festividades, conviven y nombran a sus propias autoridades de acuerdo a esa concepción que tienen de la vida. Un elemento muy importante que los distingue y les da identidad, es la lengua con la que se comunican.

3.7.1 Familias, agrupaciones y variantes

El INALI, en su "Catálogo de lenguas indígenas mexicanas" [26], divide las diferentes lenguas autóctonas de México en 3 niveles: familias, agrupaciones y variantes.

Familias

Una familia es la familia de mayor nivel dentro de la clasificación. Se define como un conjunto de lenguas cuyas semejanzas estructurales y léxicas se deben a un origen histórico común. En el catálogo se identificaron 11 familias distintas.

Álgica.
Yuto-nahua.
Cochimí-yumana.
Seri.
Oto-mangue.
Maya.
Totonaco-tepehua.
Tarasca.
Mixe-zoque.
Chontal de Oaxaca.
Huave.

Tabla 3.8: Familias lingüísticas.

Agrupaciones

Una agrupación representa el segundo nivel de clasificación, debajo de la familia. Se define como el conjunto de variantes lingüísticas comprendidas bajo el nombre dado históricamente a un pueblo indígena. En este nivel fueron identificadas 68 diferentes agrupaciones.

Akateko	Amuzgo	Awakateko	Ayapaneco	Cora
Cucapá	Cuicateco	Chatino	Chichimeco jonaz	Chinanteco
Chocholteco	Chontal de Oaxaca	Chontal de Tabasco	Chuj	Ch'ol
Guarijío	Huasteco	Huave	Huichol	Ixcateco
Ixil	Jakalteko	Kaqchikel	Kickapoo	Kiliwa
Kumiai	Ku'ahl	K'iche'	Lacandón	Mam
Matlatzinca	Maya	Mayo	Mazahua	Mazateco
Mixe	Mixteco	Náhuatl	Oluteco	Otomí
Paipai	Pame	Pápago	Pima	Popoloca
Popoluca de la Sierra	Qato'k	Q'anjob'al	Q'eqchí '	Sayulteco
Seri	Tarahumara	Tarasco	Teko	Tepehua
Tepehuano del norte	Tepehuano del sur	Texistepequeño	Tlahuica	Tlapaneco
Tojolabal	Totonaco	Triqui	Tseltal	Tsotsil
Yaqui	Zapoteco	Zoque		

Tabla 3.9: Agrupaciones lingüísticas.

Variantes

Una variante representa el último nivel de clasificación. Se define como una forma de habla que: a) presenta diferencias estructurales y léxicas en comparación con otras variantes de la misma agrupación lingüística; y b) implica para sus usuarios una determinada identidad sociolingüística, que se diferencia de la identidad sociolingüística de los usuarios de otras variantes. El catálogo identifica, por último 364 diferentes variantes. Es importante mencionar que, de acuerdo a lo especificado por el INALI, cada variante lingüística debe ser considerada como una lengua independiente.

3.7.2 Tének

Tének es una auto denominación de la agrupación lingüística huasteco, de la familia lingüística maya. Se encuentra principalmente en los estados de Tamaulipas, Veracruz y San Luis Potosí.

Figura 3.10: Mapa demográfico de algunas agrupaciones lingüísticas.

Alfabeto

Su alfabeto, según el vocabulario huasteco de San Luis Potosí [3], consiste de 33 símbolos, como se muestran a continuación:

a	**ajan**	*elote*	**o**	ot'el	*ardilla*
ā	ām	*araña*	**ō**	ōt'	*cuero*
b	bacan	*tortilla*	**p**	pacab	*caña*
c	calām	*mañana*	**q**	quetēl	*sentado*
c'	c'alam	*calabaza*	**q'**	q'uīth	*espina*
ch	chuch	*coyote*	**r**	c'ororo'	*aguardiente*
ch'	ch'uchub	*dedo*	**t**	tan	*canoa*
e	ejat	*vivo*	**t'**	t'ele'	*niño*
ē	elēb	*afuera*	**th**	thabal	*dueño*
g	Pēgru	*Pedro*	**ts**	tsan	*culebra*
h	hui'	boca	**ts'**	ts'ēn	*sierra*
i	ic'	*viento*	**u**	uts'	*piojo*
ī	īt	*nuevo*	**ū**	ūt'	*tlacuache*
j	ja'	*agua*	**x**	xomom	*guaje*
l	lem	*mariposa*	**y**	yōy	*zancudo*
m	mām	*abuelo*	**'**	to'ol	*pescado*
n	nanā'	*yo*			

Tabla 3.11: Alfabeto Tének.

Números del 1 al 10

A continuación se muestra la escritura de los números en tének del 1 al 10, según el vocabulario huasteco:

jūn	*uno*	acac	*seis*
tsāb	*dos*	būc	*siete*
ōx	*tres*	huaxic	*ocho*
tsē'	*cuatro*	belēu	*nueve*
bō'	*cinco*	lājuj	*diez*

Tabla 3.12: Números en Tének del 1 al 10.

4
Metodología

Se han definido entonces el problema, los objetivos a abarcar, el estado del arte y la información teórica relativa a este proyecto. A continuación se describe, entonces, el proceso que se ha de llevar a cabo para alcanzar los objetivos propuestos.

Además de la definición de la metodología, se enunciarán aquellos elementos importantes en el desarrollo del sistema de RAH, tales como el entrenamiento y el tratamiento de las señales.

4.1 Consideraciones previas

4.1.1 Corpus

Para realizar el proceso de RAH, es necesario contar con un corpus suficientemente grande para obtener dos conjuntos de muestras: el conjunto de entrenamiento (que será utilizado para definir los valores que se entrenarán en los clasificadores y modelos) y el conjunto de pruebas (que será utilizado para probar la efectividad del sistema).

Muestreo

El corpus utilizado en este proyecto consta de:

- Un grupo de 6 personas diferentes
- A cada persona se pidió que pronunciaran los números del 1 al 10 en tének

- Se pidió que cada número lo pronunciaran 10 veces

Para un total de 600 diferentes archivos de pronunciación de palabras en Tének.

Durante el muestreo para el corpus se consideró lo siguiente, basado en la estrategia de muestreo de Hernández [5]:

- Las condiciones de ruido deben ser lo más cercanas a nulas. Es decir, tratar de que en la grabación de los archivos del corpus no existiera ruido presente
- Separación entre la pronunciación de las palabras
- Frecuencia de muestreo de 44100 Hz
- Formato de muestra pcm
- Bit Depth de 16 bits
- Formato de archivo WAV
- Sonido Monoaural

4.1.2 Tratamiento de la señal

Filtrado de ruido

Uno de los mayores problemas al trabajar con señales de voz es la presencia de ruido (especialmente si se quiere trabajar con muestras libres de la misma). El filtrado [20] consiste en eliminar tales muestras de ruido de las señales, y aislar las palabras para obtener archivos de sonido idóneos para realizar el modelado.

Para eliminar el ruido se utilizó el software Audacity [43], que incorpora un método de reducción de ruido por identificación de sus componentes armónicos.

En el siguiente ejemplo, se puede apreciar una señal de voz, la cual contiene ruido de ambiente. El ruido puede ser apreciado en la parte final de la señal, el cual debe ser identificado en el software para extraer sus componentes armónicos .

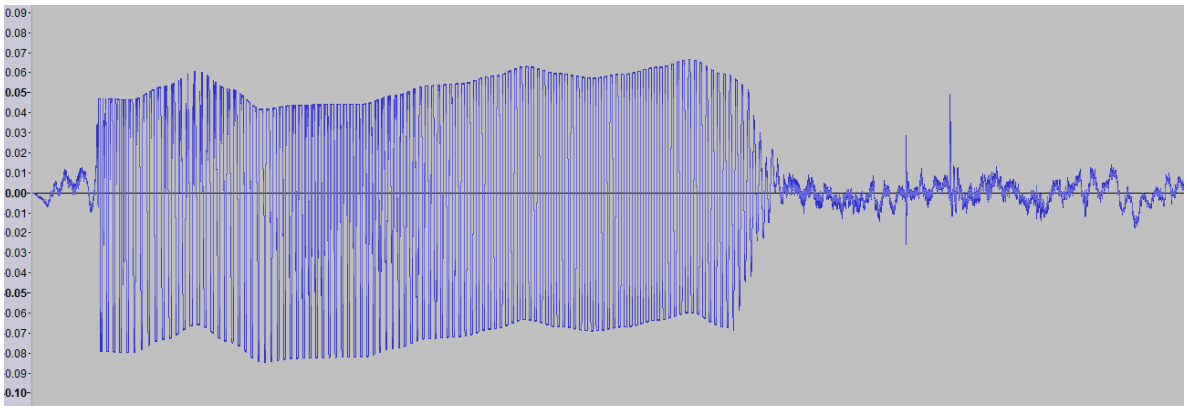


Fig. 4.1: Señal de voz con ruido presente

Para que el software pueda identificar una señal de ruido, es importante contar con la presencia de un momento en todo el archivo de audio donde se presente solamente el ruido (es decir, un momento de "silencio" en el que pueda ser aislado el ruido), con el propósito de no incluir señales que afecten negativamente en la eliminación de ruido.

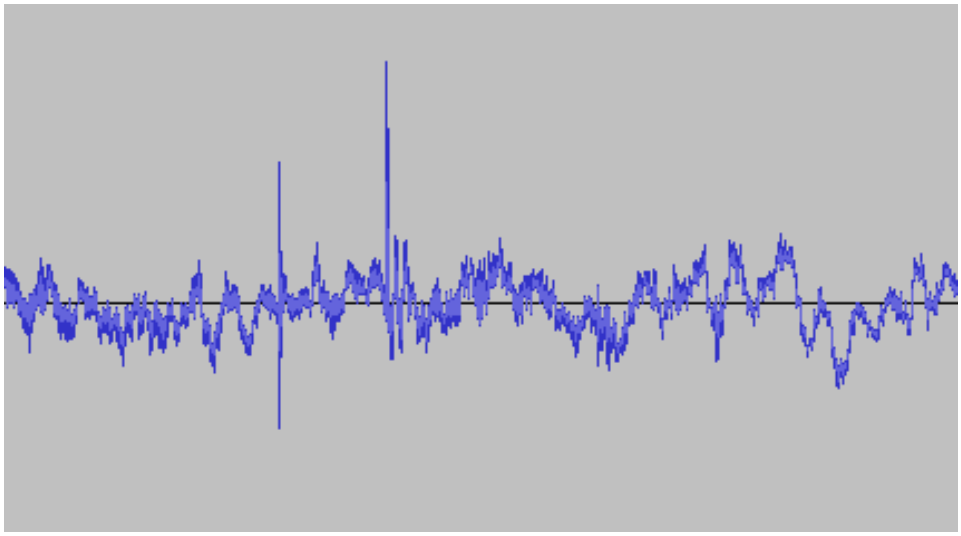


Fig. 4.2: Extracto del ruido presente en la figura 4.1

Con los componentes armónicos de la señal de ruido el software puede entonces substraerlos de toda la señal, generando así una nueva señal de voz la cual reducirá de manera significativa el ruido presente en ella, minimizando el impacto del procedimiento en la integridad original de la señal.

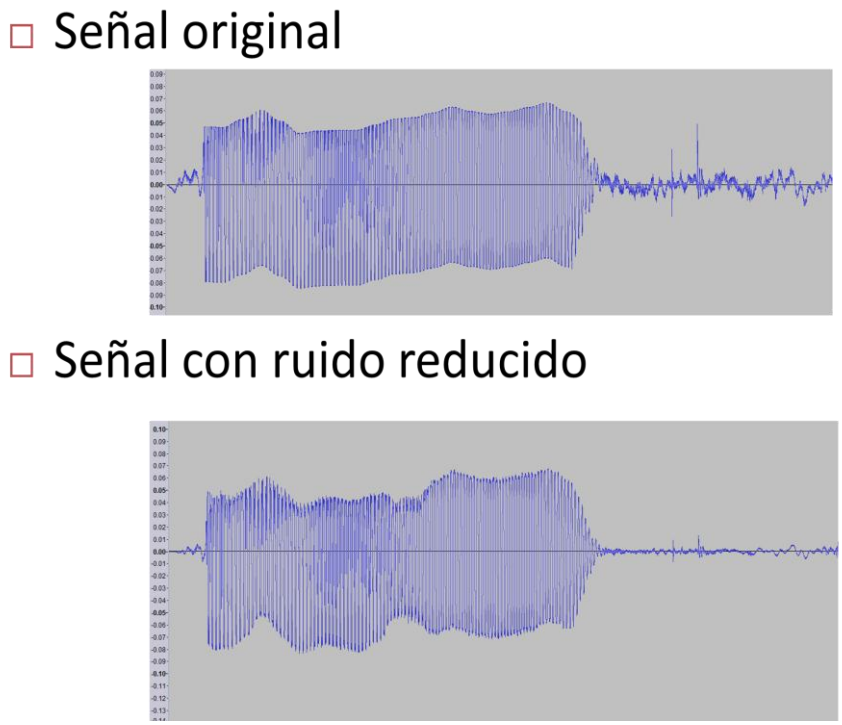


Fig. 4.3: Señal antes y después de la reducción de ruido.

