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

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

"POR MI PATRIA Y POR MI BIEN"

UN ACERCAMIENTO FONOTÁCTICO A LA LENGUA TÉENEK HACIENDO USO DEL ALFABETO FONÉTICO INTERNACIONAL

OPCIÓN I
TESIS PROFESIONAL

Que para obtener el Título de
Maestría en Ciencias de la Computación

Presenta
I.S.C. Manuel Alejandro Jiménez Quintero
G08070756

Director
Dr. Arturo Hernández Ramírez

Cd Madero, Tamps. Noviembre 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 30 de Septiembre de 2015.

OFICIO No.: U5.220/15
AREA: DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN
ASUNTO: AUTORIZACIÓN DE IMPRESIÓN DE TESIS

ING. MANUEL ALEJANDRO JIMÉNEZ QUINTERO
NO. DE CONTROL G08070756
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 :	DRA. LAURA CRUZ REYES
SECRETARIO :	DR. JUAN JAVIER GONZÁLEZ BARBOSA
VOCAL :	DR. ARTURO HERNÁNDEZ RAMÍREZ
SUPLENTE	M.C. JOSÉ APOLINAR RAMÍREZ SALDIVAR
DIRECTOR DE TESIS :	DR. ARTURO HERNÁNDEZ RAMÍREZ
CO-DIRECTOR DE TESIS:	M.C. JOSÉ APOLINAR RAMÍREZ SALDIVAR

Se acordó autorizar la impresión de su tesis titulada:

"UN ACERCAMIENTO FONOTÁCTICO A LA LENGUA TEENEK HACIENDO USO DEL ALFABETO FONÉTICO INTERNACIONAL"

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.
DIVISION DE ESTUDIOS
DE POSGRADO E
INVESTIGACION
I T C M

c.c.p.- Archivo
Minuta

MYCHC 'NLCO ' jar

Índice de contenido

Índice de ilustraciones

Índice de tablas

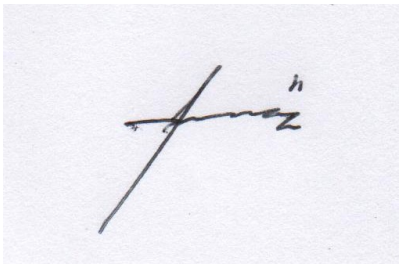
Declaración de Originalidad

Declaro y prometo que este documento de tesis es producto de mi trabajo original, y que hasta donde yo sé, no infringe los derechos de terceros, tales como derechos de publicación, derechos de autor, patentes y similares. Por lo tanto la obra es de mi exclusiva autoría y por lo tanto soy titular de los derechos que surgen de la misma.

Declaro igualmente que hago justo reconocimiento en la tesis a las personas que contribuyeron con su trabajo; dejando en claro que esta tesis es producto de mi propio trabajo con el apoyo permitido de terceros en cuanto a la concepción del proyecto, al estilo de la presentación o a la expresión escrita. Por lo mismo, las citas textuales que he incluido y en los resúmenes que he realizado de publicaciones ajenas, indico explícitamente los datos de los autores y publicaciones.

En caso de presentarse cualquier reclamación o acción por parte de un tercero en cuanto a los derechos de autor sobre la obra en cuestión, como AUTOR, asumiré toda la responsabilidad, respondiendo por cualquier reivindicación, plagio u otra clase de reclamación que al respecto pudiera sobrevenir; y saldrán exentos para todos los efectos mi director y codirector de tesis, así como el Instituto Tecnológico de Ciudad Madero y sus autoridades que actúan como terceros de buena fe.

7 de Noviembre del 2015, Ciudad Madero, Tamaulipas, México.



I.S.C. Manuel Alejandro Jiménez Quintero

Resumen

La tesis desarrollada a continuación, tiene por intención demostrar la viabilidad de una propuesta alternativa para la creación de un Identificador Automático del Lenguaje Hablado (LID, por sus siglas en inglés). Dicho LID cuenta con soporte para inferir palabras pronunciadas en el lenguaje téenek y en el idioma español. Adicionalmente y debido a la arquitectura propuesta, nuevos idiomas pueden agregarse posteriormente.

La anterior característica se consigue gracias al uso de un Reconocedor Automático del Habla (RAH). El RAH en cuestión transcribe los datos de voz al Alfabeto Fonético Internacional (AFI) —el cual es un sistema de notación fonética creado por lingüistas—, consiguiendo con esto crear una cadena de texto independiente del idioma hablado. De lo anterior se puede inferir que el *corpus* usado por éste —el RAH— está compuesto por mono-fonemas. Otra singularidad a puntualizar es que este reconocedor considera un análisis fonético utilizando los Coeficientes Cepstrales en Frecuencia de Mel (MFCC) como característica observable de la señal de audio.

Las ideas antes descritas fueron implementadas en un sistema desarrollado con un lenguaje de programación de alto nivel (Python) y nombrado con el nombre código «Kibo». Este desarrollo sistematiza y pone en orden los distintos procedimientos requeridos para lograr la identificación del idioma en que las muestras de audio fueron pronunciadas.

Por último, la decisión de trabajar con la lengua téenek fue dada por motivaciones sociales. La rápida carrera por la globalización en adición con la cada vez más cruda marginación de las poblaciones indígenas está provocando que este idioma esté en proceso de extinción. Continuar y crear nuevas investigaciones y tecnología hacia el sector indígena en estos tiempos —y en el futuro— debería ser prioridad.

Abstract

The thesis developed below, aims to demonstrate the feasibility of an alternative proposal for the creation of a Language Identifier (LID). This language identifier has support for inferring words spoken in the huasteco language and spanish. Additionally, due to the proposed architecture, new languages can be added later.

The above feature is achieved through the use of an Automatic Speech Recognizer (henceforth, RAH). The RAH in question converts the speech data into text in the International Phonetic Alphabet (from now on, AFI) —which is a system of phonetic notation created by linguists—, achieving with this create a string of text that no dependent of the language. From the above it can be inferred that the corpus used by it —the RAH—, is composed of mono-phonemes. Another singularity to stand out from this recognizer, is that it contemplate a phonetic analysis using the Mel-frequency cepstral coefficients (MFCC) as observable characteristic of the audio signal.

The aforementioned ideas were implemented on a system developed with a high-level programming language (Python) and named with the codename "Kibo". This development systematizes the required procedures to achieve the identification of the language in which the audio samples were uttered.

Finally, the decision to work with huasteco language was given by social motivations. The globalization in addition to the increasingly stark marginalization of indigenous peoples is causing this language is becoming extinct. Create more technology and carry on in this research line should be a priority for this country to guarantee the quality of life of the indigenous sector in these times and in the future.

Agradecimientos

Por más que extendiera, siempre me haría falta tiempo y palabras para poder expresar lo que necesito decir a continuación. Así que por favor disculpen que sea tan lacónico.

Deseo hacer del conocimiento del lector el inefable apoyo de quienes han sido el gran baluarte en esta etapa de mi vida.

Agradezco infinitamente a mis padres, quienes deben ser unas personas excepcionales ya que han logrado hacer de mi una persona de bien. Por ellos he llegado hasta donde estoy.

A Jessica, quien con su esfuerzo y buen corazón ha conseguido desarrollar en mi virtudes que yo consideraba inexistentes. Es mi anhelo triunfar espectacularmente en la vida para poder devolverle al menos un poco de todo lo que me ha dado.

A Alberto, a Carlos, a Lucía y a Isela por haber demostrado ser amigos incondicionales y porque me han apoyado financiera, intelectual y espiritualmente.

Doy gracias y un aplauso enorme a mi director de tesis, a mis tutores y al CONACYT. El gran soporte de parte de ellos fue indispensable para poder elaborar esta investigación.

También quisiera agradecer a Dios; a Él todo honor y toda gloria.

Por último, pido perdón a todas aquellas personas a las que no menciono por sus nombres aquí pero que gracias a sus talentos fueron esenciales para el desarrollo de esta tesis. En verdad gracias a todos ustedes: mis amigos.

«Hombre soy; nada humano me es ajeno».

— Heautontimorumenos (163 a.C.) por Publius Terentius Afer

Capítulo 1. Introducción

Con el objetivo de poner en contexto al lector, se procederá a definir, en una primera instancia, una breve cantidad de conceptos afines al procesamiento de las señales de voz; como lo son los fonemas presentes en éstas. También se describirá la forma como se representarán en este trabajo, intentando en lo posible, que el entendimiento del problema y la solución propuesta quede asentada de una manera clara y sencilla.

Las diferentes áreas del procesamiento digital de la voz se pueden listar de la siguiente manera de acuerdo con (Hernández Zepeda, 2013): reconocimiento de locutores, *speaker diarization*, reconocimiento automático del habla, identificación automática del lenguaje hablado, síntesis del habla y conversores texto-voz. No son todas, aunque sí que son las áreas donde mayor actividad existe.

La identificación automática del lenguaje a partir de enunciados orales tiene como objetivo determinar el idioma o dialecto de un hablante humano a partir de una muestra de voz (Timoshenko, 2012).

El reconocimiento automático del habla (en adelante RAH) es, en esencia, un proceso de identificación de patrones que permite a una computadora reconocer fonemas y palabras a partir de las características de una señal de voz que se presente (Lleida & Rose, 2000).

Debido a la actual necesidad global de interfaces hombre-máquina (Kraiss, 2006), los sistemas de RAH desempeñan un papel esencial en el suministro de aplicaciones de voz. Durante las últimas tres décadas, el RAH se ha convertido en una tecnología clave en las áreas de procesamiento del lenguaje hablado, así como en los sistemas de reconocimiento de voz multilingüe, sistemas de diálogo hablado, sistemas de comunicación de humano a humano, dictado de documentos, y en sistemas de minería de multimedia. Así pues, se puede notar que la tendencia de los sistemas coetáneos es brindar soporte a múltiples entradas y salidas en distintos lenguajes, especialmente si tales sistemas se emplean en mercados internacionales y/o en ambientes donde su comunidad de usuarios cuente con una gran diversidad lingüística.

Debe ser puntualizado que, aunque el objetivo del RAH es el reconocimiento del discurso del hablante, no es su finalidad comprender el mensaje comunicado, sino tan sólo expresar la muestra de voz en texto. Por lo tanto, se puede entender que el RAH es el paso natural previo a los sistemas de procesamiento de lenguaje natural, tales como Siri® (Apple Inc, 2014) o Google Now® (Google Inc, 2014).

La identificación automática del lenguaje, por otra parte, es un tema que se ha estado desarrollando de forma continua, tal como lo muestran los trabajos de (Schultz & Kirchhoff, 2006), (Timoshenko, 2012), (Alviso Vargas, 2014), entre una basta cantidad de otros trabajos.

Existen una variedad de factores que los humanos y las máquinas pueden usar para discriminar un lenguaje de otro. Los idiomas pueden ser distinguidos mediante el uso que cada uno le dé a su inventario de fonemas y sus realizaciones acústicas. Los fonemas son unidades teóricas básicas postuladas para estudiar el nivel fónico-fonológico de una lengua humana. Es decir, un fonema es cada una de las unidades segméntales postuladas para un sistema fonológico que dé cuenta de los sonidos de una lengua. Cada idioma usa un subconjunto de fonemas a partir de todo el conjunto de posibles sonidos que produce el habla. Aún cuando muchos idiomas tienen en común un subconjunto de fonemas, la frecuencia en que éstos se suceden y la forma en que éstos son realizados varía de lenguaje en lenguaje. Además, no todos los lenguajes tienen las mismas limitaciones fonotácticas, lo que debería, en teoría, facilitar la identificación del lenguaje utilizado al evidenciarse las secuencias de fonemas distintivas de cada uno.