Etiquetado de la señal

El etiquetado de la señal corresponde al procedimiento en el cual manualmente se identifican segmentos de la señal de habla con un fonema pronunciado. Este proceso representa una gran importancia ya que de ello depende el entrenamiento al modelo y, por ende, parte de la efectividad del sistema.

Para realizar el etiquetado de las señales de voz se utilizó el programa Audacity [43], el cual cuenta con herramientas para identificar y segmentar señales, con una precisión de 0.001s. En este caso, se identifican los puntos de la señal en donde se detecta un cambio en la pronunciación, esto mediante el análisis auditivo y visual de la señal. Estos puntos indicarán el inicio y el final de la pronunciación de un fonema.

4.2 Procedimiento general

Para el desarrollo de este sistema de reconocimiento de palabras, se consideró realizar una aproximación fonética al proceso, basado en las características de las aproximaciones proporcionadas por [8]. Además, basado en la estrategia propuesta por [14] se tiene el siguiente esquema general para el proceso de reconocimiento:

- Definición de la estructura fonética a utilizar.
- Segmentación de la señal de voz en las estructuras fonéticas utilizadas.
- Comparación y clasificación de las estructuras fonéticas
- Determinación de un modelo que permita definir un porcentaje de parentesco entre la señal de voz a reconocer y el diccionario de palabras entrenadas previamente.

Este sistema de reconocimiento requiere entonces de un entrenamiento previo, el cual se realiza de una manera general como se muestra a continuación:

- Definición de la estructura fonética a utilizar. Ésta debe ser la misma que la utilizada en el proceso de reconocimiento.
- Definición de un conjunto de señales de voz para ser utilizadas como entrenamiento.
- Clasificación supervisada de las estructuras fonéticas en las señales de entrenamiento.
- Clasificación supervisada de las señales de voz en las diferentes palabras a ser consideradas.
- Aplicación de las estrategias correspondientes (de acuerdo al criterio de clasificación utilizado) para convertir la información obtenida de las estructuras fonéticas en un diccionario fonético.
- Aplicación de las estrategias correspondientes (de acuerdo al tipo de modelo utilizado) para determinar el entrenamiento del modelo utilizado.

4.3 Metodología utilizada

Basado en el procedimiento general descrito anteriormente, se tomaron las siguientes consideraciones:

4.3.1 Estructura fonética.

La selección de la estructura fonética es de gran importancia, pues define el nivel de inmersión que tendrá el sistema en la detección de patrones [8]. La estructura fonética se define entonces como el elemento más pequeño que será considerado por el sistema de reconocimiento del habla, el cual puede ser desde tan general como estructurar de acuerdo a palabras o tan particular como considerar la estructura básica como un fonema pronunciado.

En este caso, se considera al fonema como el elemento a analizar en el sistema. Esto lleva entonces la consideración que en la clasificación se generará una cadena de fonemas, que después serán analizadas por un modelo.

4.3.2 Segmentado de la señal

De acuerdo a la unidad fonética considerada, el segmentado de la señal tratará de dividir la misma en las mismas unidades. En el caso del fonema, existe la siguiente consideración [8]:

Dada una señal de voz, los fonemas pronunciados en la misma no se presentan en intervalos regulares, sino que algunos fonemas se pronuncian durante un periodo mayor en el tiempo que otros.

Para solucionar este problema, surgen entonces dos consideraciones

- Segmentar la señal en ventanas pequeñas, periódicas, que después puedan ser analizadas y clasificadas, y que generen una cadena de fonemas clasificados tan larga como la cantidad de ventanas generadas, que después puedan ser interpretados en un modelo, adaptado a esta consideración.

- Implementar un segundo procedimiento de detección de cambio de fonema pronunciado en dos momentos de una señal, que permita entonces tener una cantidad menor de fonemas, correspondientes entonces a cada pronunciación de una palabra.

El desarrollo de esta metodología se realizó basado en la primera aproximación, ya que en la implementación del modelado es posible utilizar la cadena completa de fonemas para entregar un resultado favorable.

Definidas entonces las consideraciones importantes del procedimiento general, la metodología utilizada para el proceso de reconocimiento de palabras se muestra a continuación:

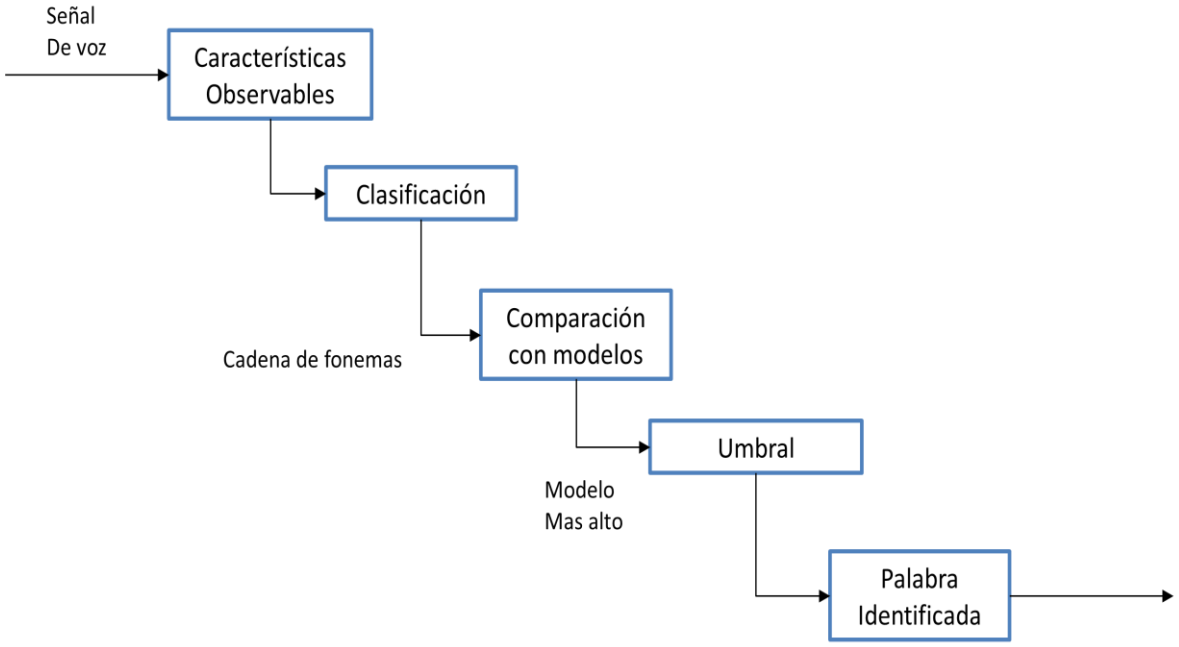


Fig. 4.4: Procedimiento general de la metodología de RAH.

4.3.3 Señal de voz

Es la señal de entrada la cual va a ser identificada. Esta señal necesita ser separada en sus componentes más básicos (fonemas en este caso), para lo cual se realiza una operación de ventaneo. El ventaneo consiste en la separación de una señal en elementos más pequeños, que puedan ser analizados independientemente.

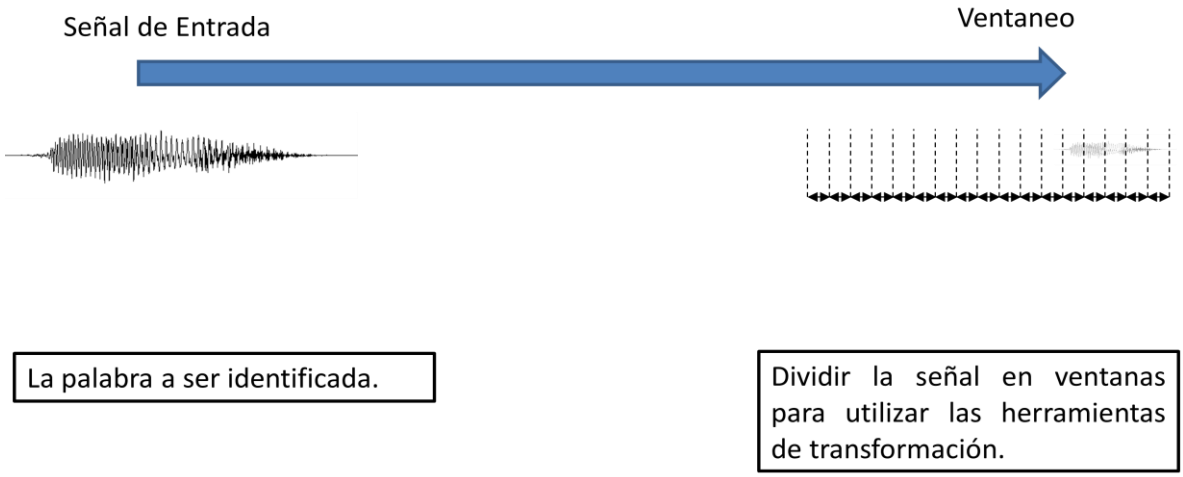


Fig. 4.5: Primera etapa del reconocimiento: Ventaneo.

4.3.4 Características observables

Una vez separada la señal en ventanas, es necesario convertir la señal en un conjunto de valores, los cuales puedan ser interpretados y clasificados por un algoritmo. Generalmente estas características observables son de orden numérico, y debido a la naturaleza de las señales de audio, es posible utilizar métodos híbridos para obtener coeficientes que ayuden a la clasificación. Se eligió utilizar MFCC como herramienta de cálculo de estas características, debido a sus previos resultados satisfactorios.

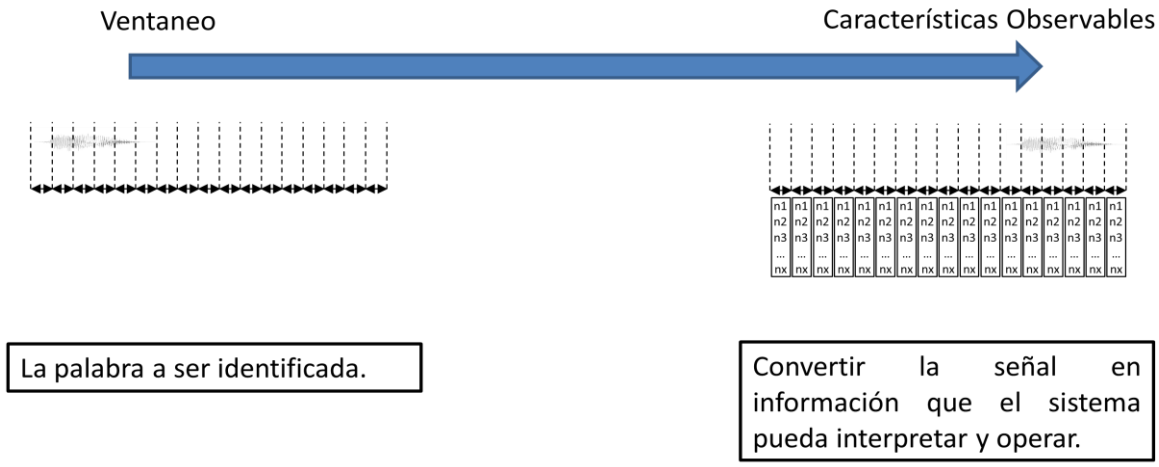


Fig. 4.6: Segunda etapa del reconocimiento: Extracción de características.

Coeficientes Cepstrales en Frecuencia de Mel

Como se mencionó anteriormente, los MFCC han obtenido buenos resultados en el desarrollo de sistemas de RAH [4], por lo tanto, han sido seleccionado como la estrategia para la extracción de características observables para el desarrollo de esta metodología. Como recordatorio, el proceso general de cálculo de MFCC se muestra en la figura 4.7:

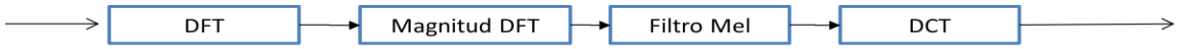


Fig. 4.7: Esquema general del proceso de cálculo de MFCC.

Algoritmo de Wojcicki

La implementación del proceso de extracción de coeficientes MFCC de la señal es una adaptación del algoritmo planteado por Kamil Wojcicki [40] en 2011.

El algoritmo de Wojcicki obtiene el espectro cepstral mel de la siguiente manera:

1. La señal es dividida en ventanas.
2. Aplicación de un filtro FIR a cada ventana de la señal de audio.

3. A cada ventana se le aplica un análisis por medio de una Transformada de Fourier Discreta.
4. A los valores de la transformada se calcula su magnitud.
5. A continuación se aplica un filtro triangular basado en la escala logarítmica de Mel.
6. A estos valores filtrados se aplica la transformada de coseno discreta.
7. Finalmente se aplica un lifter sinusoidal para producir los coeficientes MFCC.

La adaptación del algoritmo de Wojcicki se es modificado con las siguientes características:

- Para cada ventana de la señal se aplica un filtro de Hamming.
- Cada ventana representa una duración de 20ms de la señal. Sin embargo, entre una ventana y la siguiente hay una sobreposición de 10ms, esto con los siguientes objetivos: primero, representar segmentos de 10ms en toda la señal de voz y segundo, evitar ignorar los extremos de las ventanas, los cuales ya han sido reducidos por los filtros Hamming.
- Se toman 12 coeficientes para representar la señal (del segundo al treceavo)

4.3.5 Clasificación

Con la señal de voz convertida en una matriz de información numérica, es posible aplicar herramientas de reconocimiento de patrones y clasificación para agrupar las distintas formas de onda obtenidas [13]. El propósito principal de este procedimiento radica en la gran variabilidad que puede tener una onda que represente el mismo fonema, el cual se puede ver afectados por factores como el tono del hablante y el volumen de pronunciación, así como otros factores externos como el ruido de fondo.

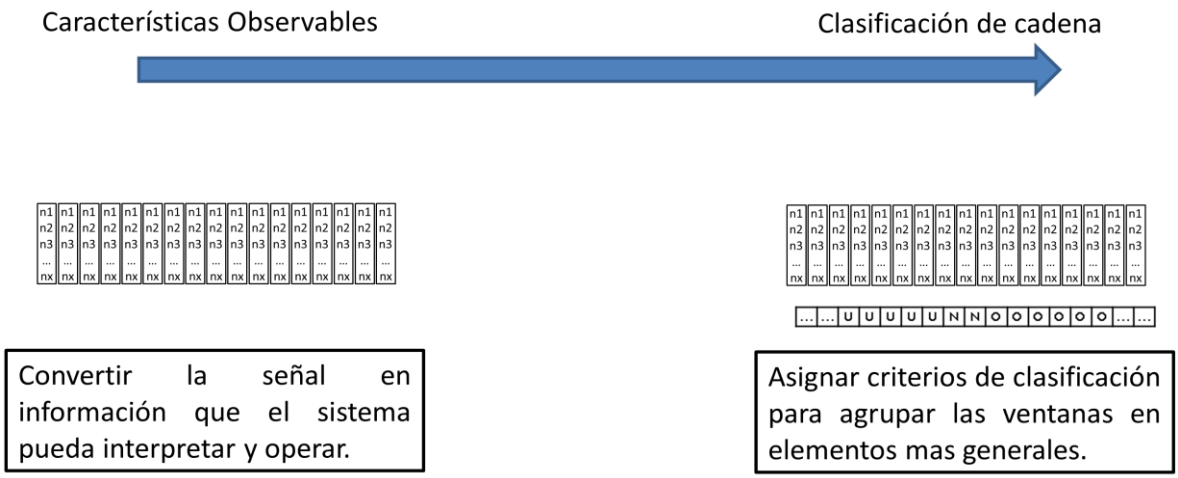


Fig. 4.8: Tercera etapa del reconocimiento; Clasificación.

Existen diferentes técnicas de clasificación, las cuales pueden ser aplicables en este punto del procedimiento. Sin embargo, las utilizadas para este sistema fueron las siguientes:

Distancias euclidianas

Esta técnica representa una manera rápida de llevar a cabo la clasificación de ventanas [41], y su aplicación es muy sencilla (ya que requiere de unos pocos cálculos matemáticos). Es aplicable debido a que las características observables son en realidad múltiples ventanas que son representadas por vectores de 12 valores, los cuales pueden ser ubicados en un espacio de 12 dimensiones.

El entrenamiento para esta técnica también es muy sencillo, puesto que para obtener los valores de entrenamiento de cada clasificación simplemente se agrupan todos aquellos elementos que comparten la misma clasificación, y se calcula el centroide de todos ellos. Este conjunto de centroides obtenidos representan entonces el conjunto entrenado que se aplicará para clasificar los nuevos elementos.

Árbol de decisión

Un árbol de decisión consiste en una serie de reglas que deben ser seguidas para poder obtener una clasificación. Sin embargo, si estas reglas no se encuentran bien definidas, la cantidad de reglas a seguir se incrementa de manera significativa, lo que puede traducirse en un tiempo de ejecución más elevado. También cuenta con el inconveniente de que el entrenamiento de esta técnica, comparada con otras, es mayor, puesto que se tiene que aplicar una serie de fórmulas para definir cada regla del árbol.

El entrenamiento de esta técnica de árbol de decisión se llevó a cabo utilizando el software WEKA [45], la cual contiene estimadores de árboles de decisión. Esta herramienta permite entonces realizar los cálculos de las reglas y generar un archivo que posteriormente puede ser leído para realizar clasificaciones futuras

Naive Bayes

Naive Bayes representa una técnica probabilística, que no solamente considera la media general de la información presentada, sino que también considera la variabilidad de la información contenida en el conjunto de entrenamiento [37], siguiendo el procedimiento probabilístico definido con anterioridad.

El entrenamiento a realizarse en Naive Bayes es similar a distancias euclidianas, considerando que la información consiste en cantidades no discretas. Primero se agrupan todos los elementos cuya clasificación es igual, y para cada atributo se calcula la media y la varianza. Entonces se tiene que para n atributos se tienen 2n valores en el conjunto de entrenamiento, los cuales serán utilizados después en la tarea de clasificación.

4.3.6 Modelado

Ya con una cadena de fonemas identificados, es posible determinar un proceso el cual permita identificar tal cadena como una palabra en específico. El

modelo, entonces, se refiere a la parte del proceso de identificación que se encarga de determinar el más probable de una serie de posibles resultados (en este caso, las palabras) dado una serie de observaciones (el cual se refiere a la cadena de fonemas observados).

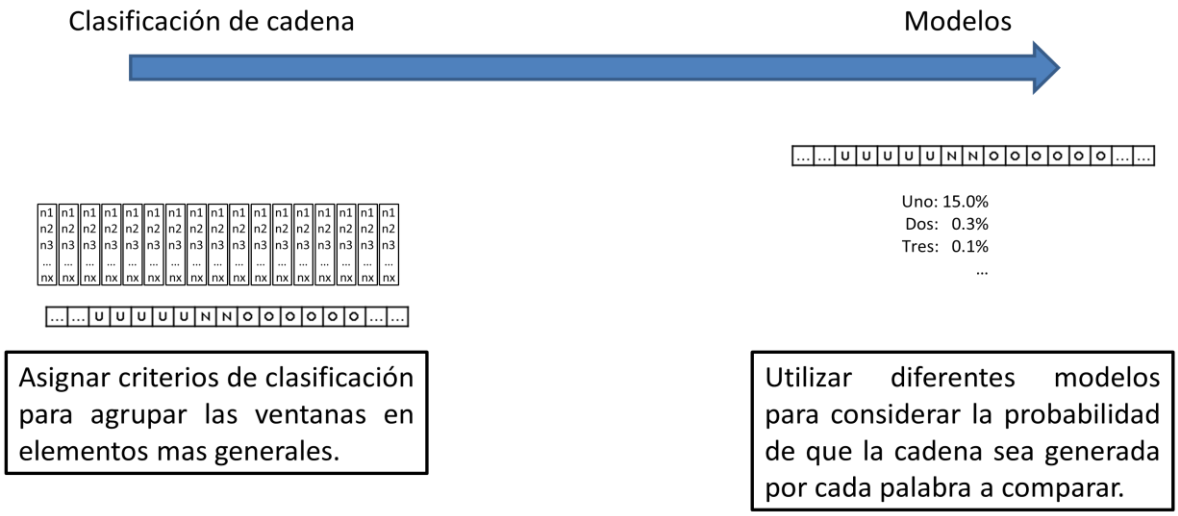


Fig. 4.9: Cuarta etapa del reconocimiento: Comparación con modelos.

Modelos Oculto de Markov

Los Modelos Ocultos de Markov cuentan con herramientas de cálculo que permite identificar y comparar diferentes modelos dada una cadena de observaciones presentadas. Esto, adaptado a la metodología de detección de palabras presentada, otorga de una solución eficaz para complementar el sistema.

El modelo se considera, entonces, como se explica a continuación [41]:

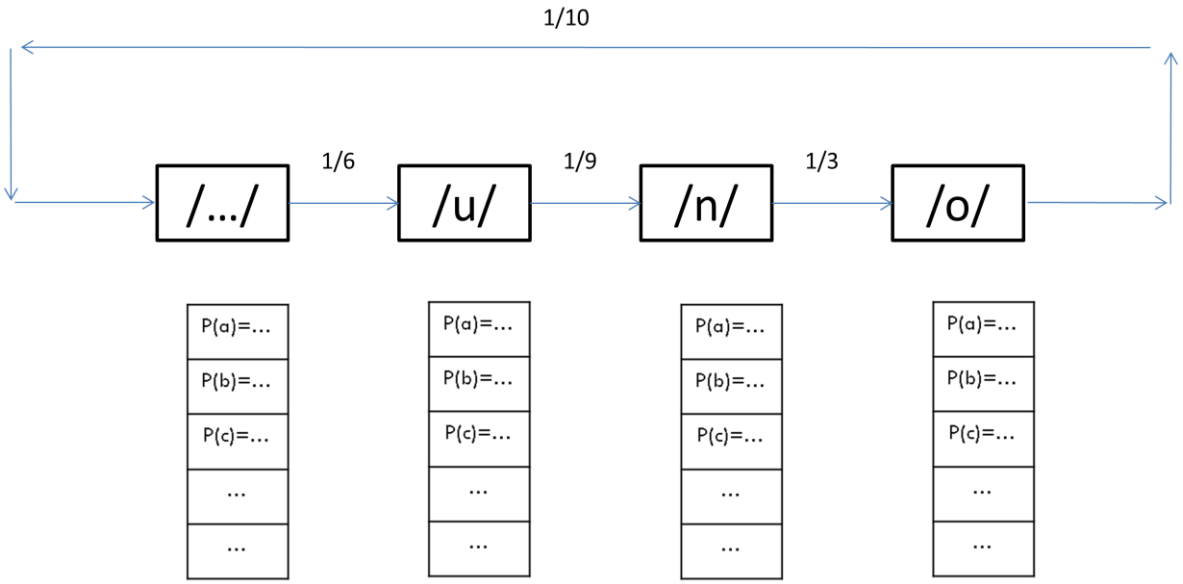


Fig. 4.10: Representación gráfica del HMM planteado para el sistema.

- Se considera primero el hecho de que para cada palabra a ser identificada en el sistema se contará con un HMM.
- Los estados del HMM (Q) representan las letras de la palabra a ser identificada, además del estado de "silencio", el cual representa el silencio al inicio y al final de cada palabra pronunciada.
- Las observaciones del HMM (V) representan los fonemas que pueden ser pronunciados.
- Las probabilidades iniciales (π) consideran el hecho de que cada señal de audio a ser identificada comienzan en una señal de silencio, por lo que la probabilidad de que el silencio sea el primer estado siempre es del 100%.
- Las probabilidades de transición (A) representan la pronunciación normal de la palabra, considerando que cada observación representan 10 ms de la señal, se considera el tiempo general en el que una persona pronuncia una letra dada antes de pronunciar la siguiente, y así consecutivamente.
- Las probabilidades de observación (B) representan los diferentes fonemas que una persona puede pronunciar en cada letra dada. Estas probabilidades de observación consideran las variaciones en la pronunciación de las palabras debido a diferencias en el tono, capacidades diferentes o costumbre de los hablantes.

Una vez definido el modelo de cada palabra a ser identificada, la identificación de palabras se realiza de la siguiente manera:

- Se cuenta con una cadena de fonemas identificados de las estrategias de clasificación anteriores
- La cadena de fonemas será analizada en cada HMM que representa cada palabra, utilizando el procedimiento foward explicado anteriormente. Este procedimiento entregará como resultado la probabilidad de que cada modelo genere una cadena igual a la que se presenta.
- Se tienen entonces una probabilidad por cada palabra de representar la cadena de fonemas. Estas probabilidades se comparan unas con otras y se obtiene la mayor de ellas
- El modelo con mayor probabilidad representa entonces a la palabra que, a consideración del sistema, es aquella que representa mejor a la cadena de fonemas. Sin embargo, si la probabilidad no supera un valor de umbral dado, se considera que ninguna palabra es lo suficientemente probable como para representar la cadena (lo que daría como resultado un error).

4.4 Entrenamiento

El entrenamiento representa un elemento muy importante al proceso de identificación de palabras, ya que dota del conocimiento y la capacidad de decidir al sistema.

Para la metodología presentada, el entrenamiento consiste en un proceso de los cuales se obtendrá la información entrenada en dos momentos: en la generación de archivos de entrenamiento para la clasificación de fonemas, y en la generación de modelos para determinar la palabra detectada.