También es posible distinguir lenguajes mediante sus reglas semánticas y sintácticas (Timoshenko, 2012). Sin embargo, hacer uso de éstas implica un alto costo computacional, lo que dificulta su implementación en dispositivos móviles y de recursos limitados.

(House, 1977), a pesar de no usar datos de voz sino sólo transcripciones, demostraron la viabilidad del uso de características acústicas derivadas de amplias categorías fonéticas del habla para identificar idiomas. Tiempo después, (Li & Edwards, 1980) refinaron esa idea para aplicarla a datos de voz reales, utilizando un esquema de segmentación amplia para clasificar los datos en seis clases de fonética acústica: núcleos silábicos, sonantes no vocales, murmullo vocal, fricación sonora, fricación sin voz y segmentos de silencio y baja energía.

Ahora bien, partiendo de las ideas anteriores, es posible aventurarse a considerar un hecho la viabilidad de desarrollar un método para el reconocimiento automático del habla que utilice el reconocimiento de fonemas resultante del tratamiento de datos de voz expresados por hablantes humanos, para poder encontrar la identidad de cada muestra de audio. Sin embargo, las investigaciones actuales han sido enfocadas mayormente en la transcripción de datos de voz a fonemas propios de cada lenguaje. En el presente trabajo se procederá a etiquetar un subconjunto de fonemas, los cuales serán transcritos a un sistema de notación fonética estandarizado (para el caso particular de esta investigación se utilizará el Alfabeto Fonético Internacional [en adelante sólo AFI] debido a que otorga en forma regularizada, precisa y única la representación de los sonidos de cualquier lenguaje oral (International Phonetic Association, 1999)) a partir de una muestra de voz y, de ser posible, realizar la identificación del idioma utilizado en ésta tomando como base esta misma transcripción. Esta hipotética identificación se realizaría mediante la comparación de las palabras previamente transcritas al AFI, las cuales serían obtenidas de una base de datos convenientemente estructurada para este caso. Así pues, el LID dependería de qué tan extensa sea esta base de datos y de su correcta clasificación, resultando en un sistema que podría brindar más sencillamente soporte para más idiomas sin la necesidad de obtener grandes muestras de audio de hablantes de tales lenguajes. A grandes rasgos, sólo se necesitaría un corpus para la transcripción al AFI y una base de datos planos que sustente cada lenguaje al que se pretenda dar soporte.

La presente investigación pretende sentar las bases para un sistema LID cuyas principales ventajas sean su modularidad y escalabilidad. En una primer etapa, un RAH transcribirá los datos de voz de una muestra desconocida a texto, representando los fonemas presentes en la muestra con el AFI. El primer idioma soportado en éste será el huasteco (téenek); lengua hablada en el norte de la costa del golfo de México (más específicamente en los estados de San Luis Potosí, Veracruz y Tamaulipas), y que tan sólo en el 2010, era hablada por más de 161,120 personas (Instituto Nacional de Estadística Geografía e Informática (México), 2011).

Aunque esta investigación será desarrollada con la intención de crear un sistema LID que use un transcriptor de datos de voz (en idioma huasteco) al AFI para su posterior identificación, se deja abierta la posibilidad de crear un sistema de identificación automática del habla que soporte muchos más idiomas.

1.1. Definición del problema

Se requiere desarrollar un sistema para lograr la correcta transcripción de datos de voz a un subconjunto de los fonemas soportados por la especificación del AFI y que sirva de base para la posterior creación de un identificador automático de lenguaje, cuya primera meta sea permitir la correcta identificación del idioma utilizado por un hablante de una lengua autóctona de México. El problema, entonces, reside en definir la estructura, métodos y factores de discriminación de las distintas características acústicas soportadas por el AFI con el fin de obtener una transcripción de calidad, independiente del hablante.

1.2. Objetivo general

Desarrollar un sistema LID que haga uso de transcripciones de muestras de audio provenientes de un hablante del idioma téenek a un subconjunto de fonemas del AFI. Tales muestras de audio se tratan de un conjunto de palabras almacenadas en archivos independientes.

1.2.1. Objetivos específicos

- Definir la estrategia con la cual se abordará el problema de transcribir los datos de voz a texto, ideando tanto los métodos de representación adecuados como los métodos de detección que mejor se ajusten a la identificación de fonemas.

- Obtener algunas muestras de audio con palabras utilizadas en la lengua téenek, con el fin de crear el modelo acústico del sistema.
- Realizar el entrenamiento del sistema.
- Ejecutar un conjunto de pruebas con el fin de medir la calidad de la transcripción realizada.
- Comprobar el grado de efectividad para reconocer los sonidos acústicos correctamente.
- Realizar pruebas en el sistema para encontrar puntos de mejora.

1.3. Justificación

Los sistemas LID varían en niveles de complejidad computacional y requerimientos para el entrenamiento de los datos, todo dependiendo del enfoque y tipo de información usado para distinguir los idiomas.

Una de las características principales para discriminar los idiomas son los fonemas. Cada lenguaje usa un subconjunto de fonemas de un conjunto de todos los posibles sonidos.

Otras características importantes utilizadas para discriminar los idiomas son las reglas de sintaxis y semántica que gobiernan a cada uno de ellos, tratándolos como patrones de oraciones. El problema de este enfoque es que requiere de una existencia amplia de reconocedores de habla para cada vocabulario, por lo que se necesita también considerar costos computacionales al momento de realizar las fases de entrenamiento y prueba.

Por otro lado, existe una reciente línea de investigación que utiliza solamente la acústica de las muestras de audio para realizar la tarea del LID; concretamente lo que utiliza son las bajas frecuencias de las señales de audio para caracterizar al idioma. Este proceso se basa en una hipótesis que establece que en las bajas frecuencias existe información sobre las características suprasegmentales que utilizamos al hablar.

La mayoría de los sistemas LID se basan en información espectral extraída a través de análisis espectrales de tiempo corto de la señal de habla, así como también propiedades acústicas como las unidades de sonido y sus secuencias. En adición a la información espectral, algunos sistemas LID también incorporan información prosódica (Hernández Zepeda, 2013).

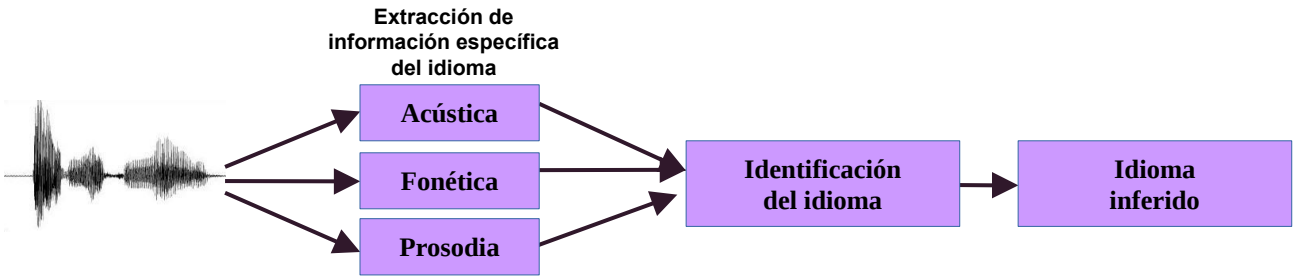
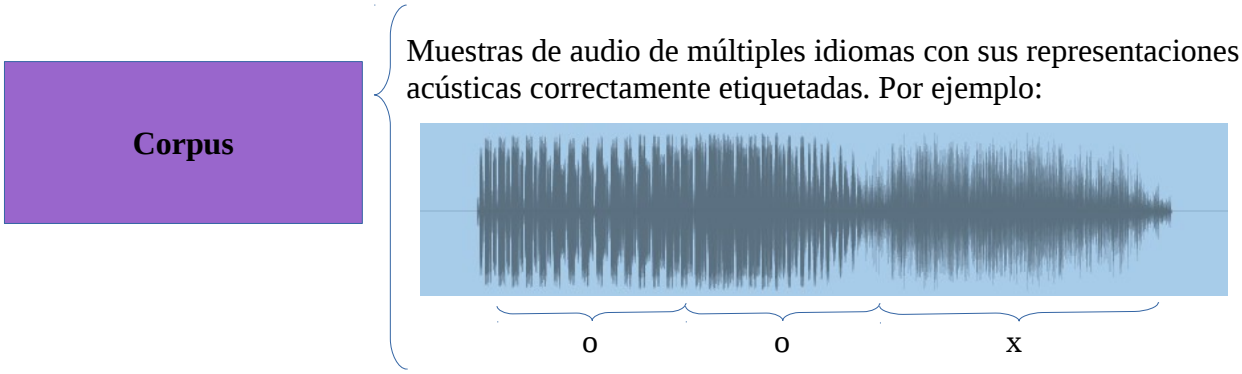


Ilustración 1: *Arquitectura general de un sistema LID usando diferente información discriminatoria.*

Actualmente, un sistema LID que dependa de la existencia de las transcripciones de los fonemas para su fase de entrenamiento probablemente se vería anclado a dar soporte a un único idioma, dado que identificar más lenguajes requeriría un mayor consumo de recursos.



AFI	Ortografía Maya	Español
o:ʃ	oox	tres

Ilustración 2: *Ejemplo de corpus de un sistema LID que dependa de la existencia de las transcripciones de los fonemas para su fase de entrenamiento.*

El *corpus* se compondría, entonces, de las transcripciones de las palabras del idioma téenek. A mayor cantidad de palabras, más robusto sería y el sistema LID sería más estable.

El problema de este enfoque, es que el dar soporte a un segundo idioma significaría crear un

nuevo corpus para contener las transcripciones de palabras de este nuevo idioma. Además, los procesos de extracción de información cambiarían. Básicamente, se estaría creando un sistema LID por cada idioma. Así sucesivamente en caso de agregar nuevos idiomas.

Aunque esta investigación tiene por intención desarrollar un sistema LID para el idioma huasteco, también se busca contar con soporte para el español por su relación regional y cultural tan estrecha. Por lo mismo, para resolver el problema anteriormente mencionado, se propone una arquitectura alternativa de un sistema LID en el cual se pueda brindar soporte para ambos idiomas. Para lograr esto, se reemplaza el *corpus* basado en transcripciones de palabras de un idioma en particular para reemplazarlo por uno que se componga de transcripciones de palabras a un alfabeto fonético estandarizado; en este caso, se hará uso del Alfabeto Fonético Internacional (AFI). Esta arquitectura propuesta se representa en la ilustración 3.