A continuación se explica el proceso, el cual se representa de manera general en la figura 4.11

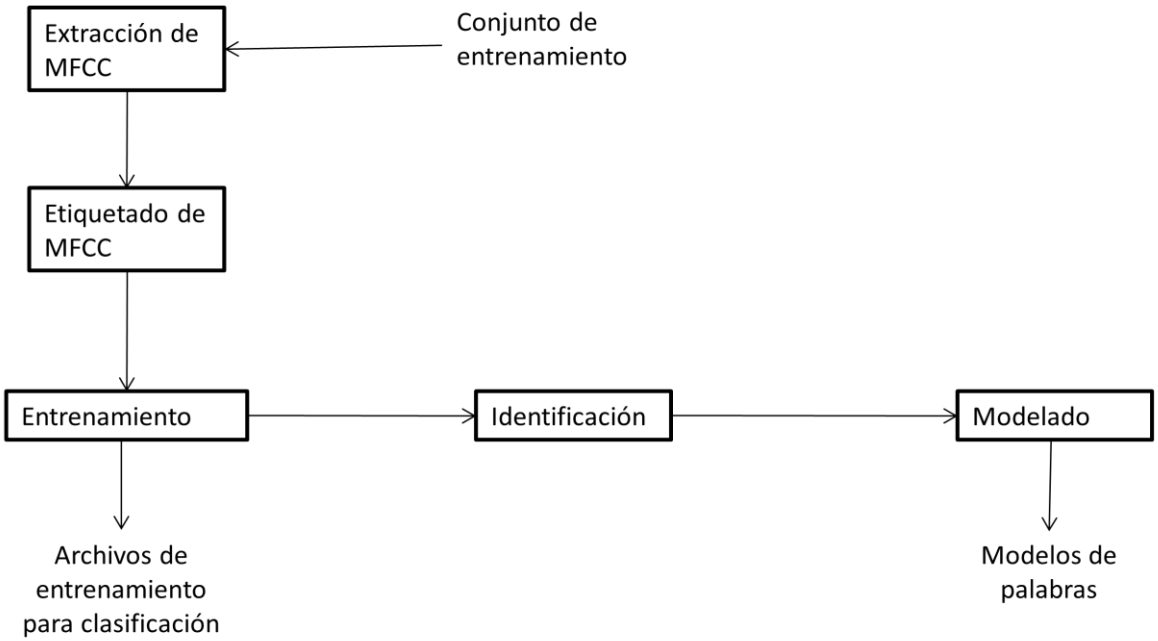


Fig. 4.11: Esquema general del proceso de entrenamiento.

- Conjunto de entrenamiento. Para realizar el entrenamiento es contar con un conjunto de señales previamente identificadas, que puedan representar la mayor variación posible de la información (esto con el fin de poder volver el sistema independiente del locutor).
- Extracción de MFCC. Se refiere al proceso en el cual cada señal en el conjunto de entrenamiento será segmentada en ventanas, sobre las cuales se calcularán sus coeficientes MFCC. De esta manera, para cada señal en el conjunto de entrenamiento, se tiene una matriz de Mx12, donde M representa el número de ventanas obtenidas (de cada ventana se obtendrán 12 coeficientes MFCC).
- Etiquetado de MFCC. Basado en el etiquetado a las palabras realizado anteriormente, se asigna un fonema a cada vector de MFCC. De esta manera se cuenta con una clasificación supervisada de todos los MFCC obtenidos del conjunto de entrenamiento
- Entrenamiento. Cada técnica de clasificación a ser utilizada realiza su procedimiento de entrenamiento de manera diferente. Durante el entrenamiento de clasificación se utiliza el proceso de entrenamiento correspondiente a la técnica de clasificación. En este caso, el proceso de entrenamiento de las estrategias utilizadas (distancias euclidianas, árboles de decisión y Naive Bayes) ya han sido explicadas anteriormente.

- De ese proceso se entrenamiento se obtiene el primer archivo, que contiene la información de entrenamiento para que el sistema posteriormente pueda clasificar señales a identificar.
- Identificación. Para obtener los parámetros de los modelos a ser utilizados, es necesario someter las señales del conjunto de entrenamiento al proceso de identificación, hasta el punto de obtener las cadenas de fonemas. Durante el proceso de identificación se separarán, entonces, las señales del conjunto de entrenamiento de acuerdo a la palabra que representan. De esta manera, se conoce previamente qué palabra representa cada cadena.
- Modelado. Como se mencionó anteriormente, cada palabra a ser identificada debe ser representada por un HMM. En este caso, para cada HMM se considera lo siguiente:
    - Los estados se obtienen al analizar la palabra que se representa.
    - Las observaciones se obtienen de todos los fonemas detectados.
    - Las probabilidades iniciales se obtienen del supuesto que se inicia del silencio, por lo tanto el estado del silencio representa un 100% de la probabilidad inicial, y el resto de los estados representan un 0%.
    - Las probabilidades de transición se obtienen del análisis de cada señal. Si se considera que cada ventana representa un periodo igual de tiempo, y que en cada palabra existe una transición de un estado al siguiente, entonces la probabilidad de transición de un estado al siguiente es de 1 entre el número de ventanas en las que se observó el estado, y la probabilidad de mantenerse en el mismo estado es de 1 menos la probabilidad de transición de un estado al siguiente.
    - Las probabilidades de observación se obtienen de un análisis de los fonemas detectados en la identificación con la letra que representan. Estos valores se estiman y se obtienen vectores de probabilidades de observación de cada estado para cada fonema.

Este modelado se realiza para cada palabra, obteniendo así el HMM de cada palabra a ser identificado.

Una vez realizado el entrenamiento del sistema, este mismo se encuentra listo para poder sometido a pruebas y análisis del mismo.

4.5 Desarrollo del sistema

El sistema utilizado para aplicar la metodología consiste en una serie de etapas, las cuales se encuentran numeradas de acuerdo al orden empleado para preparar el sistema de RAH. Las etapas se muestran a continuación:

Fig. 4.12: Sistema de Reconocimiento de palabras, junto con su entrenamiento, dividido en 10 etapas.

Dentro del sistema, las etapas 1 a 9 se refieren al proceso de entrenamiento, mientras que la etapa 10 consiste en el sistema de reconocimiento final.

1. Raws

"Raw" es la traducción al inglés de la palabra "crudo", y es generalmente utilizado en aquellos elementos los cuales son captados directamente del medio, antes de ser procesados. En el caso de este sistema, los RAWs se refieren a los archivos de audio captados directamente del micrófono y guardados, antes de pasar por el proceso de reducción de ruido.

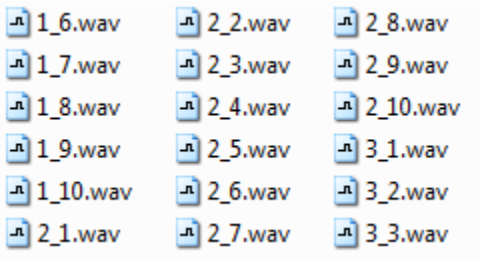


Fig 4.13: Archivos Raws.

Los archivos están ordenados bajo el siguiente formato:

X_Y.wav

donde:

- X se refiere al identificador de la persona que esta hablando.
- Y se refiere al número de muestra tomada de la persona X.

Es importante mencionar que cada archivo RAW contiene toda una secuencia de los números del 1 al 10 en tének.

2. Palabras listas

Las palabras listas hacen mención a los archivos de voz ya tratados y separados, los cuales siguen el siguiente formato de nombre:

X_Y_Z.wav

donde:

- X se refiere al identificador de la persona que está hablando.
- Y se refiere al número de repetición tomada de la persona X.
- Z se refiere a la palabra pronunciada.

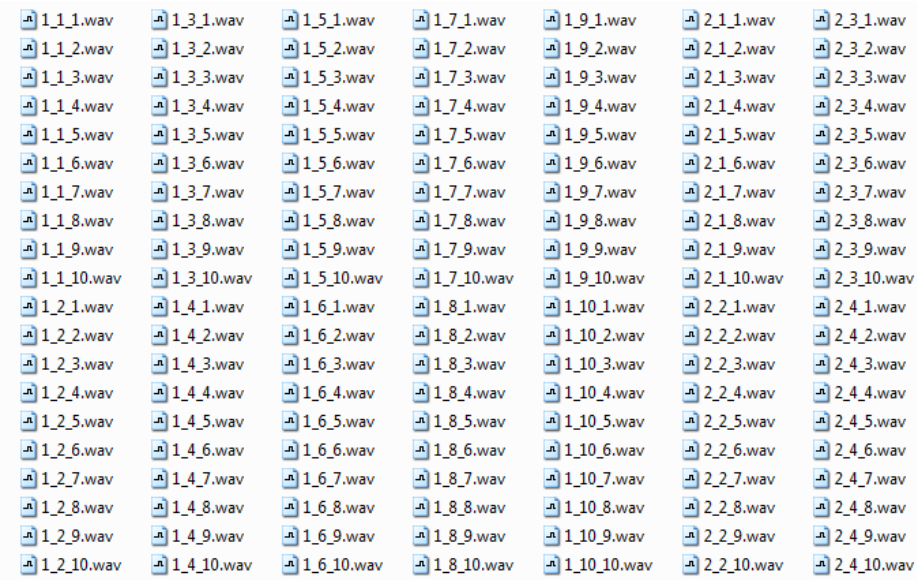


Fig. 4.14: Archivos tratados y separados, con su formato.

3. Etiquetados

Esta etapa contiene el etiquetado manual de los fonemas que se realiza a los archivos. Dentro de ellas se contiene un listado a dos valores que indican el tiempo en el que inicia un fonema dado, junto al fonema al que se refiere.

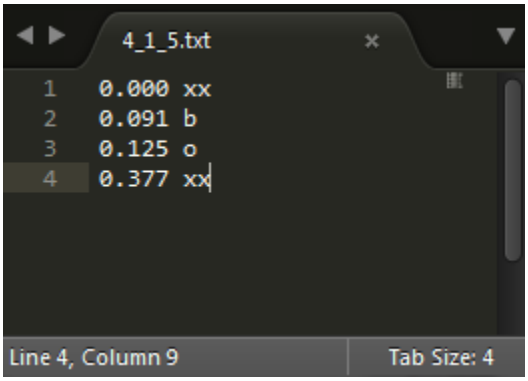
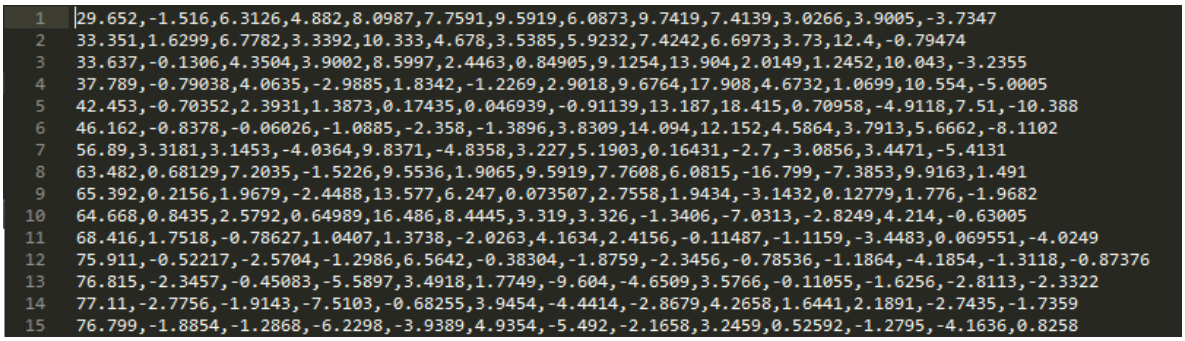


Fig. 4.15: Contenido de un archivo de etiquetas.

La nomenclatura de los archivos de etiquetado (así como la de las etapas 4 y 5) sigue el formato de los archivos de la sección de Palabras listas

4. MFCC de palabras

Esta sección consta del procedimiento de tomar los archivos de la sección 2, y utilizar el algoritmo de Wojcicki modificado para extraer los MFCC de cada archivo. Estos MFCC se guardarán en un nuevo archivo de extensión ".CSV"

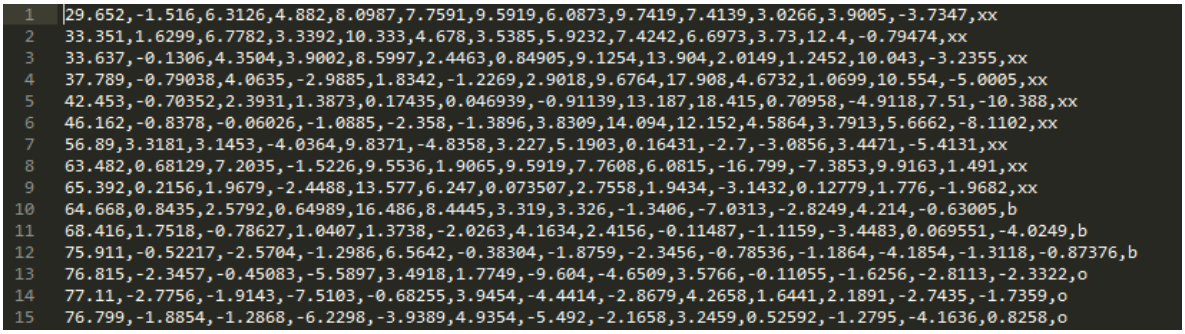


```
29.652,-1.516,6.3126,4.882,8.0987,7.7591,9.5919,6.0873,9.7419,7.4139,3.0266,3.9005,-3.7347
33.351,1.6299,6.7782,3.3392,10.333,4.678,3.5385,5.9232,7.4242,6.6973,3.73,12.4,-0.79474
33.637,-0.1306,4.3504,3.9002,8.5997,2.4463,0.84905,9.1254,13.904,2.0149,1.2452,10.043,-3.2355
37.789,-0.79038,4.0635,-2.9885,1.8342,-1.2269,2.9018,9.6764,17.908,4.6732,1.0699,10.554,-5.0005
42.453,-0.70352,2.3931,1.3873,0.17435,0.046939,-0.91139,13.187,18.415,0.70958,-4.9118,7.51,-10.388
46.162,-0.8378,-0.06026,-1.0885,-2.358,-1.3896,3.8309,14.094,12.152,4.5864,3.7913,5.6662,-8.1102
56.89,3.3181,3.1453,-4.0364,9.8371,-4.8358,3.227,5.1903,0.16431,-2.7,-3.0856,3.4471,-5.4131
63.482,0.68129,7.2035,-1.5226,9.5536,1.9065,9.5919,7.7608,6.0815,-16.799,-7.3853,9.9163,1.491
65.392,0.2156,1.9679,-2.4488,13.577,6.247,0.073507,2.7558,1.9434,-3.1432,0.12779,1.776,-1.9682
64.668,0.8435,2.5792,0.64989,16.486,8.4445,3.319,3.326,-1.3406,-7.0313,-2.8249,4.214,-0.63005
68.416,1.7518,-0.78627,1.0407,1.3738,-2.0263,4.1634,2.4156,-0.11487,-1.1159,-3.4483,0.069551,-4.0249
75.911,-0.52217,-2.5704,-1.2986,6.5642,-0.38304,-1.8759,-2.3456,-0.78536,-1.1864,-4.1854,-1.3118,-0.87376
76.815,-2.3457,-0.45083,-5.5897,3.4918,1.7749,-9.604,-4.6509,3.5766,-0.11055,-1.6256,-2.8113,-2.3322
77.11,-2.7756,-1.9143,-7.5103,-0.68255,3.9454,-4.4414,-2.8679,4.2658,1.6441,2.1891,-2.7435,-1.7359
76.799,-1.8854,-1.2868,-6.2298,-3.9389,4.9354,-5.492,-2.1658,3.2459,0.52592,-1.2795,-4.1636,0.8258
```

Fig. 4.16: Extracto de MFCC de un archivo.

5. MFCC con etiquetas

En esta sección se toman los archivos de la sección 4, y a cada conjunto de MFCC se le asigna un fonema basado en el etiquetado de la sección 3, considerando el hecho de que cada conjunto de MFCC representa 10ms de la señal de audio



```
29.652,-1.516,6.3126,4.882,8.0987,7.7591,9.5919,6.0873,9.7419,7.4139,3.0266,3.9005,-3.7347,xx
33.351,1.6299,6.7782,3.3392,10.333,4.678,3.5385,5.9232,7.4242,6.6973,3.73,12.4,-0.79474,xx
33.637,-0.1306,4.3504,3.9002,8.5997,2.4463,0.84905,9.1254,13.904,2.0149,1.2452,10.043,-3.2355,xx
37.789,-0.79038,4.0635,-2.9885,1.8342,-1.2269,2.9018,9.6764,17.908,4.6732,1.0699,10.554,-5.0005,xx
42.453,-0.70352,2.3931,1.3873,0.17435,0.046939,-0.91139,13.187,18.415,0.70958,-4.9118,7.51,-10.388,xx
46.162,-0.8378,-0.06026,-1.0885,-2.358,-1.3896,3.8309,14.094,12.152,4.5864,3.7913,5.6662,-8.1102,xx
56.89,3.3181,3.1453,-4.0364,9.8371,-4.8358,3.227,5.1903,0.16431,-2.7,-3.0856,3.4471,-5.4131,xx
63.482,0.68129,7.2035,-1.5226,9.5536,1.9065,9.5919,7.7608,6.0815,-16.799,-7.3853,9.9163,1.491,xx
65.392,0.2156,1.9679,-2.4488,13.577,6.247,0.073507,2.7558,1.9434,-3.1432,0.12779,1.776,-1.9682,xx
64.668,0.8435,2.5792,0.64989,16.486,8.4445,3.319,3.326,-1.3406,-7.0313,-2.8249,4.214,-0.63005,b
68.416,1.7518,-0.78627,1.0407,1.3738,-2.0263,4.1634,2.4156,-0.11487,-1.1159,-3.4483,0.069551,-4.0249,b
75.911,-0.52217,-2.5704,-1.2986,6.5642,-0.38304,-1.8759,-2.3456,-0.78536,-1.1864,-4.1854,-1.3118,-0.87376,b
76.815,-2.3457,-0.45083,-5.5897,3.4918,1.7749,-9.604,-4.6509,3.5766,-0.11055,-1.6256,-2.8113,-2.3322,o
77.11,-2.7756,-1.9143,-7.5103,-0.68255,3.9454,-4.4414,-2.8679,4.2658,1.6441,2.1891,-2.7435,-1.7359,o
76.799,-1.8854,-1.2868,-6.2298,-3.9389,4.9354,-5.492,-2.1658,3.2459,0.52592,-1.2795,-4.1636,0.8258,o
```

Fig. 4.17: MFCC con su fonema relacionado.

6. Archivos entrenamiento

Con la información de los archivos de la sección 5 se crean archivos mayores, conteniendo el conjunto completo de archivos de entrenamiento para las siguientes secciones.

7. Entrenamiento

Esta sección se encargará de tomar la información de los archivos de la sección 6, y dependiendo del método de clasificación se realizarán las acciones correspondientes para generar los archivos de entrenamiento que utilizará el sistema de reconocimiento. Las figuras 4.18, 4.19 y 4.20 muestran ejemplos de la información contenida en los archivos de entrenamiento para clasificación por distancias euclidianas, árbol de decisión y Naive Bayes, respectivamente

```
-2.8257,-2.9262,-3.5511,-0.38988,-2.4771,3.6323,2.9368,-3.512,0.33759,-1.3731,-1.2881,-0.59684,a
2.073,2.4317,0.3625,3.7158,3.0863,4.0755,-0.51909,-0.23131,3.2187,-1.6571,0.028287,-1.1016,b
-4.0761,6.6581,1.5703,-4.3397,-2.0984,2.3519,-4.7274,0.5732,0.88419,-3.2603,3.0046,-1.5135,e
```

Fig. 4.18: Ejemplo de entrenamiento para clasificación por distancias euclidianas.

```
1,<=,-2.7147,
2,<=,4.0052,
3,<=,-3.5632,
1,<=,-7.6008,
2,<=,-8.254,
4,<=,10.971, j
4,>,10.971, xx
2,>,-8.254,
1,<=,-11.476,
11,<=,3.9876,
```

Fig. 4.19: Fragmento de entrenamiento para un árbol de decisión.

```
-1.4739,15.6064,-3.7947,19.7274,-4.2759,33.5707,0.89101,24.7004,-2.2103,25.1971,-0.
048243,72.1097,2.1161,60.6948,-0.56697,60.8063,-0.35079,30.7977,-1.2944,17.4133,-0.32273,32.6668,
0.12003,25.3124,a,318
1.7102,18.3572,1.9771,20.577,-0.49087,30.3154,2.0215,28.2775,1.3616,30.1837,2.9808,44.3964,1.3518,46.8893,
0.553,80.829,1.9052,36.72,-2.551,20.9953,0.58091,18.3559,-0.57487,13.7245,b,121
-3.3962,11.9563,8.1214,23.5029,3.0381,26.5555,-3.5315,26.2413,-0.83514,23.4464,
0.53703,41.6363,-5.7856,28.6392,0.80444,31.3178,-1.2508,29.9964,-4.
0238,25.5601,2.683,25.2547,-1.2391,18.7115,e,216
```

Fig. 4.20: Fragmento de entrenamiento para Naive Bayes.

8. Identificación

Identificación se refiere al proceso de programación de los clasificadores que se encargarán de asignar un fonema dado un vector de MFCC no identificado. Cada técnica de clasificación tendrá entonces una subsección correspondiente en la etapa de identificación. Sin embargo, el requisito general de todos los clasificadores es el de requerir los mismos vectores de entrada y entregar los mismos vectores de salida, esto con el propósito de aplicar estos clasificadores en el sistema final. Para obtener mayor información sobre la manera en que se realizaron estos clasificadores, el apéndice B enuncia la programación de los métodos utilizados.

La salida general de todos los clasificadores es una cadena de fonemas identificados, tal como se muestra en al figura 4.x21

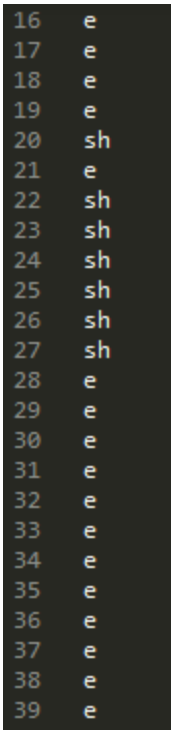


Fig. 4.21: Fragmento de una cadena de fonemas identificados.

9. Modelado

Esta etapa corresponde a la creación de los HMM de cada palabra a ser considerada en el sistema, basado en la identificación obtenida en la etapa 8 y el etiquetado realizado en la etapa 3. La figura 4.22 muestra el ejemplo del modelo de una palabra, con la que se explicará el mismo.

```
3   jun,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
    ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
4   12,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
    ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
5   0,0,0,0,0,0,0,0,0,0,0,1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
    ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
6   0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
    0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
    0,0,0,0,0,0,0,0,0,0.94737,0,0,0,0.052632,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
    ,0,0,0,0,0,0,0,0,0,0,0.040816,0,0,0.95918,0,0,0,0,0,0,0,0,0,0,0,0.02139,0.97861
7   0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
    0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
    0,0,0.19737,0.039474,0.039474,0.065789,0,0,0.013158,0.23684,0,0,0.078947,0.32895,0
    ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.13265,0.05102,0.020408,0.010204,0
    ,0,0,0.020408,0.11224,0,0.5,0.15306,0.06383,0,0.090426,0,0,0.021277,0,0.021277,0,0
    ,0.058511,0.74468
```

Fig. 4.22: HMM representando a la palabra "jun" (uno).

La primera línea (3) indica la palabra, la cual funcionará como identificador para la detección

La segunda línea (4) indica el número de fonemas involucrados, contando el silencio. Para explicar con mejor detalle las siguientes líneas, se asignará a este número la variable **n**.

La tercera línea (5) representa **π**, el vector de probabilidad de estados iniciales. Se consideran entonces **n** valores, donde el último de ellos representa la probabilidad de iniciar con el silencio. Este **n**-ésimo valor siempre tendrá una probabilidad de 1 (dado que se considera que todas las palabras aisladas inician en un estado de silencio).

La cuarta línea (6) representa ***A***, la matriz de probabilidades de transiciones entre estados. En esta línea se consideran $\mathbf{n}^2$ valores

La quinta línea (7) representa ***B***, la matriz de probabilidades de observaciones de cada estado. En esta línea se consideran $\mathbf{n}^2$ valores.

10. Sistema final

Esta etapa incluye el sistema de identificación, el cual utilizará los archivos de entrenamiento generados en la etapa 7, las técnicas de clasificación programadas en la etapa 8 y los modelos generados en la etapa 9

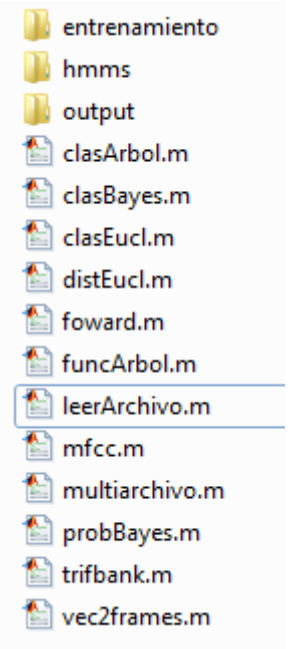


Fig. 4.23: Contenido del sistema de identificación, etapa 10.

Los archivos mostrados en la figura 4.23 se explican a continuación:

- leerArchivo.m corresponde al programa principal, que realizará la labor de identificación de palabras.
- multiarchivo.m es una instancia que permite ejecutar leerArchivo.m múltiples veces.
- los archivos clasArbol.m, clasBayes.m y clasEucl.m corresponden a las técnicas de clasificación programadas en la etapa 8.
- los archivos distEucl.m, funcArbol.m y probBayes.m son funciones auxiliares para las técnicas de clasificación.
- el archivo mfcc.m corresponde al algoritmo de Wojcicki modificado
- los archivos trifbank.m y vec2frames.m son archivos auxiliares de mfcc.m
- foward.m contiene la función que permite aplicar el procedimiento foward de los HMM.