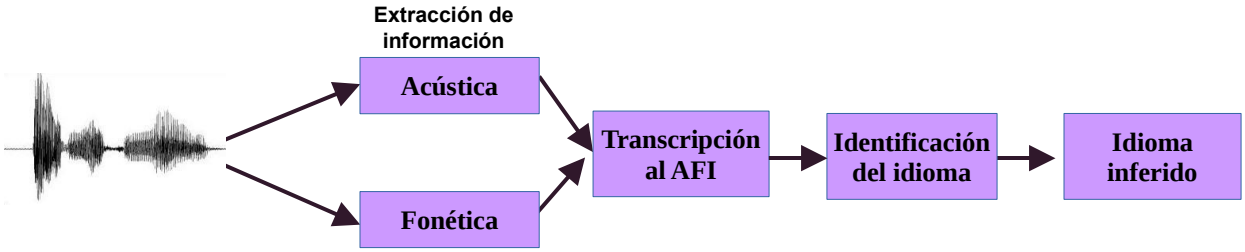
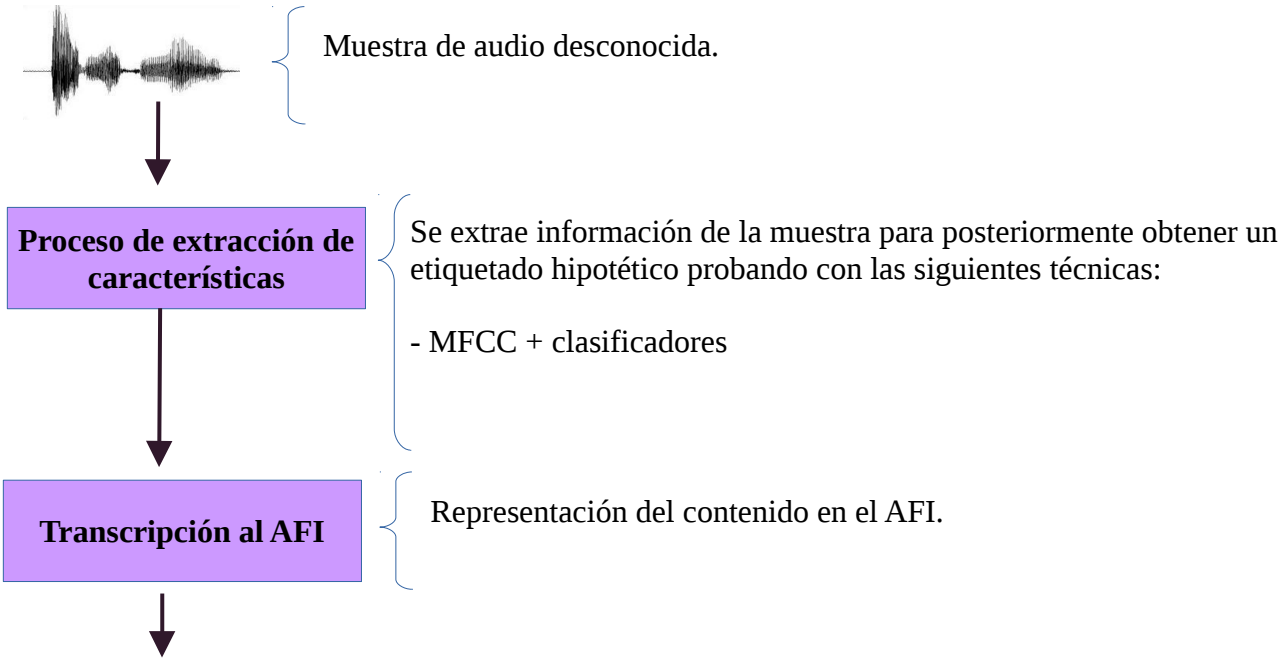


Ilustración 3: *Arquitectura alternativa de un sistema LID usando diferente información discriminatoria.*

El proceso del sistema LID propuesto sería el siguiente:



Proceso de identificación del idioma

Se realizan comparaciones de palabras previamente transcritas al AFI provenientes de distintos idiomas y la transcripción obtenida del proceso anterior.

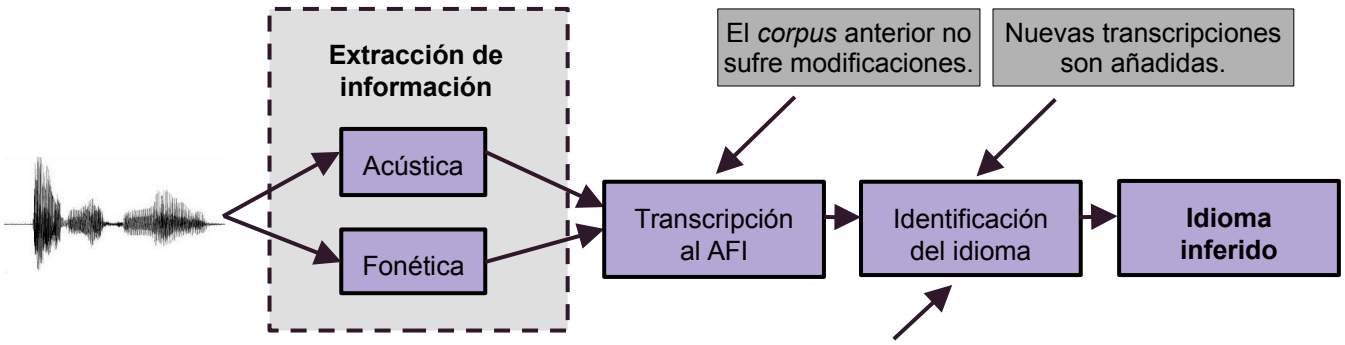
Idioma inferido

Presentación del probable idioma correspondiente a la muestra desconocida.

Resumiendo, el *corpus* es, entonces, una base de datos de MFCC obtenidos a partir de trozos de muestras de voz. Es decir, MFCC correspondientes a «a», a «b'», a «ʧ», etcétera. Como son representaciones acústicas del AFI, no es necesario que las grabaciones de donde se obtengan éstas sean especificas de un idioma, sino tan sólo que contengan las representaciones en sí.

De la muestra de voz a identificar se obtienen las representaciones acústicas del AFI y, mediante una serie de clasificadores, lógica difusa u otros algoritmos de segmentación, se logra obtener la representación completa en AFI de toda la palabra. Obtenida ésta, sólo resta realizar una sencilla búsqueda en una base de datos diseñada con una estructura «representación en AFI/representación en un idioma en particular/idioma» para poder identificar el idioma en el que la palabra fue expresada.

De esta forma, se logra crear un sistema LID preparado para agregar soporte para nuevos idiomas, reutilizando la mayoría de los procesos y el *corpus*. En la ilustración 4 se muestra un ejemplo de este sistema LID con soporte para 3 idiomas.



AFI	Palabra	Idioma
sjete	siete	Español
b'uːk	buuk	HSF
sɛvən	seven	Inglés

Ilustración 4: *Sistema LID propuesto para la identificación de tres idiomas diferentes.*

Hasta el momento no se han reportado implementaciones de algún RAH que realice

transcripciones al AFI. De igual manera, tampoco existe un LID que identifique un idioma a partir de una transcripción desde este alfabeto.

La decisión de escoger al idioma téenek como primer idioma soportado en este RAH ha sido dada por motivos sociales. En México, tal como sucede en el resto del mundo, las minorías sociales han sido relegadas dada la tendencia actual a la globalización (United Nations & Department of Public Information, 2009). En este caso, las minorías sociales a las que se hace referencia son las comunidades de hablantes de lenguas indígenas.

Aún habiendo aumentado la población de hablantes de éstas lenguas, el número de personas bilingües (y que hacían a su vez de traductores) ha disminuido (United Nations & Permanent Forum on Indigenous Issues, 2009), lo que ha ayudado a incrementar la fragmentación. Y, lo que es aún peor, la calidad de vida de estas personas se ha visto drásticamente reducida, puesto que los empleadores y/o encargados de las instituciones de apoyo social son incapaces de comprenderlos. Aunado a esto, los ciudadanos de las zonas urbanizadas son incapaces de entender las necesidades de quienes emigran en busca de oportunidades. Además, la justicia para estas personas es impartida de forma deficiente, puesto que de verse inculpados en algún crimen, serían incapaces de defenderse al no poderse comunicar efectivamente. Debido a todo esto, se ha llegado a relacionar la población indígena con la pobreza (United Nations & Permanent Forum on Indigenous Issues, 2009).

Por si fuera poco, se espera que para dentro de los próximos 100 años el 90% de los idiomas podrían desaparecer. Los idiomas hablados en las comunidades indígenas simplemente ya no se transmiten de una generación a otra. La mayoría de los gobiernos son conscientes de esta crisis de la lengua, pero los fondos que se asignan suelen destinarse solamente a dejar constancia de su existencia y muy poco va a parar a programas de revitalización de los idiomas. La lengua, por otra parte, no es sólo un medio de comunicación, sino que suele estar vinculada con la tierra o la región que tradicionalmente han ocupado en los pueblos indígenas; es un componente esencial de la identidad colectiva e individual de la persona y, por consiguiente, da un sentido de pertenencia y comunidad. Cuando el idioma muera, ese sentido de comunidad se verá severamente deteriorado (United Nations & Permanent Forum on Indigenous Issues, 2009).

Ningún país en el mundo puede darse el lujo de permitir que su acervo lingüístico se deteriore. Precisamente es por esta razón que se ha dado prioridad a una lengua indígena. Si bien, el trabajo a realizar posiblemente no derive en un impacto contundente en la sociedad, sí sienta un precedente para el desarrollo de futuros trabajos sobre el tema.

Capítulo 2. Alcances y limitaciones del proyecto

2.1. Alcances

- Desarrollar un sistema de LID que soporte las siguientes características:
  - Independencia del locutor.
  - Reconocimiento de muestras de voz que contengan únicamente una palabra.
- Voz con una clara pronunciación y dicción.
- Compilar un corpus con los siguientes requerimientos:
  - Muestras de audio provenientes de una misma cantidad de hombres y mujeres.
- Entrenar efectivamente el total de fonemas contendidos en las muestras de audio que hacen uso del alfabeto de la lengua téenek (Larsen, Instituto Lingüístico de Verano, A.C., 1997), y que a su vez, es un subconjunto del AFI, de manera que el sistema pueda reconocerlos correctamente.
- Crear una arquitectura que permita incluir posteriormente soporte para el resto de las representaciones acústicas del AFI.
- Desarrollar una interfaz que facilite el uso de este sistema.

2.2. Limitaciones

- Esta investigación sólo pretende enfocarse en un subconjunto de las representaciones acústicas del AFI. En especifico, los fonemas resultantes de cinco palabras en el idioma téenek. Sin embargo, sería conveniente brindar soporte al total de fonemas utilizados en este lenguaje.
- Las muestras de audio sólo pueden contener una palabra, dejando para trabajo futuro aquellas —muestras— que contengan frases completas.

Capítulo 3. Estado del arte

3.1. Breve historia del reconocimiento de voz por computadora

Se considera que el reconocimiento de voz por computadora es una tarea muy compleja debido a todos los requerimientos que le son implícitos (Oropeza Rodríguez & Suárez Guerra, 2009). Además del alto orden de los conocimientos que en ella se conjugan, deben tenerse nociones de los factores inmersos que propician un evento de análisis individual (estados de ánimo, salud, etc.). Por tanto, en los RAH, ya sea para tareas específicas o generales, es inmensa la cantidad de aspectos a tener en cuenta. La historia esencial de los sistemas de reconocimiento de voz se puede resumir con las siguientes premisas (Oropeza Rodríguez & Suárez Guerra, 2009):

- Los inicios (1950 – 1960).
    - Bell Labs. Reconocimiento de dígitos aislados mono-locutor.
    - RCA Labs. Reconocimiento de 10 sílabas mono-locutor.
    - University College in England. Reconocedor fonético.
    - MIT Lincoln Lab. Reconocedor de vocales independiente del hablante.
- Los fundamentos (1961 - 1970): Comienzo en Japón (NEC labs).
    - Dynamic Time Warping (Alineación Dinámica en Tiempo). Vintsyuk (Soviet Union).
    - CMU (Carnegie Mellon University). Reconocimiento del Habla Continua. HAL 9000.
- Las primeras soluciones (1971 - 1980): El mundo probabilístico.
    - Reconocimiento de palabras aisladas.
    - IBM: desarrollo de proyectos de reconocimiento de grandes vocabularios.
    - Gran inversión en los EE. UU.: proyectos DARPA.
    - Sistema HARPY (CMU), primer sistema con éxito.
- Reconocimiento del Habla Continua (1981 - 1990): Expansión, algoritmos para el habla continua y grandes vocabularios.
    - Explosión de los métodos estadísticos: Modelos Ocultos de Markov.
    - Introducción de las redes neuronales en el reconocimiento de voz.
    - Sistema SPHINX.

- Empieza el negocio (1991 - 2000): Primeras aplicaciones: ordenadores y procesadores baratos y rápidos.
    - Sistemas de dictado.
    - Integración entre reconocimiento de voz y procesamiento del lenguaje natural.
- Una realidad (2000 - actualidad): Integración en el Sistema Operativo.
    - Integración de aplicaciones por teléfono y sitios de Internet dedicados a la gestión de reconocimiento de voz (Voice Web Browsers).
    - Aparece el estándar VoiceXML.

3.2. Antecedentes

(Hauenstein, 1996), al comparar el rendimiento en un SRAH híbrido (HMM-NN: Hidden Markov Models-Neural Networks) utilizando sílabas y fonemas como unidades básicas para el modelo, encuentra que ambos sistemas presentan ventajas que se pueden aprovechar de manera combinada.

(Su-Lin Wu, Shire, Greenberg, & Morgan, 1997) propusieron la integración de información a nivel de sílabas dentro de los reconocedores automáticos del habla para mejorar el rendimiento y aumentar la robustez. La razón de error alcanzada fue del 10% para un corpus de voz de dígitos del corpus de OGI (Oregon Graduate Institute). (Su-Lin Wu, Kingsbury, Morgan, & Greenberg, 1998), adicionalmente, reportan resultados del orden del 6.8% para un corpus de dígitos proveniente de conversaciones telefónicas, haciendo uso de un sistema híbrido fonema-sílaba.

(R. J. Jones, Downey, & Mason, 1997) experimentaron con los modelos ocultos de Markov (HMM - Hidden Markov Models) para obtener las representaciones de las unidades a nivel de sílaba, encontrando que se puede mejorar sustancialmente el rendimiento del sistema de RAH en una base de datos de tamaño mediano al compararlos con modelos mono-fónicos. Logrando un 60% de reconocimiento que lo compararon con un 35% que se obtiene al utilizar mono-fonemas, dejando en claro que las aplicaciones prácticas deben de conformarse por un sistema híbrido.

3.3. Aplicaciones actuales de los sistemas de RAH

En la actualidad, los sistemas de RAH se han ido popularizando y expandiendo rápidamente. Esto se debe, en su mayoría, gracias a la conjunción con sistemas de procesamiento de lenguaje natural, los cuales crean poderosas herramientas que se vuelven cada vez más útiles y eficaces con cada revisión.

Algunos ejemplos de aplicaciones de RAH son CMU Sphinx (Carnegie Mellon University, 2014), Julius (Kawahara Lab., Kyoto University, Information-technology Promotion Agency, Japan, Shikano Lab., Nara Institute of Science and Technology, & Julius project team, Nagoya Institute of Technology, s. f.), Kaldi (Povey et al., 2011), Dragon Dictate (Nuance Communications, Inc., 2014), Siri Personal Assistant (Apple Inc, 2014) y Google Now (Google Inc, 2014), por mencionar algunos. Los últimos dos, cabe mencionar, son en la actualidad sistemas punteros en conjuntar sistemas de RAH con procesadores de lenguaje natural.

3.4. Reconocimiento de voz multilenguaje basado en mono-fonemas

El reconocimiento basado en fonemas se está convirtiendo en la nueva dirección a investigar, ya que son la construcción básica del sonido producido por el lenguaje humano (Yusnita et al., 2011). La mayoría de los lenguajes pueden ser descritos en términos de fonemas, la unidad estructural más pequeña del habla. Se considera, además, que los fonemas son casi los mismos para todos los lenguajes (Hunt & Speech Applications Group, Sun Microsystems Laboratories, 1997). Partiendo de esta idea, se creó el Alfabeto Fonético Internacional, el cual entrega un símbolo para representar cada fonema. Esto es de considerar, ya que, por ejemplo, incluso el idioma inglés se apoya de este alfabeto para representar los fonemas que las 26 letras de su idioma no les es posible representar.

La idea de un sistema de reconocimiento de voz multilenguaje empezó hace poco, como se puede ver en (Manikandan, Venkataramani, Preeti, Sananda, & Sadhana, 2009) y en (Anderson, Dalsgaard, & Barry, 1994), donde se analizó la identificación de fonemas individuales, discriminando entre los distintos grupos de fonemas. Aunque cabe señalar que sólo fueron probados lenguajes europeos con grandes similitudes.

De cualquier modo, el reconocimiento de mono-fonemas resulta ser mucho más versátil que las actuales técnicas que requieren casi un infinito número de palabras para reconocer varios lenguajes, lo que hace a éstos poco prácticos de manejar. Además, mediante el reconocimiento de mono-fonemas, se deja abierta la posibilidad de adaptar nuevos lenguajes.

Debe hacerse notar que este enfoque no cuenta aún con un buen ratio de exactitud, siendo el resultado mejor reportado de 96%, aunque con una muestra bastante pequeña (Runstein & Violaro, 1996).

3.5. Uso del Alfabeto Fonético Internacional en los sistemas RAH

El uso del AFI en los sistemas actuales de RAH es prácticamente nulo. Existen muy tímidas aproximaciones para su uso, tal como se evidencia en (Attanayake, 2012), donde se ha creado una tabla de mapeo para ser usada en el software CMU Sphinx. Sin embargo, este intento de implementación carece de una utilidad compleja ya que CMU Sphinx hace uso de una tabla de mapeo por cada lenguaje, aún tratándose del AFI, debido a las limitaciones propias del software, el cual sólo brinda soporte a alfabetos basados en ASCII, tales como Arpabet (Colaboradores de Wikipedia, 2014b) o SAMPA (UCL Division of Psychology & Language Sciences & Wells, 2005).

Actualmente, la mayoría de los sistemas mencionados brinda soporte para SAMPA como medio de representación de fonemas. No obstante, SAMPA requiere de una tabla de símbolos ASCII para cada lenguaje diferente y del cual éste tenga soporte. Estos lenguajes son: árabe, bosnio, búlgaro, cantonés, croata, checo, danés, holandés, inglés, estonio, francés, alemán, griego, hebreo, húngaro, italiano, noruego, polaco, portugués, rumano, ruso, escocés, serbio, eslovaco, esloveno, español, sueco, tailandés y turco.

Debido a lo anterior, se creó una extensión al alfabeto antes mencionado, llamado X-SAMPA (John Christopher Wells & Department of Phonetics and Linguistics, University College London, 2000), el cual se diseñó con el objetivo de unificar los distintos alfabetos SAMPA y extenderlos para lograr cubrir todos los rasgos que ahora cubre el Alfabeto Fonético Internacional (AFI). Así pues, X-SAMPA es, en esencia, una interpretación en ASCII del AFI.

Se podría decir que X-SAMPA es la mayor aproximación del AFI (aún no siendo en sí éste) en sistemas de RAH hasta el momento, empero, no pasa de ser una representación de fonemas sin la utilidad que se pretende dar en esta investigación, además de seguir estando alejada de ser una forma de evitar pasar por diferentes conversiones de alfabetos hasta llegar a X-SAMPA.

3.6. Modelos más efectivos para identificar el habla

(Colás Pasamontes, 2001) lista los métodos más efectivos para identificar el habla:

- Comparación de Plantillas o Patrones utilizando técnicas de Programación Dinámica (DTW),
- Modelos Ocultos de Markov (HMM) y,
- Redes Neuronales (NN).

3.6.1. Comparación de plantillas o patrones utilizando técnicas de programación dinámica (DTW)

Este modelo consiste en comparar el patrón a reconocer (de entrada) con una serie de plantillas o patrones que representan a las unidades a reconocer. La plantilla no es más que un conjunto de características acústicas ordenadas en el tiempo, y la comparación de patrones incluye un alineamiento temporal no lineal y una medida de distancia. Esta técnica, utilizada tanto para resolver problemas de reconocimiento de habla continua como aislada e incluso con una cierta independencia del locutor, se conoce como DTW (Dynamic Time Warping).

3.6.2. Modelos Ocultos de Markov (HMM)

El modelado estocástico de la señal de habla soluciona el problema que presentaba la técnica de alineamiento de plantillas, proporcionando los mejores resultados hasta la fecha tanto para el reconocimiento de habla aislada como continua y para independencia del locutor.

En el fondo la filosofía de comparación de patrones subyace en este tipo de aproximación al problema pero difiere 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.

Ahora, se utiliza un algoritmo de alineamiento no lineal (Programación Dinámica) conocido como algoritmo de Viterbi, capaz de «alinear» la secuencia de vectores de entrada o índices de un codebook con el conjunto de patrones estocásticos (HMM) que representan las palabras del diccionario, en forma de la probabilidad de que esa secuencia sea observada (generada) por los distintos Modelos Ocultos de Markov.

3.6.3. Redes Neuronales (NN)

Las redes neuronales son estructuras de procesamiento paralelo de información, formadas por numerosos nodos simples conectados entre sí mediante pesos y agrupados en diferentes capas, entre las que se deben distinguir la capa de entrada y la capa de salida. Debido a su naturaleza intrínsecamente no lineal, a su capacidad de clasificación, y sobre todo a la capacidad que tienen para aprender una determinada tarea a partir de pares observación-objetivo sin hacer suposición alguna sobre el modelo subyacente, se han convertido en una de las herramientas más atractivas para la solución del problema del reconocimiento de habla. Hoy en día se han conseguido resultados comparables a los obtenidos con otros métodos ya clásicos como los HMM.

3.7. Lenguas autóctonas de México

El Instituto Nacional de Lenguas Indígenas (INALI), elaboró el «catálogo de lenguas indígenas mexicanas» (Consejo Nacional del Instituto Nacional de Lenguas Indígenas & Gaxiola Moraila, 2007), bajo mandato federal.

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. 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, relativa a la presente síntesis, la atención se centró en la diversidad lingüística correspondiente al habla propia de los pueblos indígenas arraigados en el territorio nacional.

Considerando las investigaciones realizadas hasta el presente, así como las consultas y los propios estudios realizados por el INALI para la elaboración del Catálogo, la realidad lingüística del país es mucho más compleja de lo que en términos generales se ha creído hasta ahora. Además, ha resultado impreciso, al parecer desde siempre, el uso que se le ha dado al concepto lengua en torno a la diversidad lingüística mexicana; por ejemplo, a partir de la época virreinal, o quizá desde antes, se difunde la creencia de que los pueblos indígenas hablan «una sola lengua» (altamente uniforme en todos sus componentes), sin advertir, las más de las veces, la existencia de distintas clases de variantes lingüísticas, explicables bien sea por razones geográficas, genealógicas o sociales, como ocurre en todo el mundo.

3.8. Investigación actual sobre conservación de lenguas autóctonas de México

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 (Unión Editorialista, S.A. de C.V., 2013).

También se publicó la convocatoria del Premio de Literaturas Indígenas de América que se entregó durante la edición 2013 de la Feria Internacional del Libro en Guadalajara (Unión Editorialista, S.A. de C.V., 2013).

Para 2018 se fijó la meta de tener traducida la Constitución en las 68 lenguas; se elaborarán diccionarios y se estandarizará la escritura de las lenguas. Hasta ahora, el INALI tiene ocho normas terminadas de Chiapas, y se tienen en proceso 18 más, de 12 estados de la República. El director del Instituto menciona que se tiene una visión para 2030, en donde las lenguas tengan presencia en todos los ámbitos públicos o privados (Unión Editorialista, S.A. de C.V., 2013).