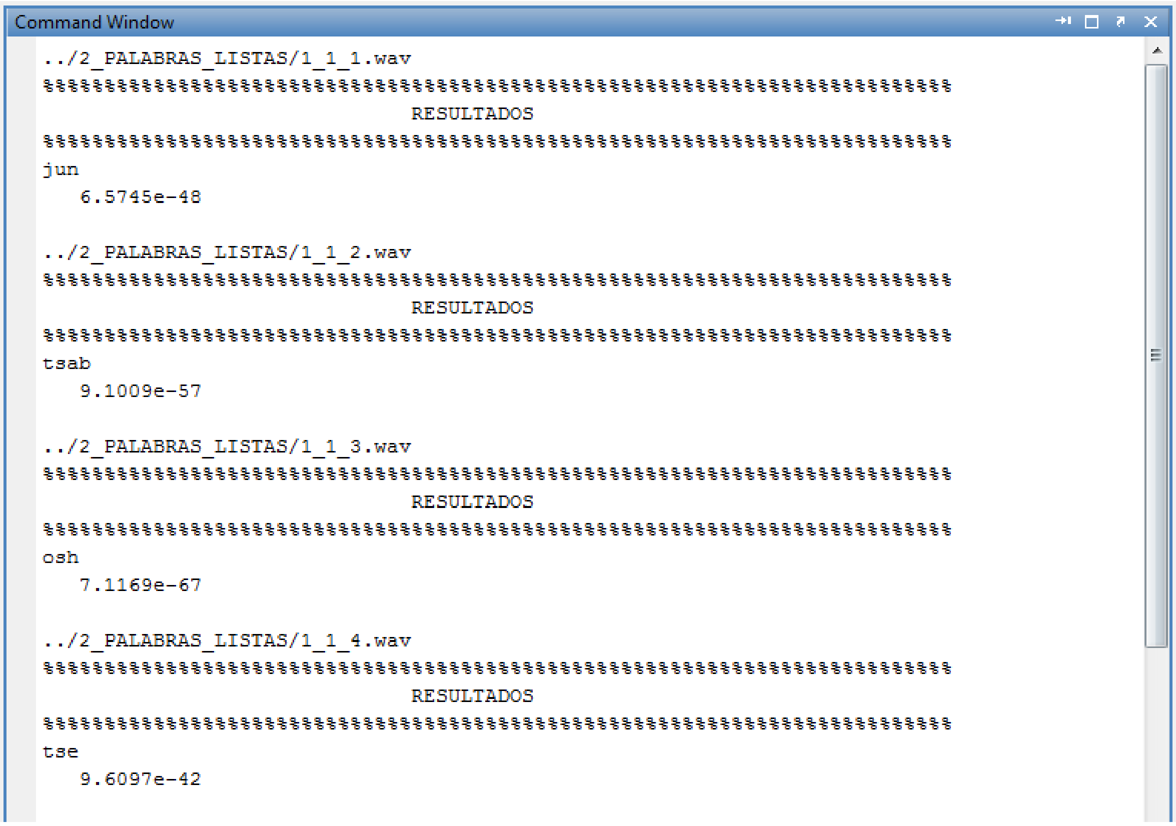


Fig. 4.24: Resultados para múltiples aplicaciones del proceso de identificación.

En el apéndice A se encuentra la programación general del programa principal de reconocimiento de palabras.

5
Evaluación

La evaluación es el proceso que determina la efectividad del modelo planteado, así como también dota de elementos de análisis que permiten tomar un curso de acción basado en los resultados obtenidos.

5.1 Objetivos

La evaluación del sistema tiene como objetivo fundamental determinar la efectividad de la metodología aplicada en la solución del problema. También, como objetivos adicionales se contemplan obtener observaciones importantes relativas al desempeño del sistema, en sus diferentes variaciones, de acuerdo a las pruebas de desempeño, con la intención de determinar una estrategia a seguir en el desarrollo de un sistema cada vez más completo de reconocimiento del habla.

5.2 Criterios de evaluación

Para evaluar la metodología utilizada, se utilizará el conjunto de pruebas, el cual consiste en un grupo de archivos de palabras tomadas del corpus realizado anteriormente. Es importante mencionar que los archivos utilizados para el conjunto de pruebas deben no ser utilizados en el conjunto de entrenamiento.

Este conjunto de pruebas serán sometidas al sistema de reconocimiento de palabras. Posteriormente, se identificarán manualmente las palabras del conjunto de pruebas. Si el sistema de reconocimiento detectó correctamente la palabra, se considerará un acierto. El desempeño del sistema consistirá entonces en una relación porcentual entre el número de palabras detectadas correctamente del conjunto de palabras analizadas.

5.3 Modelos a evaluar

Los modelos a ser evaluados siguen el esquema presentado en la metodología, donde se distinguen:

- Característica observables.
- Técnicas de clasificación.
- Modelado.

Cada variación en los elementos característicos del modelado supone un modelo diferente, que puede ser evaluado. De esta manera, una combinación de distintas características observables, técnicas de clasificación y modelados suponen un conjunto de estrategias a ser evaluadas.

5.4 Definición de las pruebas

Las pruebas se realizaron por medio de combinaciones de los siguientes criterios:

Característica observable:

- MFCC (de acuerdo al algoritmo de wojcicki).

Criterio de clasificación:

- Distancias Euclidianas.
- Árbol de Decisión.
- Naive Bayes con probabilidades equivalentes. Se asume que cada fonema representado en el método de Naive Bayes tiene la misma probabilidad de ser observado. Por ejemplo, si se representan 10 fonemas, cada fonema tiene la probabilidad de 1/10 de ser observado.
- Naive Bayes con probabilidades basadas en entrenamiento. La probabilidad de observación de cada fonema es obtenido mediante la

relación de porcentaje de observación del mismo en los conjuntos de entrenamiento. Es decir, si un fonema es observado un 50% de los casos, tiene una probabilidad de 50%.

Modelado:

- HMM sin error en su entrenamiento (limpio). Es decir, en este modelado se aseguró que en el conjunto de entrenamiento no existieran errores.
- HMM con posible error en su entrenamiento (sucio). En este modelado, se incluye en el entrenamiento aquellos casos donde hay un posible error, o pueda ser considerado de manera diferente.

Locutores:

- Locutor entrenado, es decir, que en el conjunto de entrenamiento se encuentre información obtenida del mismo locutor con el que se esta realizando la prueba.
- Locutor no entrenado, es decir, que en el conjunto de entrenamiento no se encuentre información obtenida del mismo locutor con el que se esta realizando la prueba.

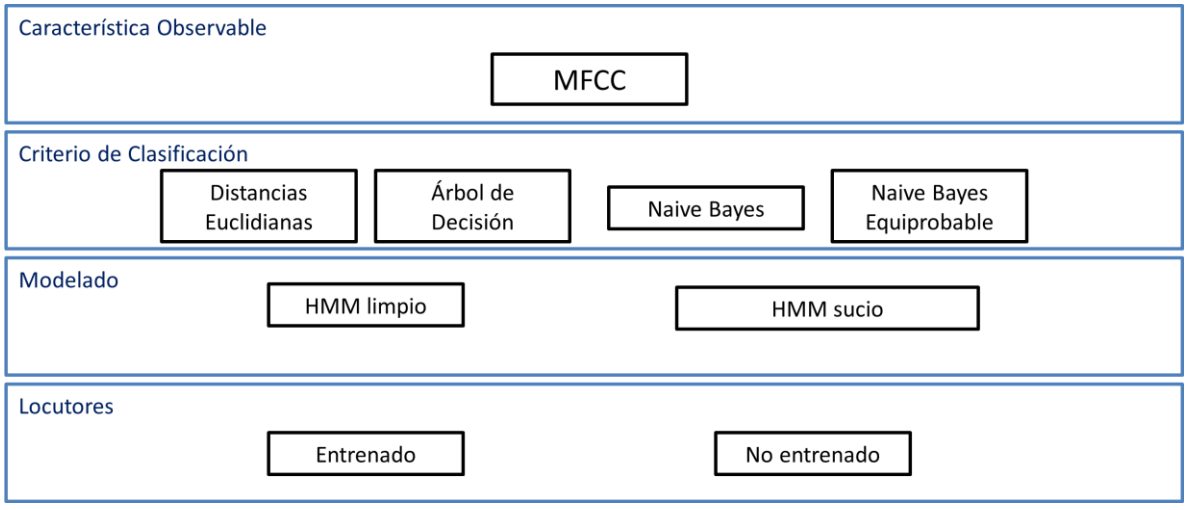


Fig. 5.1: Criterios utilizados para generar estrategias de sistemas de RAH.

5.5 Resultados

Las tablas 5.2, 5.3, 5.4 y 5.5 muestran los resultados de las pruebas realizadas. Cada tabla representa una estrategia de reconocimiento utilizada, con columnas de reconocimiento de acuerdo al tipo de locutor y a la estrategia de entrenamiento de los HMM.

MFCC – Distancias Euclidianas – HMM

Palabra	Loc. entr. HMM limpio.	Loc. no ent. HMM limpio.	Loc. entr. HMM sucio.	Loc. no entr. HMM sucio.
Jūn	37.5%	10%	42.5%	10%
Tsāb	55%	60%	52.5%	60%
Ōx	87.5%	40%	85%	30%
Tsē	52.5%	70%	72.5%	100%
Bō	52.5%	40%	60%	40%
Acac	60%	0%	70%	0%
Būc	82.5%	30%	80%	30%
Huaxic	67.5%	80%	90%	90%
Belēu	85%	0%	87.5%	30%
Lājuj	70%	80%	72.5%	80%
Total	**65%**	**41%**	**71.25%**	**47%**

Tabla 5.2: Resultados de la combinación de MFCC- Distancias Euclidianas - HMM

MFCC – Árbol de decisión – HMM

Palabra	Loc. entr. HMM limpio.	Loc. no ent. HMM limpio.	Loc. entr. HMM sucio.	Loc. no entr. HMM sucio.
Jūn	50%	10%	40%	0%
Tsāb	12.5%	0%	50%	10%
Ōx	42.5%	10%	72.5%	10%
Tsē	52.5%	40%	90%	60%
Bō	47.5%	50%	52.5%	30%
Acac	15%	10%	77.5%	30%
Būc	27.5%	20%	40%	10%
Huaxic	22.5%	0%	72.5%	20%
Belēu	15%	0%	75%	40%
Lājuj	7.5%	0%	40%	20%
Total	**29.25%**	**14%**	**61%**	**23%**

Tabla 5.3: Resultados de la combinación de MFCC - Árbol de decisión - HMM.

MFCC – Naive Bayes (equiprobable)– HMM

Palabra	Loc. entr. HMM limpio.	Loc. no ent. HMM limpio.	Loc. entr. HMM sucio.	Loc. no entr. HMM sucio.
Jūn	45%	40%	45%	40%
Tsāb	62.5%	80%	67.5%	80%
Ōx	55%	40%	80%	40%
Tsē	75%	70%	92.5%	100%
Bō	60%	60%	60%	60%
Acac	82.5%	100%	77.5%	100%
Būc	70%	80%	62.5%	90%
Huaxic	87.5%	80%	97.5%	100%
Belēu	75%	60%	75%	100%
Lājuj	47.5%	20%	67.5%	20%
Total	**66%**	**63%**	**72.5%**	**73%**

Tabla 5.4: Resultados de la combinación de MFCC- Naive Bayes (equiprobable) - HMM

MFCC – Naive Bayes (no equiprobable) – HMM

Palabra	Loc. entr. HMM limpio.	Loc. no ent. HMM limpio.	Loc. entr. HMM sucio.	Loc. no entr. HMM sucio.
Jūn	75%	80%	72.5%	80%
Tsāb	85%	40%	80%	30%
Ōx	72.5%	80%	85%	60%
Tsē	85%	100%	90%	100%
Bō	67.5%	100%	67.5%	80%
Acac	92.5%	100%	92.5%	100%
Būc	62.5%	10%	62.5%	0%
Huaxic	80%	100%	87.5%	100%
Belēu	92.5%	100%	92.5%	100%
Lājuj	70%	70%	70%	70%
Total	**78.25%**	**78%**	**80%**	**72%**

Tabla 5.5: Resultados de la combinación de MFCC- Naive Bayes (no equiprobable) - HMM

5.6 Análisis

De acuerdo a los resultados obtenidos, y a las observaciones realizadas durante la evaluación, se obtuvieron las siguientes conclusiones:

- De las técnicas de clasificación, Naive Bayes presenta mejores resultados, esto por la flexibilidad que presenta ante la variación de los atributos del elemento a ser clasificado (a diferencia de las otras dos técnicas).
- Entrenar el sistema con información de los posibles errores permite al éste considerar los mismos en el proceso de detección de palabras, lo que le permite ser más robusto ante el ruido presente en una señal dada.
- Un problema fundamental que no había sido considerado, y que se muestra en los resultados es la detección de palabras con una composición fonética similar (por ejemplo, entre bo y buc).
- La técnica utilizada para obtener la cadena de fonemas impacta en gran manera en el desempeño del sistema de RAH, por lo que es necesario probar con otras técnicas de clasificación para considerar una mejora en el porcentaje de detección.

6
Conclusiones

6.1 Objetivos Cumplidos

Al inicio de este proyecto se presentaron ciertos objetivos a cumplirse durante el desarrollo del mismo. Es importante mencionar, entonces, que los objetivos presentados fueron cumplidos satisfactoriamente. A continuación se muestran los resultados de acuerdo al cumplimiento de objetivos.