3.9. Trabajos relacionados a la identificación de habla en lenguas autóctonas de México

3.9.1. Técnicas para el reconocimiento de voz en palabras aisladas en la lengua náhuatl (2009) (Flores Paulín, 2009)

Esta tesis, presentada por Juan Carlos Flores Paulín, 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 Mel (MFCC).

3.9.2. Corpus de las Lenguas Indígenas Tének, Náhuatl y Xi'iuy para la Identificación Automática del Lenguaje Hablado (2013) (Hernández Zepeda, 2013)

Elaborada por Carlos Arturo Hernández Zepeda, 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.

3.9.3. Reconocimiento de habla en palabras aisladas en lenguas indígenas de San Luis Potosí (2013) (Alviso Vargas, 2014)

Esta tesis fue escrita por Jesús Álviso Vargas. Su investigación tuvo como producto final un RAH para el idioma téenek. Para lograrlo, usó diferentes técnicas de clasificación y modelos probabilísticos, así como los Coeficientes Cepstrales Mel como forma de representación de las muestras de voz.

3.9.4. On the Development of Speech Resources for the Mixtec Language (2013) (Caballero-Morales, 2013)

(Caballero-Morales, 2013) 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.

3.10. Idioma huasteco o téenek

El idioma téenek es hablado en el norte del estado de Veracruz y en algunos municipios de San Luis Potosí. También es posible encontrar hablantes de esta lengua en el sur de Tamaulipas (véase la ilustración 5). Esta lengua no se ha transmitido a las nuevas generaciones desde hace ya varios años, por eso es una lengua indígena en peligro de desaparición. De acuerdo con el censo de 2011 (Instituto Nacional de Estadística Geografía e Informática (México), 2011), hay un total de 166 952 hablantes de huasteco; aunque es difícil confiar en la exactitud de la cifra.

Ilustración 5*: Localización aproximada de donde puede encontrarse hablantes del idioma huasteco. Mapa creado mediante la tecnología de Google Maps (©2015 Google).*

El huasteco del sureste (HSF), junto con el de San Luís Potosí (HVA), el veracruzano (HUV) y el extinto chicomucelteco (COB) forman la rama huastecana de la familia maya. Los lugares donde se hablan las primeras tres lenguas están marcados en la ilustración 5. El chicomucelteco (COB), ahora extinto, se habló en Chiapas (Kondic, 2012).

3.11. Propuesta de sistema ortográfico para el idioma huasteco y representación de las mismas en AFI

Las lenguas huastecas hasta ahora han sido escritas con más o menos cinco ortografías diferentes. La creación de un sistema ortográfico para esta lengua fue una parte de un proyecto de documentación y descripción del huasteco del sureste realizado por (Kondic, 2012). A continuación se presentan las grafías que fueron propuestas para la escritura del HSF (huasteco del sureste) y que en adelante servirán para esta tesis en particular, fomentando la estandarización de la misma. Además, se incluye la representación de los mismos en AFI con la finalidad de tener una visión más amplia del tamaño del subconjunto que éstos abarcan en el alfabeto completo.

AFI	Grafía	AFI	Grafía	AFI	Grafía
a	a	k	k	t	t
aː	aa	k’	k’	t’	t’
b’	b	kw	kw	θ	th
ʧ	ch	k’w	k’w	t͡s	tx
ʧ’	ch’	l	l	u	u
e	e	m	m	uː	uu
eː	ee	n	n	w	w
x	j	o	o	ʃ	x
i	i	oː	oo	y	y
iː	ii	p	p	ʔ	’

Tabla 1*. Correspondencia entre sonidos y grafías del HSF.*

Capítulo 4. Marco teórico

4.1. La voz

La voz es una forma de energía de naturaleza analógica. Puesto que es en la laringe donde se forma el sonido de la voz, estamos hablando de ondas sonoras producidas por diferentes presiones de aire, mismas que están dadas por todo el conjunto de órganos que intervienen en el proceso, desde la nariz hasta los pulmones, y que determinan el tipo de sonoridad de la voz (Contreras Morales, 2007).

Por lo tanto, la voz puede ser cuantificada electrónicamente debido a que se trata de un fenómeno relacionado con la presión y eso es algo que se puede medir. Las frecuencias de la voz quedan dentro del rango de cero a 10 KHz para efectos de acaparar todas las frecuencias que ésta genera (Contreras Morales, 2007).

Idealmente, una señal de voz es representada por medio de una gráfica, donde se pueden apreciar tres elementos fundamentales:

1. La amplitud, que representa la intensidad de la voz, también conocido como volumen.
2. 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.
3. El tiempo, el cual es una de las formas más comunes de representar una gráfica como variable independiente *(x)*.

4.2. Reconocimiento automático del habla

La comunicación oral ha sido, al menos de momento, la forma más efectiva que ha tenido el ser humano para comunicarse. Sin embargo, en el área computacional, la comunicación de forma escrita es la que prevalece debido a que aún no se ha encontrado una manera eficiente para lograr tal empresa. Aunque eso no significa que no se haya estado trabajando en ello.

Uno de los primeros pasos en el camino a conseguir una comunicación total con las máquinas es la creación de un reconocedor automático del habla (RAH); el cual consiste en un sistema, que mediante distintos procedimientos, sea capaz de captar, procesar e identificar un conjunto de palabras pronunciadas por un hablante.

4.2.1. Niveles de comprensión de un sistema de RAH

(Navarro Mesa, 2005) afirma 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 que se pueden ver en la ilustración 6.

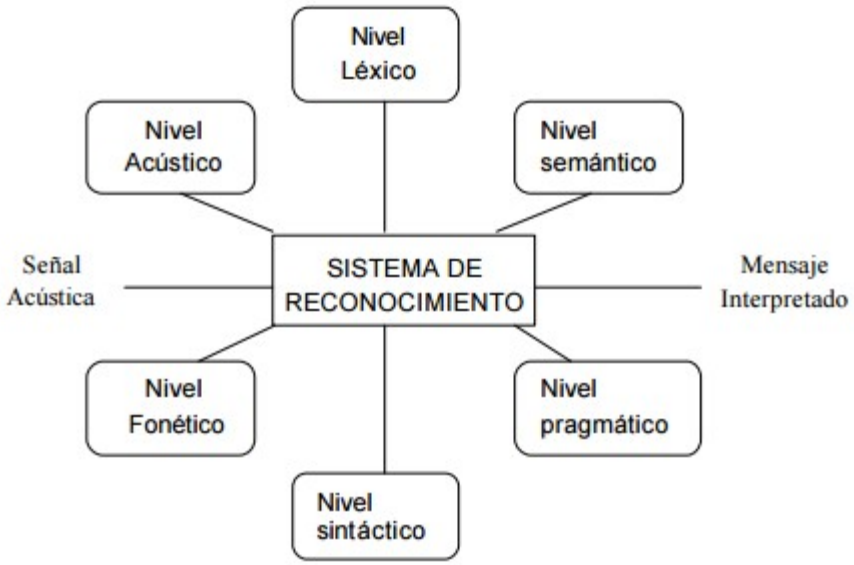


Ilustración 6: *Planteamiento general del reconocimiento automático del habla.*

- 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, actúa 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.

4.2.2. Restricciones de los sistemas de RAH

Dado a la severa complejidad de los sistemas de RAH, se ha buscado restringir éstos con la finalidad de facilitar el proceso, y eventualmente, ir aumentando el nivel de complejidad de los mismos. (Navarro Mesa, 2005) clasifica estas restricciones en dependencia del locutor, tipo de habla y talla del léxico.

Dependencia del locutor

La dependencia 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:

1. Mono-locutores, diseñados para el funcionamiento de un solo locutor.
2. Multilocutores, los cuales se componen de un conjunto restringido de locutores
3. 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 sistemas mono-locutores 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 sus 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.

4.2.3. Métodos de representación de una señal de voz

El hecho de trabajar con vectores de características, sin duda, constituye un elemento clave en un sistema de RAH. El concepto mismo de vector de rasgos o de características es lo que dota al reconocimiento automático de su elegancia y de su enorme potencial práctico al reducir la extraordinaria y diversa 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 de la siguiente forma tal como lo describe (Navarro Mesa, 2005):

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.

Los coeficientes del predictor son los coeficientes de ponderación utilizados en la combinación lineal. La utilidad de este método se fundamenta tanto en su habilidad para proporcionar estimaciones precisas de los parámetros de voz como en su relativa rapidez de cálculo. El método de predicción lineal es matemáticamente preciso, y, además, es sencillo y simple de implementar mediante algoritmos matemáticos y en plataformas hardware.

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 técnicas basadas en las Transformadas de Fourier, las técnicas de procesado homomórfico y el análisis cepstral. Respecto al *cepstrum* hay que decir que su carácter no paramétrico no está del todo claro, si bien sus coeficientes aportan una envolvente del espectro, éstos no se ajustan a un modelo predeterminado.

Un banco de filtros puede ser considerado como un modelo aproximado de las etapas iniciales de transducción en el sistema auditivo humano. Existen dos motivaciones principales que impulsan y justifican la utilización de esta técnica de representación. Primera, la posición de máximo desplazamiento a lo largo de la membrana basilar para estímulos tales como tonos puros, es proporcional al logaritmo de la frecuencia del tono. Segunda, experimentos sobre la percepción humana han demostrado que las componentes de frecuencias determinadas de un sonido complejo, incluidas dentro de una cierta banda de frecuencias particular, no pueden ser identificadas individualmente. Sin embargo, cuando una de las componentes de este sonido cae fuera de esta banda, puede ser distinguido individualmente. Esta banda de frecuencias es conocida como banda crítica.

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

El análisis cepstral es un caso particular de procesado homomórfico. Desde su introducción, a principios de los años setenta, las técnicas homomórficas de procesado de señal han sido ampliamente utilizadas en aplicaciones de reconocimiento del habla. Los sistemas homomórficos son una clase de sistemas no lineales que obedecen a un principio general de superposición.

Los sistemas homomórficos son de gran utilidad para el procesado de voz porque ofrecen un método eficaz para separar la estructura fina y los formantes del espectro de la señal de voz. En el modelo lineal de producción de voz, el espectro compuesto de la señal de voz, expresado mediante

transformada de Fourier, consiste en una señal de excitación (producida en el sistema subglotal), filtrada mediante un filtro lineal variante en el tiempo que representa la configuración del tracto vocal.

El espectro del tracto vocal puede ser separado de la señal de excitación utilizando técnicas homomórficas de procesado de señal. Este método no es válido para todas las clases de sonidos de voz, como los sonidos fricativos, donde la excitación se produce por encima de la glotis.

El proceso de separar las componentes cepstrales en estos dos factores se denomina liftado (*liftering* en inglés, derivado de *filtering*, filtrado) y consiste sencillamente en un enventanado. El análisis cepstral permite, de este modo, convertir la ecuación de convolución en suma.

Para obtener el *cepstrum*, en primer lugar se calculan las magnitudes espectrales logarítmicas, para calcular, posteriormente, la transformada inversa de Fourier del espectro logarítmico.

Generalmente, debido a que el *cepstrum* es calculado utilizando un operador no lineal (función logaritmo), se le considera especialmente sensible a ciertos tipos de ruido y distorsiones en la señal. Por ello, para aplicaciones en medios ruidosos se prefieren parámetros cepstrales derivados de un estimador espectral de alta resolución, como, por ejemplo, un análisis de predicción lineal.

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 la noción de 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.

4.3. Corpus

El término *corpus* de habla, según (Hernández Zepeda, 2013), 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 Zepeda, 2013) define *corpus* de habla 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, metadatos y/o documentación almacenados en un medio digital».

4.4. Ruido

Se denomina ruido a toda señal no deseada que se mezcla con la señal útil que se quiere transmitir. Es el resultado de diversos tipos de perturbaciones que tiende a enmascarar la información cuando se presenta en la banda de frecuencias del espectro de la señal, es decir, dentro de su ancho de banda.

El ruido se debe a múltiples causas: a los componentes electrónicos (amplificadores), al ruido térmico de los resistores, a las interferencias de señales externas, etc. Es imposible eliminar totalmente el ruido, ya que los componentes electrónicos no son perfectos. Sin embargo, es posible limitar su valor de manera que la calidad de la comunicación resulte aceptable.

4.5. Transformada discreta de Fourier

Una sucesión periódica puede ser representada por series de Fourier. Con la correcta interpretación, la misma representación puede ser aplicada a sucesiones de duración finita. La representación de Fourier resultante para sucesiones de duración finita es lo que se conoce como la transformada discreta de Fourier (TDF).

Se puede representar una secesión de duración finita de largo N por una sucesión periódica con periodo N, un periodo de la cual es idéntica a la sucesión de duración finita.

Consideremos una sucesión de duración finita $x(n)$ de largo N de forma que $x(n)=0$ excepto en el intervalo $0 \leq n \leq (N-1)$. Claramente una secuencia de largo M menor que N también puede considerarse de largo N, teniendo amplitud cero los últimos $(N-M)$ puntos del intervalo. La sucesión periódica correspondiente de periodo N, para la cual $x(n)$ es un periodo, será denotada por $\hat{x}(n)$ y está dada por

$$\hat{x}(n)=\sum_{r=-\infty}^{\infty} x(n+rN)$$

Dado que $x(n)$ es de largo finito N no hay sobrelapamiento entre los términos $x(n+rN)$ para diferentes valores de r. Así, la ecuación anterior puede ser escrita alternativamente como

$$\hat{x}(n)=x(\mathrm{n}\ \%\ \mathrm{N})$$

donde $\%$ indica la operación *módulo*. La sucesión de duración finita $x(n)$ es obtenida a partir de $\hat{x}(n)$ extrayendo un periodo; esto es:

$$x(n) = \begin{cases} \tilde{x}(n), & 0 \leq n \leq N-1 \\ 0, & \text{de otro modo} \end{cases}$$

Por conveniencia en la notación, es útil definir la sucesión rectangular $R_N(n)$ dada por

$$\mathrm{R_N(n)} = \begin{cases} 1, & 0 \leq n \leq N-1 \\ 0, & \text{de otro modo} \end{cases}$$

Con esta notación la ecuación de arriba puede escribirse como

$$x(n) = \hat{x}(n) R_N(n)$$

Los coeficientes de la series de Fourier discreta $\hat{X}(k)$ de la sucesión periódica $\hat{x}(n)$ son en sí mismas una sucesión periódica con periodo N. Para mantener una dualidad entre los dominios del tiempo y frecuencia, se escogerán los coeficientes de Fourier que se asocian con la sucesión de duración finita correspondiente a un periodo de $\hat{X}(k)$. Así, con $X(k)$ denotando los coeficientes de Fourier que se asocian con $x(n)$, $X(k)$, $\hat{X}(k)$ están relacionados por

$$\hat{X}(k) = X(k \% N)$$

$$X(k) = \hat{X}(k) R_N(k)$$

$\hat{X}(k)$ y $\hat{x}(k)$ están relacionadas por

$$\hat{X}(k) = \sum_{n=0}^{N-1} \hat{x}(n) W_N^{kn}$$

$$\hat{x}(k) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}(k) W_N^{-kn}$$

Ya que las sumas en las ecuaciones anteriores involucran solamente el intervalo entre 0 y $N-1$, se deduce que

$$X(k) = \begin{cases} \sum_{n=0}^{N-1} x(n) W_N^{kn}, & 0 \leq k \leq N-1 \\ 0, & \text{de otro modo} \end{cases}$$

$$x(n) = \begin{cases} \frac{1}{N} \sum_{k=0}^{N-1} X(k) W_N^{-kn}, & 0 \leq n \leq N-1 \\ 0, & \text{de otro modo} \end{cases}$$

El par de transformadas dadas por las ecuaciones se conocen como la transformada discreta de Fourier (TDF), con la primer ecuación representando la transformada de análisis y la segunda ecuación la transformada de síntesis.

4.6. Transformada de coseno discreta

La transformada de coseno discreta (TCD) es una transformada basada en la Transformada de Fourier discreta, pero utilizando únicamente números reales. La transformada de coseno discreta expresa una sucesión finita de varios puntos como resultado de la suma de distintas señales sinusoidales (con distintas frecuencias y amplitudes). Como la TDF la TCD trabaja con una serie de números finitos, pero mientras la TCD solo trabaja con cosenos la TDF lo hace con exponenciales complejas.

Formalmente, la transformada de coseno discreta es una función lineal invertible $f : \mathscr{R}^N \rightarrow \mathscr{R}^N$, (donde $\mathscr{R}$ denota el conjunto de los números reales), o en forma equivalente a una matriz cuadrada de NxN . Las variantes más usadas son la TCD-I y la TCD-II. La TCD-III se conoce popularmente como la TCDI (transformada inversa). Cada una de estas posibles variaciones es debida a la periodicidad y el tipo de simetría aplicada a las muestras originales. A continuación se enuncian las fórmulas más usadas para la TCD:

TCD-I:

$$x_0 + (-1¿^j x_{n-1}) + \sum_{k=1}^{n-2} x_k \cos\left[\frac{\pi}{n-1} j\right]$$

$$f_j = \frac{1}{2}¿$$

TCD-II:

$f_j = \sum_{k=0}^{n-1} x_k \cos\left[\frac{\pi}{n} j\left(k + \frac{1}{2}\right)\right]$, es la forma más típicamente utilizada

TCD-III:

$$f_j = \frac{1}{2} x_0 + \sum_{k=1}^{n-1} x_k \cos\left[\frac{\pi}{n}\left(j+\frac{1}{2}\right)k\right]$$

TCD-IV:

$$f_j = \sum_{k=0}^{n-1} x_k \cos\left[\frac{\pi}{n}\left(j+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\right]$$

4.7. Coeficientes Cepstrales en la Frecuencia Mel (MFCC)

Un problema que surge inmediatamente al trabajar con archivos de audio es la gran variabilidad de los valores en el dominio de la amplitud. Tratar de identificar patrones en una secuencia de valores es una tarea titánica, además que no representa la forma en cómo el oído humano escucha, que es en el dominio de la frecuencia en lugar en el dominio de la amplitud. Los coeficientes cepstrales en la frecuencia Mel (MFCC, *Mel* Frequency *Cepstral* Coefficients) tratan de resolver este problema. Para lograr esto se intercambian segmentos/ventanas de la señal original con 13 valores/coeficientes. En la práctica una configuración común es reducir 160 muestras de amplitud a 13 valores, que significa que una señal de audio original queda representada por aproximadamente 79,950 valores.

Los MFCC no solo están motivados por su efecto en comprimir los datos, sino también tienen una motivación matemática, biológica y de procesamiento de señales (Meza Ruiz, 2013).

4.7.1. El dominio de la frecuencia

Como su nombre lo sugiere la frecuencia tiene que ver con periodicidad, en este caso, periodicidad matemática. En términos matemáticos toda función en el tiempo puede ser representada por una combinación de funciones periódicas (estas son funciones que se repiten de manera infinita, por ejemplo una senoide). El dominio de la frecuencia va a permitir identificar que «tonos»/frecuencias la componen; como no se trata de una función periódica, está limitada en el tiempo, se obtendrán estos «tonos» para diferentes segmentos de la señal de audio. Es decir, se obtendrán como cambian los tonos que componen a la señal conforme avanza el tiempo.

4.7.2. El algoritmo

Los MFCC van a representar a elementos en el dominio de la frecuencia, sin embargo, se van a utilizar alguna intuición fonética para poderlos reducir en número a valores muy representativos. Los pasos a seguir son los siguientes:

1. Dividir la señal en ventanas del mismo tamaño (se recomienda que estén intercaladas).
2. De preferencia se usa una ventana de Hamming.
3. Para cada ventana calcular la transformada discreta de Fourier.
4. Con los valores de la transformada de Fourier calcular el periodograma que representa la energía por cada valor de frecuencia.
5. Filtrar las frecuencias y energías del periodograma usando un banco de filtros Mel.
6. Calcular el logaritmo de la energía por cada uno de los filtros.
7. Calcular la transformada coseno discreta de las energías.
8. Quedarse con los valores 2 a 13 de esta transformada, que representan a los 13 MFCC.

4.7.3. Ventaneo

Debido a que la señal de audio no es periódica, no es posible calcular los componentes en el dominio de la frecuencia (ya que las herramientas que se utilizan sólo son aplicables en señales periódicas). Para solucionar este problema, se divide la señal en varios segmentos y se supone que para cada segmento, la señal es periódica. A estos segmentos se le conocen como ventanas.

Una vez que se tienen los componentes para una ventana es posible moverse a la siguiente ventana en el tiempo y calcular los componentes para esa nueva ventana. De manera práctica se obtendrán una secuencia de cómo evolucionan los componentes en el tiempo.

La ilustración 7 representa la primera y segunda ventana de una grabación de manera continua con diferentes colores.

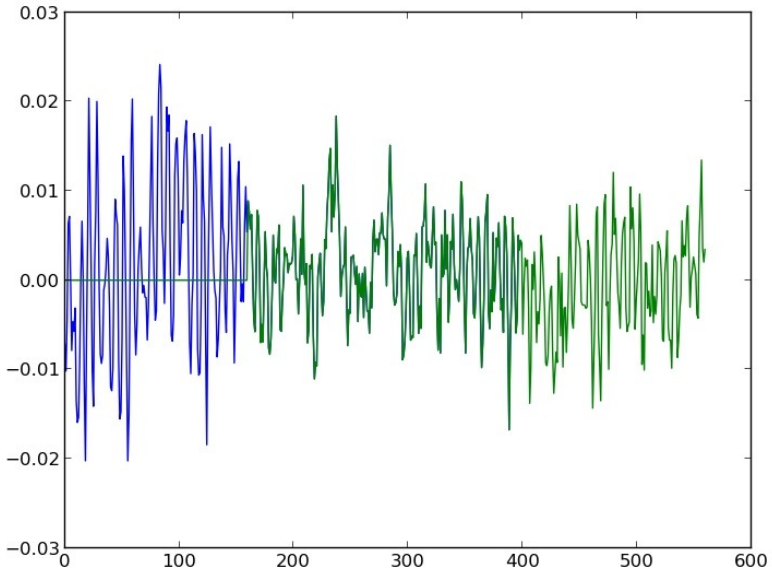


Ilustración 7: *Ventaneo de una señal de voz.*

Como se puede observar ambas ventanas se sobreponen y comparten valores entre sí (del 160 al 400).

4.7.4. Hamming

La ventana Hamming ayuda a suavizar la información en la ventana de la señal original, esto ayudará a filtrar frecuencias espurias que aparecen por cortar abruptamente la señal. La ilustración 8 muestra la ventana Hamming, una ventana de alguna señal y la señal después de pasarla por la ventana Hamming.