Objetivo general

El objetivo general fue cumplido, ya que el sistema desarrollado permite identificar un conjunto de palabras, con un porcentaje de detección aceptable.

Objetivos específicos

- Se definió un conjunto de estrategias para desarrollar el sistema.
- Se obtuvo un conjunto de muestras de la lengua Tének.
- Se utilizó parte de este conjunto de muestras para entrenar el sistema.
- Se realizaron pruebas para probar la efectividad del sistema.
- Para realizar el sistema, se tomaron las recomendaciones de otros autores que han desarrollado sistemas similares con resultados positivos.

6.2 Conclusiones

Realizar un sistema de RAH es un proceso que, de primera instancia, se sabe que no es sencillo. Sin embargo, es un proceso que, una vez finalizado, podrá resolver muchos problemas relacionados con un sector de la población que tristemente no es considerado generalmente.

Durante el desarrollo del sistema de RAH se encontraron muchas consideraciones que no se habían tomado en cuenta inicialmente. Algunas de ellas son del carácter técnico del sistema mismo, mientras que otras corresponden al carácter social y cultural que se vive en el país.

El primer problema sociocultural importante es la cantidad limitada de hablantes de una lengua autóctona. La mayoría de estas personas viven en regiones remotas, de difícil acceso, donde en el mejor de los casos se encuentran comunicadas solamente por caminos de terracería, y una cantidad muy pequeña de hablantes existen en regiones urbanas.

El segundo problema es la disposición. Si bien encontrar hablantes es complicado, el brindar apoyo para el desarrollo de un sistema como el presente es una tarea aún más difícil. En el caso de hablantes en zonas urbanas, en muchas ocasiones la vergüenza de hablar la lengua es muy grande como para admitir en casos que la hablan, mientras que en otros casos es necesario un incentivo (generalmente de carácter económico) para lograr cierta cooperación.

Sin embargo, con los resultados obtenidos, se demuestra que la identificación del habla en lenguas autóctonas es un proyecto posible, con la capacidad de mejorar de manera significativa su desempeño, dados los recursos apropiados.

6.3 Aportaciones

El desarrollo de un sistema funcional de reconocimiento de palabras, en una lengua que no ha sido considerada con anterioridad, presentando resultados de hasta un 80% de exactitud representa la mayor aportación del desarrollo de este proyecto.

El sistema, por sí mismo, puede ser adaptado para considerar mas palabras y ser utilizado como un detector de palabras, para apoyar en casos tales como la detección de síntomas y las indicaciones básicas, solucionando problemas de comunicación

Esta tesis representa uno de los primeros pasos dirigidos a la aplicación de tecnologías de procesamiento del habla en poblaciones indígenas del país, por lo que también aporta una base para el desarrollo de trabajos futuros.

De igual manera, el desarrollo de un corpus pensado para el proceso de RAH es también una aportación importante, ya que en este momento no existe un corpus diseñado para ser utilizado en reconocimiento del habla.

Por último, de este proyecto surgió el desarrollo de tres artículos técnicos sobre el reconocimiento del habla y la metodología utilizada en esta tesis.

6.4 Trabajo futuro

Debido a que este es un trabajo inicial de RAH, existen muchas áreas de oportunidad donde se puede dedicar una investigación para mejorar este sistema. A continuación se enuncian algunos puntos considerados como trabajo futuro (los cuales no son necesariamente todos, pero son los que se han observado) :

Corpus

Es importante el desarrollo de un corpus más completo para poder ser utilizado en sistemas de RAH, y debido a las complicaciones encontradas, es un trabajo que involucra dedicación. Una alternativa consiste en la adaptación del corpus Entendámonos, de [4], el cual contiene muestras en lengua tének, pero sin el etiquetado de palabras correspondiente.

Criterios de clasificación de fonemas

Mejorar la clasificación de fonemas otorgará mejores resultados en la detección de palabras. Se probaron tres estrategias de clasificación, pero actualmente se están desarrollando proyectos de clasificación (independientes del RAH) con técnicas como las redes neuronales, las cuales están dando resultados satisfactorios. Entonces, es importante considerar esas técnicas, implementarlas en el sistema actual y analizar los resultados obtenidos con ellas.

Reconocimiento del habla con ruido

Es importante analizar una señal de habla con ruido, ya que para el desarrollo de herramientas tales como un traductor, se necesitará tomar la señal de voz sin ser tratada antes, por lo que el ruido será un factor presente e importante en la detección. Para este problema es importante analizar las estrategias de detección y reducción de ruido, o la implementación de modelos con señales de ruido.

Reconocimiento del habla continuo

Una vez obtenidos resultados satisfactorios con un sistema de habla de palabras aisladas, es importante observar su comportamiento en el habla continuo. Esta es la meta del sistema de reconocimiento del habla, y la que marcará su aplicación para el desarrollo de herramientas que solucionen los problemas mencionados en un inicio.

Referencias

Libros

[1] Ian H Witten et al. "Data Mining Practical Machine Learning Tools and Techniques". Third Edition. Elsevier Inc. 2011

[2] Joseph Mariani. "Spoken Language Processing". ISTE Ltd. 2008

[3] Ramón Larsen et al. "Vocabulario huasteco del Estado de San Luis Potosí". Instituto Lingüistico de Verano, México D.F., México, 1997

Tesis

[4] Juan Carlos Flores Paulín, "Técnicas para el reconocimiento de voz en palabras aisladas en la lengua náhuatl", Instituto Politécnico Nacional, 2009.

[5] Carlos Arturo Hernández Zepeda, "Corpus de Lenguas Indígenas Mexicanas para la Identificación Automática del Lenguaje Hablado", Instituto Tecnológico de Ciudad Madero, 2013

[6] Marco A. Camejo, "Reconocimiento Automático del Habla mediante un modelo híbrido basado en Modelos Ocultos de Markov y Redes Neuronales Artificiales. Caso de estudio: Habla venezolana", Universidad de los Andes Mérida, Venezuela, Junio 2008.

[7] Mikael Nilsson et al, "Speech Recognition using Hidden Markov Model, perfomance evaluation in noisy enviroment". Blekinge Institute of Technology, Suecia, 2002.

[8] Juan Luis Navarro Mesa, "Procesador Acústico: El Bloque de Extracción de Características". Universidad de Las Palmas de Gran Canaria.

[9] Tomás Navarrete Gutiérrez, "Detección de anomalías en la carga de un procesador, utilizando modelos ocultos de Markov", Instituto Tecnológico de Morelia, 2007

[10] Gildardo Contreras Morales, "Digitalizador de voz", Universidad de Guadalajara.

[11] "PFC-HMM y algoritmo EM", Biblioteca de ingeniería de la Universidad de Sevilla, España.

Artículos

[12] Santiago-Omar Caballero, "On the Development of Speech Resources for the Mixtec Language" The scientific world journal, Volumen 2013, Febrero 2013.

[13] Diego H. Milone, "Modelos ocultos de Markov para el reconocimiento automático del habla", 2004.

[14] Lawrence R. Rabiner "A tutorial on hidden markov models and selected applications in speech recognition.", 1990.

[15] Richard A. O'Keefe, "An introduction to Hidden Markov Models", 2004.

[16] Phil Blunsom, "Hidden Markov Models", 2004

[17] Chadawan Ittichaichareon, Siwat Suksri y Thaweesak Yingthawornsuk, "Speech Recognition using MFCC". International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012), pp 135-138, Julio 2012.

[18] D. R. Tomassi et al., "Evaluación de técnicas clásicas de reducción de ruido en señales de voz". Revista Argentina de Bioingeniería, Vol XX, No X, Diciembre 2005.

[19] Fabian M. Hernandez, "El concepto de distancia y su aplicación en estadística multivariada", Revista AMAI, México

[20] José María Drake Moyano. "Ruidos e interferencias: Técnicas de reducción". Universidad de Cantabria. España. 2005.

[21] Malagón Luque, Constantino. "Clasificadores bayesianos. El algoritmo Naïve Bayes". Universidad Nebrija. Madrid, España. 2003

Publicaciones virtuales

[22] "La población indígena en México", INEGI, 2004

[23] "10 datos de los pueblos indígenas", Revista Quo México, 2013

[24] "Pueblos indígenas", Organización de las Naciones Unidas. http://www.un.org/es/globalissues/indigenous/

[25] "La situación de los pueblos indígenas en el mundo", Organización de las Naciones Unidas, http://www.cinu.org.mx/pueblosindigenas/docs/Informe_Completo_Ingles.pdf

[26] "Catálogo de las Lenguas Indígenas Nacionales", INALI, 2010, http://www.inali.gob.mx/clin-inali/

[27] "Diccionario Español - Huasteco (Tének)", Educein Culturas. 2011. http://educeinculturas.blogspot.mx/

[28] "Atención a indígenas en materia penal y penitenciaria", Comisión Nacional para el Desarrollo de los Pueblos Indígenas, 2014, http://www.cdi.gob.mx/index.php?option=com_content&view=article&id=3082

[29] "Lengua Tének, entre las que están en riesgo de desaparecer", Revista/Diario digital emsavalles. Febrero 2013. http://www.emsavalles.com/revtxt.php?r=2923

[30] "No hablar español, el 'delito' por el que indígenas han pisado la cárcel", CNN en español, Octubre, 2013. http://mexico.cnn.com/nacional/2013/10/12/no-hablar-espanol-el-delito-por-el-que-indigenas-han-pisado-la-carcel

[31] Artículo: "A la baja, cantidad de hablantes de lenguas indígenas en México", Informador, Jalisco, 2013. http://www.informador.com.mx/jalisco/2013/473192/6/a-la-baja-cantidad-de-hablantes-de-lenguas-indigenas-en-mexico.htm

[32] "About Siri", Apple online support, 2014, http://support.apple.com/en-us/ht4992

[33] "Apple - iOS 8 - Siri", Apple , 2014, https://www.apple.com/mx/ios/siri/ -- imagen de Siri

[34] "Ok Google y la búsqueda por voz.", Google, 2014 https://support.google.com/websearch/answer/2940021?hl=es

[35] "Utilizar reconocimiento de voz en Windows XP", Windows Corporation, http://support.microsoft.com/kb/306901/es

[36] Bruno López Takeyas, "Algoritmo ID3", Instituto tecnológico de Nuevo Laredo, México

[37] "Naive Bayes", SciKit Learn. http://scikit-learn.org/stable/modules/naive_bayes.html

[38] Ivan Vladimir Meza Ruiz, "MFCCs", Universidad Nacional Autónoma de México, 2013, http://turing.iimas.unam.mx/~ivanvladimir/es/post/MFCC/ .

[39] "Modelos Ocultos de Markov", wikipedia, http://es.wikipedia.org/wiki/Modelo_oculto_de_M%C3%A1rkov

[40] Wojcicki, Kamil. “Mel frequency Cepstral Coefficient feature extraction”. Matlab Central. 2011

[41] Arturo Camacho "Reconocimiento del habla", ArturoCamachoClases, https://www.youtube.com/playlist?list=PLH2VuIvVYXkZmQIJ1Vy2p-KaBCwRmx4-H