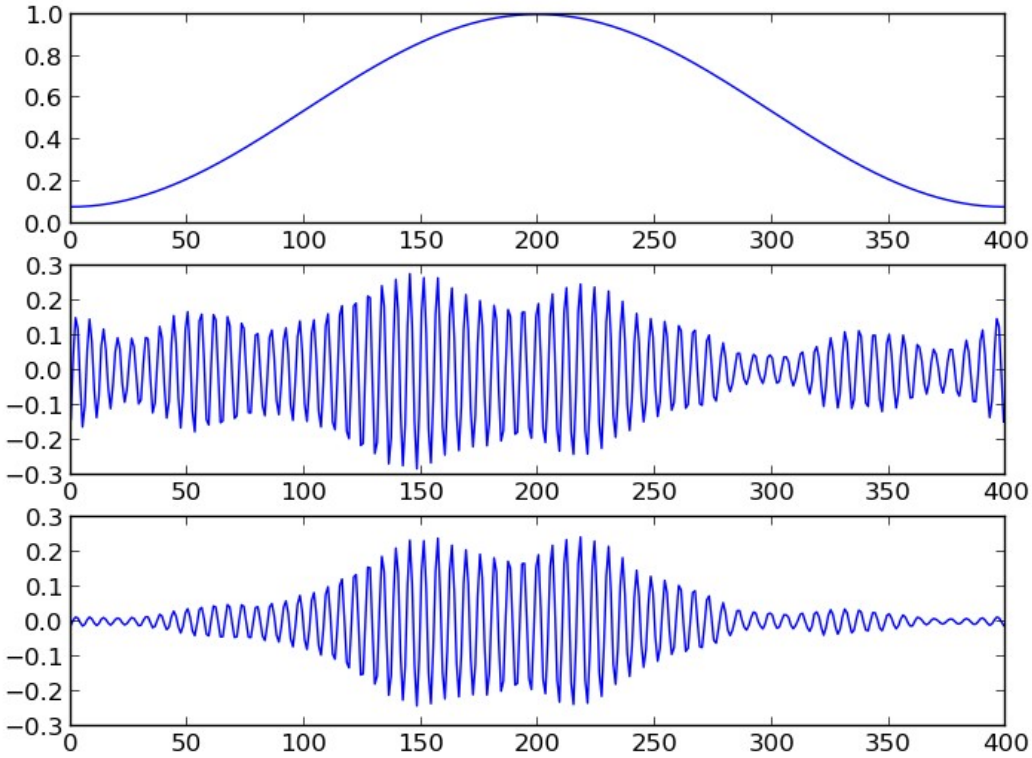


Ilustración 8: *Gráfica de la ventana de Hamming y su aplicación ventana audio.*

4.7.5. Transformada Discreta de Fourier

La transformada discreta de Fourier es la operación que permite calcular los componentes en el dominio de la frecuencia. Desafortunadamente, la transformada da componentes en el dominio de los números imaginarios, que son los que se utilizan para representar información sobre la amplitud de las funciones periódicas por grupos de frecuencias. Sin embargo en lo que hay interés es en la energía en esas frecuencias, por lo que se eleva al cuadrado la información real e imaginaría para extraer la magnitud de la energía. Con esta información, se podrá sacar el periodograma para esa ventana de la siguiente forma:

Primero se calcula la transformada de Fourier rápida para números reales y luego se eleva al cuadrado tanto la parte real como imaginaria y se normaliza por el tamaño de la ventana. Al realizar la graficación de los coeficientes obtenidos de las energías de la transformada de Fourier resulta una gráfica como la de la ilustración 9.

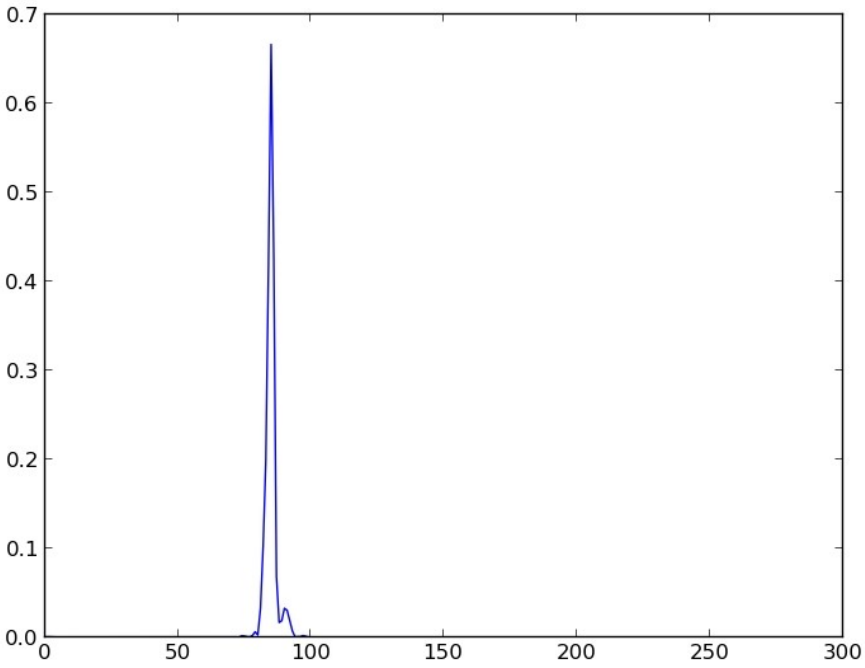


Ilustración 9: *Periodograma de una señal de voz.*

La ventana particular de la ilustración 8 tiene componentes alrededor de los recipientes número 85. Además, el tamaño de las frecuencias es menor a 512, de hecho es la mitad, esto porque los valores son simétricos y representan frecuencias negativas, por lo que es posible obviar esos valores. Ahora si se calcula cada uno de los periodogramas para todas las ventanas en nuestra señal y se gráfica con respecto al tiempo se obtiene un resultado semejante a la ilustración 10.

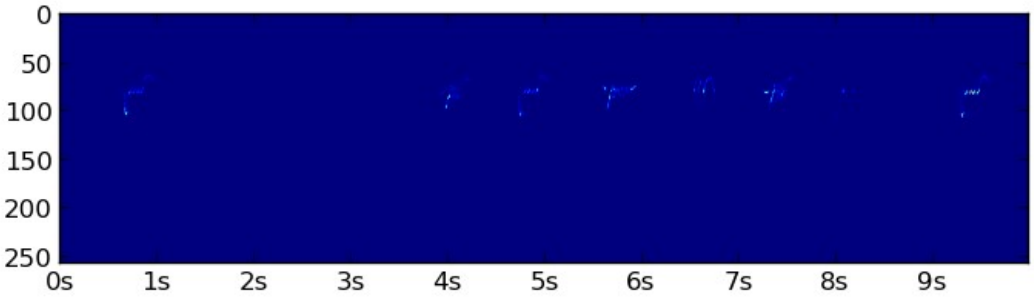


Ilustración 10: *Espectro de periodogramas.*

Lo que se ve es que en altas frecuencias no hay mucha información, pero sí a bajas. También hay que notar que se puede describir toda la señal con 257,000 valores.

4.7.6. Filtros Mel

Un filtro Mel es un filtro triangular que ayuda a obtener información de una banda de frecuencia. Estas bandas están basadas en la percepción del oído humano, en donde frecuencias bajas son más granulares y con mayor peso y frecuencias altas más amplias pero con menor peso. La ilustración 11 muestra las bandas para 257 recipientes de frecuencias:

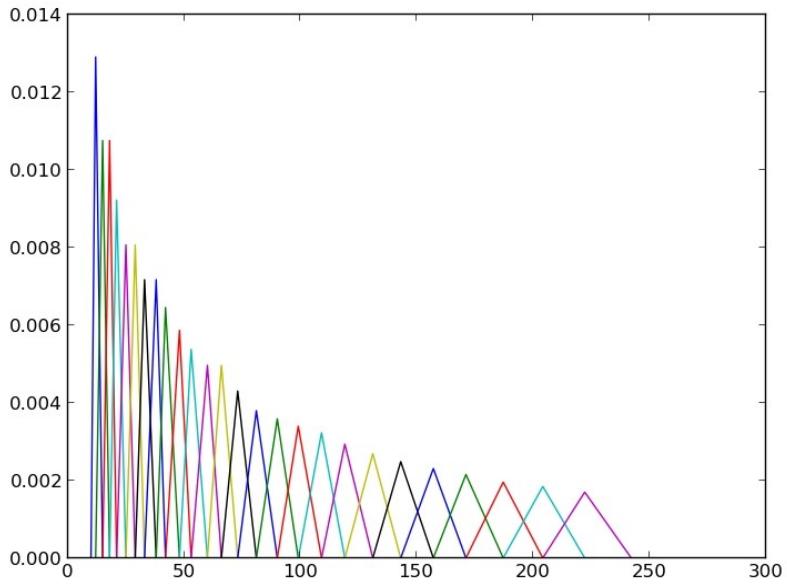


Ilustración 11: *Filtros Mel.*

Como se puede apreciar, un filtro Mel se concentra en las frecuencias bajas y abarca pocas frecuencias, conforme crece las frecuencias comienza abarcar más frecuencias pero la altura del filtro disminuye. Esta forma de los filtros tienen que ver en cómo el ser humano percibe el volumen en frecuencias altas contra el volumen en frecuencias bajas.

4.7.7. Transformada coseno discreta

Es la escala *log* de la energía de los coeficientes Mel lo que convierte a esta señal en un coeficiente

cepstral. Por eso, el siguiente paso es aplicar una transformada de coseno a los coeficientes Mel. La intuición de este paso es desde el punto de vista de procesamiento de señales, que va a permitir comprimir dicha señal en elementos más informativos. De hecho, de años de experiencia en el campo de reconocimiento de voz, de esta transformada coseno se tiende a conservar del segundo al treceavo valor, el resto es ignorado. A continuación, en la ilustración 12, se representa el proceso de las transformadas y los filtros.

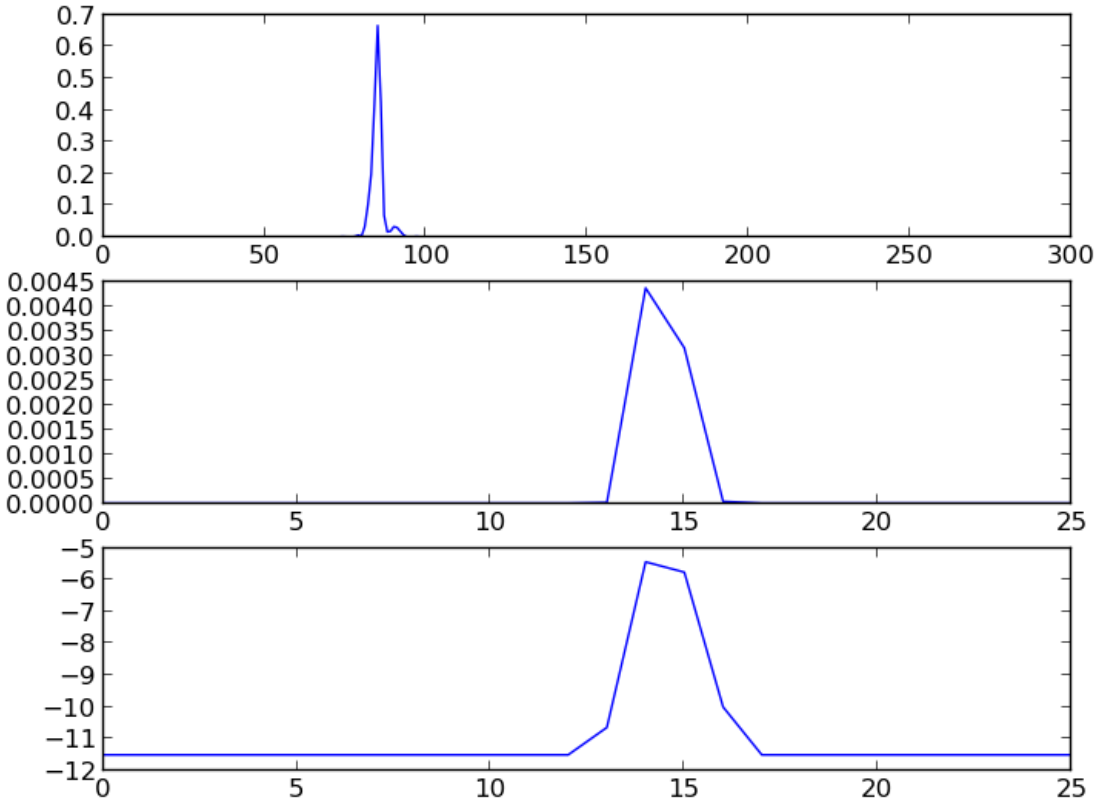


Ilustración 12: *Energías de la DFT, filtrado Mel y DCT de la señal.*

4.8. Alfabeto Fonético Internacional

El Alfabeto Fonético Internacional (AFI) es un sistema de notación fonética creado por lingüistas. Su propósito es otorgar en forma regularizada, precisa y única la representación de los sonidos de cualquier lenguaje oral (International Phonetic Association, 1999), y es usado por lingüistas, logopedas y terapeutas, maestros de lengua extranjera, lexicógrafos y traductores (Daniels & Bright, 1996). En su forma básica (en 2005) tiene aproximadamente 107 símbolos básicos y 55 modificadores (International Phonetic Association, 1999).