Herramientas y software

[42] Matlab webpage. http://www.mathworks.com/products/matlab/

[43] Audacity webpage. http://audacity.sourceforge.net/

[44] Sublime text webpage. http://www.sublimetext.com/

[45] Weka webpage. http://www.cs.waikato.ac.nz/ml/weka/

A

Programa de Identificación de Palabras

A continuación se muestra el algoritmo utilizado para realizar el proceso de identificación de palabras mencionado en la metodología:

```
function
salida=leerArchivo(archivo,entrenamiento,entrenamiento2,modelo,tipoClasif
icacion,nCoef)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%% DEFINICION DE VARIABLES                                          %%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Tw = 20;                % analysis frame duration (ms)
Ts = 10;                % analysis frame shift (ms)
alpha = 0.95;           % preemphasis coefficient
M = 20;                 % number of filterbank channels
C = nCoef;                 % number of cepstral coefficients
L = 22;                 % cepstral sine lifter parameter
LF = 250;               % lower frequency limit (Hz)
HF = 3400;              % upper frequency limit (Hz)
wav_file = archivo;     % audio file

lectura_file=entrenamiento;  %El archivo de lectura del conj de
entrenamiento
fonemas_file=entrenamiento2; %El archivo de lectura del arbol de desicion
modelos_file=modelo;

%%%%% PASO 1: LEER ARCHIVO DE AUDIO
 [ speech, fs, nbits ] = wavread( wav_file );
%%%%% PASO 2: CALCULAR LOS MFCC DEL ARCHIVO DE AUDIO
[ mfccs2, FBEs, frames ] = mfcc( speech, fs, Tw, Ts, alpha, @hamming, [LF
HF], M, C+1, L );

%%%%% PASO 3: PREPARAR MATRIZ PARA CLASIFICACION
mfccs=transpose(mfccs2);
mfccs(:,1)=[]; %Eliminar el primer elemento de energias de los mfcc
```

```
%%%%% PASO 4: CLASIFICACION DE CADENA DE FONEMAS
if(tipoClasificacion==0)
    [clasificacion clasNum nFonemas]=clasEucl(mfccs,lectura_file,nCoef);
elseif(tipoClasificacion==1)
    [clasificacion clasNum
nFonemas]=clasArbol(mfccs,lectura_file,fonemas_file,nCoef);
else
    [clasificacion clasNum nFonemas]=clasBayes(mfccs,lectura_file,nCoef);
end

%%%%% PASO 5: LECTURA DE LOS MODELOS OCULTOS DE MARKOV
fid=fopen(modelos_file);
A=textscan(fid,'%s');
fclose(fid);

[model_rows model_columns] = size (A{1,1});

%Pasar toda la info a una tabla
for i=1:model_rows
   CC=regexp(A{1,1}{i,1},',','split');
   [hlpRows hlpCols]=size(CC);
   for j=1:(hlpCols)
       B{i,j}=CC{j};
   end
end

%Poner toda la info en tablas diferentes para facil acceso
nModelos=str2num(B{1,1});
nombreModelo=cell(1,nModelos);
pi=cell(1,nModelos);
trans=cell(1,nModelos);
obs=cell(1,nModelos);

for i=1:nModelos
    numBase=((i-1)*5)+3;
    nEstados=str2num(B{numBase+1,1});
    pi{1,i}=zeros(1,nEstados);
    trans{1,i}=zeros(nEstados);
    obs{1,i}=zeros(nEstados,nFonemas);

    nombreModelo{1,i}=B{numBase,1};%leer el nombre del modelo

    for j=1:nEstados %leer pi
       pi{1,i}(1,j)=str2num(B{numBase+2,j});
    end

    for j=1:nEstados %leer trans
       for k=1:nEstados
           trans{1,i}(j,k)=str2double( B{numBase+3,(j-1)*nEstados+k}  );
       end
    end

    for j=1:nEstados %leer obs
        for k=1:nEstados
            obs{1,i}(j,k)=str2double( B{numBase+4,(j-1)*nEstados+k} );
```

```
        end
    end
end

%%%%% PASO 6:  CALCULO DE PROBABILIDAD DE LA CADENA PARA CADA MODELO
probs=zeros(1,nModelos);

for i=1:nModelos
    prob(1,i)=foward(clasNum,pi{1,i},trans{1,i},obs{1,i});
end
%%%%% PASO 7: RESULTADOS
nombreMin=nombreModelo{1,1};
min=prob(1,1);

for i=1:nModelos
   if(prob(1,i)>min)
        nombreMin=nombreModelo{1,i};
        min=prob(1,i);
    end
end

disp('RESULTADOS')

disp(nombreMin)
disp(min)
salida=nombreMin;
end
```

B

Clasificadores

En este apéndice se mostrarán los algoritmos utilizados para realizar el proceso de clasificación de las tres técnicas utilizadas: Distancias euclidianas, árbol de decisión y Naive Bayes.

Algoritmo de distancias Euclidianas:

```
function [clasificacion clasNum
nFonemas]=clasDistEucl(mfccs,training_file,nCoef)

% Uso: Entregar dos vectores de clasificacion y el numero de fonemas
entregados dada una tabla de mfccs, un archivo de lectura y el # de
coeficientes %
%      Uno de los vectores entrega las letras, otro entrega los numeros
relacionados a cada letra
%
%%%%% ENTRADAS %%%%%
% mfccs - Una matriz n * m conteniendo un conjunto de mfcss ('n'
conjuntos de 'm' coeficientes mfcc) %
% training_file - el archivo de entrenamiento %
% nCoef - el num de coeficientes mfccs (debe de ser coherente con la
matriz mfccs) %
%
%%%%% SALIDAS %%%%%
% clasNum - el vector de las letras identificadas, en numerico (A es 1, B
es 2 y asi... de acuerdo al orden en el archivo de entrenamiento) %?????
% nFonemas - el numero de mfcss clasificados (coherente con 'n' de mfccs
en la entrada)????? %

[mfccs_rows mfccs_cols]=size(mfccs);
%%%%% IMPORTAR EL ARCHIVO DE LOS PROMEDIOS DEL ENTRENAMIENTO %%%%%
[file msg]=fopen(training_file);
A=textscan(file,'%s');
[train_rows train_columns] = size (A{1,1});

for i=1:train_rows
   C=regexp(A{1,1}{i,1},',','split');
   for j=1:(nCoef+1)
```

```
        B{i,j}=C{j};
    end
end

%EN ESTE PUNTO TENGO:
%A - Matriz auxiliar que tiene los archivos de entrada
%B - Matriz donde estan todos los coeficientes de entrenamiento <-- ESTE
ES EL QUE SE USA
%C - Matriz auxiliar para sacar los coeficientes

%%%%% CLASIFICACION DE FONEMAS POR DISTANCIAS EUCLIDIANAS
%%%%%
%creamos dos matrices donde guardar los resultados, una de letras y otra
de numeros
    clasificacion=cell(mfccs_rows,1);
    clasNum=zeros(mfccs_rows,1);

    for i=1:mfccs_rows
        %empezamos suponiendo que el primero es el mas cercano
        selected=1;
        %%%%% SE UTILIZA UNA FUNCION EXTERNA LLAMADA distEucl.m %%%%%
        minDist=distEucl(mfccs(i,:),B(1,:),nCoef);
        %Revisamos el resto de las opciones de clasificacion, si hay una
dist. euclidiana menor
        %actualizamos para que ese sea el valor menor elegido
        for j=2:train_rows
           testDist= distEucl(mfccs(i,:),B(j,:),nCoef);
           if(testDist<minDist)
               selected=j;
               minDist=testDist;
           end
        end
        %Ya con el valor menor, asignamos al conjunto de mfcc
        clasificacion{i}=B{selected,nCoef+1};
        clasNum(i)=selected;
    end
    nFonemas=mfccs_rows;
    disp('fin');
end
```

Algoritmo de Árbol de decisión:

```
function [clasificacion clasNum
nFonemas]=arbolDesicion(mfccs,training_file,coef_file,nCoef)
% Uso: Entregar dos vectores de clasificacion y el numero de fonemas
entregados dada una tabla de mfccs, un archivo de lectura y el # de
coeficientes %
%      Uno de los vectores entrega las letras, otro entrega los numeros
relacionados a cada letra
%
%%%%% ENTRADAS %%%%%
```

```
% mfccs - Una matriz n * m conteniendo un conjunto de mfcss ('n'
conjuntos de 'm' coeficientes mfcc) %
% training_file - el archivo de entrenamiento.  %
% coef_file - el archivo que contiene el numero de coeficientes a
clasificar
% nCoef - el num de coeficientes mfccs (debe de ser coherente con la
matriz mfccs) %
%
%%%%% SALIDAS %%%%%
% clasNum - el vector de las letras identificadas, en numerico (A es 1, B
es 2 y asi... de acuerdo al orden en el archivo de entrenamiento) %?????
% nFonemas - el numero de mfcss clasificados (coherente con 'n' de mfccs
en la entrada)????? %
%
[mfccs_rows mfccs_cols]=size(mfccs);
%Esto es para sacar filas y columnas de la matriz de mfcss que voy a
revisar %

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%% IMPORTAR EL ARCHIVO DE ENTRENAMIENTO
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
[file msg]=fopen(training_file);
A=textscan(file,'%s');
[train_rows train_columns] = size (A{1,1});

for i=1:train_rows
   C=regexp(A{1,1}{i,1},',','split');
   for j=1:(4)
       B{i,j}=C{j};
   end
end

[file msg]=fopen(coef_file);
A=textscan(file,'%s');
[coef_rows coef_columns] = size (A{1,1});

for i=1:coef_rows
   C=regexp(A{1,1}{i,1},',','split');
   for j=1:(1)
       D{i,j}=C{j};
   end
end

%EN ESTE PUNTO TENGO:
%A - Matriz auxiliar que tiene los archivos de entrada
%B - Matriz donde esta el arbol de desicion <-- ESTE ES EL QUE SE USA
%C - Matriz auxiliar para sacar los coeficientes
%D - Matriz con los diferentes coeficientes%


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%% CLASIFICACION DE FONEMAS POR ARBOL DE DESICION            %%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
%creamos dos matrices donde guardar los resultados, una de letras y otra
de numeros
    clasificacion=cell(mfccs_rows,1);
    clasNum=zeros(mfccs_rows,1);

    for i=1:mfccs_rows
        %inicializamos el valor selected en 1 (buscar en el primer nodo)
        selected=1;
        sub_A=mfccs(i,:);
        sub_B=B;

        terminar=0;
        filaArbol=1;

        while(terminar==0)
            %CICLO DE IR NAVEGANDO POR EL ARBOL%
            arbol_signo=B{filaArbol,2};
            arbol_coef=str2double(B{filaArbol,1});
            arbol_valor=str2double(B{filaArbol,3});

            valor_test=str2double(sub_A(arbol_coef));

            coincidencia=0;
            if(arbol_signo=='<=')
                if(valor_test<=arbol_valor) %revisar <=
                    coincidencia=1;
                    end
            else
                if(valor_test>arbol_valor) %revisar >
                    coincidencia=1;
                    end
            end

            if(coincidencia==1)
                if(isempty(B{filaArbol,4})) %revisar que no se haya
llegado al final de la rama%
                    filaArbol=filaArbol+1;
                else %se llego al final de la rama y se tiene una
clasificacion %
                    letra=B{filaArbol,4};
                    terminar=1;
                end
            else
                nextFila=0;
                while(nextFila==0)
                   filaArbol=filaArbol+1;
                   arbol_ncoef=str2double(B{filaArbol,1});
                   arbol_nvalor=str2double(B{filaArbol,3});

                   if((arbol_coef==arbol_ncoef) &&
(arbol_valor==arbol_nvalor))
                       nextFila=1;
                   end
                end
            end
        end
```

```
        %Asignamos la letra y el numero a la salida
        clasificacion{i}=letra;
        for j=1:coef_rows
            if(letra==D{j,1})
                classnum(i)=j;
            end

        end
    end
    nFonemas=mfccs_rows;
end
```

Algoritmo de Naive Bayes:

```
function [clasificacion clasNum
nFonemas]=classNaiveBayes(mfccs,training_file,nCoef)
% Uso: Entregar dos vectores de clasificacion y el numero de fonemas
entregados dada una tabla de mfccs, un archivo de lectura y el # de
coeficientes %
%      Uno de los vectores entrega las letras, otro entrega los numeros
relacionados a cada letra
%
%%%%% ENTRADAS %%%%%
% mfccs - Una matriz n * m conteniendo un conjunto de mfcss ('n'
conjuntos de 'm' coeficientes mfcc) %
% training_file - el archivo de entrenamiento %
% nCoef - el num de coeficientes mfccs (debe de ser coherente con la
matriz mfccs) %
%
%%%%% SALIDAS %%%%%
% clasNum - el vector de las letras identificadas, en numerico (A es 1, B
es 2 y asi... de acuerdo al orden en el archivo de entrenamiento) %?????
% nFonemas - el numero de mfcss clasificados (coherente con 'n' de mfccs
en la entrada)????? %
[mfccs_rows mfccs_cols]=size(mfccs);

%%%%% IMPORTAR EL ARCHIVO DE LOS PROMEDIOS DEL ENTRENAMIENTO %%%%%
[file msg]=fopen(training_file);
A=textscan(file,'%s');
[train_rows train_columns] = size (A{1,1});
for i=1:train_rows
   C=regexp(A{1,1}{i,1},',','split');
   for j=1:((2*nCoef)+2)
       B{i,j}=C{j};
   end
end
%EN ESTE PUNTO TENGO:
%A - Matriz auxiliar que tiene los archivos de entrada
%B - Matriz donde estan todos los valores de entrenamiento <-- ESTE ES EL
QUE SE USA
%C - Matriz auxiliar para sacar los coeficientes
```

```
%%%%% CLASIFICACION DE FONEMAS POR NAIVE BAYES            %%%%%
%creamos dos matrices donde guardar los resultados, una de letras y otra
de numeros
    clasificacion=cell(mfccs_rows,1);
    clasNum=zeros(mfccs_rows,1);

    for i=1:mfccs_rows
        %empezamos suponiendo que el primero es el mas cercano
        selected=1;
        %%%%% SE UTILIZA UNA FUNCION EXTERNA LLAMADA probBayes.m %%%%%
        maxPRob=probBayes(mfccs(i,:),B(1,:),nCoef);

        %Revisamos el resto de las opciones de clasificacion, si hay una
probabilidad mayor
        %actualizamos para que ese sea el valor mayor elegido
        for j=2:train_rows
           testProb= probBayes(mfccs(i,:),B(j,:),nCoef);
           if(testProb>maxPRob)
               selected=j;
               maxPRob=testProb;
           end
        end
        %Ya con el valor mayor, asignamos al conjunto de mfcc
        clasificacion{i}=B{selected,(2*nCoef)+1};
        clasNum(i)=selected;
    end
    nFonemas=mfccs_rows;
end
```