Los símbolos del Alfabeto Fonético Internacional están divididos en tres categorías: letras (que indican sonidos «básicos»), diacríticos (que especifican esos sonidos) y suprasegmentales (que indican cualidades tales como velocidad, tono y acentuación). Estas categorías están divididas en secciones menores: las letras están divididas en vocales y consonantes (International Phonetic Association, 1999), y los diacríticos y suprasegmentales están divididos según si indican articulación, fonación, tono, entonación o acentuación (International Phonetic Association, 1999).

Aunque el AFI fue creado para representar solo aquellas cualidades del habla que son relevantes para el idioma en sí (como la posición de la lengua, modo de articulación, y la separación y acentuación de palabras y sílabas) (International Phonetic Association, 1999), un conjunto extendido de símbolos llamados AFI Extendido (Extended IPA en inglés) ha sido creado por fonólogos para marcar cualidades del habla que no tienen un efecto directo en el significado (como el crujido de dientes, ceceo (sigmatismo), y sonidos efectuados por personas con paladar hendido o labio leporino) (Daniels & Bright, 1996).

4.8.1. Descripción

El principio general del AFI es otorgar un símbolo por cada sonido. Esto significa que el AFI no usa combinaciones de letras a menos que el sonido representado pueda ser visto como una secuencia de dos o más sonidos. El AFI usualmente tampoco tiene letras separadas para dos sonidos si ningún idioma conocido distingue entre ellos, y no usa letras que representen múltiples sonidos, en el modo en que «x» representa el conjunto de consonantes [ks] en español. Además, en el AFI ninguna letra tiene valores que dependan del contexto, como la «c» en la mayoría de los idiomas europeos.

Los símbolos del AFI son 107 letras para consonantes y vocales (International Phonetic Association, 1999), 31 diacríticos que especifican esos sonidos, y 19 suprasegmentales, que indican cualidades tales como duración, tono, acento y entonación.

4.8.2. Símbolos y sonidos

El Alfabeto Fonético Internacional ha sido basado deliberadamente en las letras del alfabeto latino, usando tan pocas formas no latinas como sea posible (International Phonetic Association, 1999). La Asociación creó el AFI para que los sonidos de la mayoría de las consonantes tomadas del alfabeto latino correspondieran a «uso internacional» (International Phonetic Association, 1999). Estas consonantes son [b], [d], [f], [ɡ], [k], [l], [m], [n], [p], [s], [t], [v], y [z]. Las otras consonantes del alfabeto latino, [c], [h], [j], [q], [r], [w], [x], y [y], corresponden a los sonidos que representan en otros idiomas:

AFI	pronunciado como en
[c]	kinyarwanda, IAST transliteración del sánscrito, irlandés (en algunos contextos)
[h]	La mayoría de las lenguas germánicas
[j]	La mayoría de las lenguas germánicas y eslavas. Como la y del francés yeux o del inglés yes, y un sonido un poco más fuerte que la i del español viuda.
[q]	quechua de Cuzco-Collao, aimara; inuktitut; transliteración del árabe
[r]	Lenguas eslavas, la mayoría de las lenguas romances, como en español rr en Perro.
[w]	inglés. Como la u en el español huelga
[x]	«x» rusa en el alfabeto cirílico, como j en español.
[y]	francés, alemán, holandés, finés, anglosajón y las lenguas escandinavas; griego antiguo «Y» (ípsilon, upsilon), como en la u francesa y en la mayoría de los casos como la ü alemana.

***Tabla 2**. Pronunciación de acuerdo a los sonidos en otros idiomas (Colaboradores de Wikipedia, 2014a).*

Las vocales del alfabeto latino ([a], [e], [i], [o], [u]) corresponden a las vocales del español.

Los símbolos derivados del alfabeto griego incluyen [β], [ɣ], [ɛ], [θ], [ɸ], y [χ]. De éstas, las únicas que cercanamente corresponden a las letras griegas de las que se derivan son [ɣ] y [θ]. Aunque [β], [ɛ], [ɸ], y [χ] indiquen sonidos similares a beta, épsilon, fi (phi), y ji (chi), no corresponden exactamente. La letra [ʊ], aunque visualmente similar a la vocal griega «υ», ípsilon (upsilon), es realmente una consonante.

Los valores fónicos de las modificaciones de los grafemas de los caracteres latinos pueden inferirse fácilmente de las letras originales (International Phonetic Association, 1999). Por ejemplo, las letras con un gancho girado a la derecha en la parte inferior representan consonantes retroflejas; las mayúsculas pequeñas generalmente notan consonantes uvulares. Aparte del hecho de que ciertas clases de modificaciones en la forma de las letras correspondan a ciertos tipos de modificaciones del sonido que representan, no hay manera de deducir el valor fónico que representa un símbolo solamente por la forma de ese símbolo (por contraste con lo que sucede en *visible speech* [el sistema de caracteres inventado por el profesor Alexander Melville Bell para representar todos los sonidos que pueden ser pronunciados por los órganos del habla]).

Además de las letras mismas hay variedad de símbolos secundarios que se pueden usar en transcripción. Se pueden combinar diacríticos con las letras del AFI para transcribir valores fonéticos de articulaciones secundarias. Hay también símbolos especiales para rasgos suprasegmentales, tales como acentuación y tono.

4.8.3. Letras

El Alfabeto Fonético Internacional divide sus símbolos de letra en tres categorías: consonantes infraglotales o egresivas (pulmónicas), consonantes supraglotales o ingresivas (no pulmónicas), y vocales (International Phonetic Association, 1999).

Consonantes infraglotales o egresivas (pulmónicas)

Las consonantes egresivas son aquellas que se articulan exhalando aire desde los pulmones. Casi todas las consonantes se encuentran en esta categoría, ordenadas en la siguiente tabla de manera que las columnas indican el punto de articulación, y las filas el modo de articulación. Las consonantes a la izquierda representan sonidos sordos y las consonantes a la derecha, un sonido sonoro.

CONSONATES INFRAGLOTALES	bilabial		labio-dental		dental		alveolar		post-alveolar		retrofleja		palatal		velar		uvular		faríngea		glotal	
	srd	snr	srd	snr	srd	snr	srd	snr	srd	snr	srd	snr	srd	snr	srd	snr	srd	snr	srd	snr	srd	snr
oclusiva	p	b					t	d			ʈ	ɖ	c	ɟ	k	g	q	ɢ			ʔ	
nasal		m		ɱ				n				ɳ		ɲ		ŋ		ɴ				
vibrante múltiple		ʙ						r										ʀ				
vibrante simple								ɾ				ɽ										
fricativa	ɸ	β	f	v	θ	ð	s	z	ʃ	ʒ	ʂ	ʐ	ç	ʝ	x	ɣ	χ	ʁ	ħ	ʕ	h	ɦ
fricativa lateral							ɬ	ɮ														
aproximante				ʋ				ɹ				ɻ		j		ɰ						
aproximante lateral								l				ɭ		ʎ		ʟ						

***Ilustración 13**. Consonantes infraglotales o egresivas (pulmónicas) (Colaboradores de Wikipedia, 2014a).*

Coarticulación

Las consonantes coarticuladas son sonidos en los que dos consonantes individuales son pronunciadas al mismo tiempo.

ʍ	Aproximante velar labializada sorda
w	Aproximante velar labializada sonora
ɥ	Aproximante palatal labializada sonora
ɕ	Fricativa postalveolar palatalizada (alveolo-palatal) sorda
ʑ	Fricativa postalveolar palatalizada (alveolo-palatal) sonora
ɧ	Fricativa "palatal-velar" sorda

Tabla 3*. Sonidos de las constantes coarticuladas (Colaboradores de Wikipedia, 2014a).*

Nota: [ɧ] es descrito como un «[ʃ] y [x] simultáneo» (Ladefoged & Maddieson, 1996). Sin embargo, este análisis está discutido.

Existen otras consonantes coarticuladas usualmente denotadas mediante diacríticos.

Africadas y oclusivas de doble articulación

Las africadas y las oclusivas de doble articulación se representan por dos símbolos unidos por una barra de ligadura, abajo o arriba de los símbolos. Las seis africadas más comunes son representadas alternativamente por ligaduras, aunque no representa el uso oficial del AFI, debido al gran número de ligaduras que se requeriría para representar todas las africadas de esta forma. Una tercera forma de transcripción de las africadas que se ve en ocasiones es el uso de caracteres volados, verbigracia ts para t͡s, siguiendo el modelo de kx ~ k͡x. Los símbolos para las oclusivas palatales, [c, ɟ] se usan a menudo por conveniencia para [t͡ʃ, d͡ʒ] o africadas similares, incluso en las publicaciones oficiales en AFI, por lo que deben ser interpretados con mucho cuidado.

Barra de ligadura	Ligadura	Descripción
t͡s	ʦ	africada alveolar sorda
d͡z	ʣ	africada alveolar sonora
t͡ʃ	ʧ	africada postalveolar sorda
d͡ʒ	ʤ	africada postalveolar sonora
t͡ɕ	ʨ	africada palato-alveolar sorda
d͡ʑ	ʥ	africada palato-alveolar sonora
t͡ɬ	–	africada lateral-alveolar sorda
k͡p	–	oclusiva velo-labial sorda
g͡b	–	oclusiva velo-labial sonora
ŋ͡m	–	oclusiva velo-labial nasal sorda

Tabla 4*. Africadas y oclusivas de doble articulación (Colaboradores de Wikipedia, 2014a).*

Consonantes supraglotales o ingresivas (no pulmónicas)

Clics	Implosivas	Eyectivas
ʘ Bilabial	ɓ Bilabial	ʼ *Por ejemplo:*
ǀ Laminal alveolar ("dental")	ɗ Alveolar	pʼ Bilabial
ǃ Apical (post-) alveolar ("retrofleja")	ʄ Palatal	tʼ Alveolar
ǂ Laminal postalveolar ("palatal")	ɠ Velar	kʼ Velar
ǁ Lateral coronal ("lateral")	ʛ Uvular	sʼ Fricativa alveolar

Tabla 5*. Consonantes supraglotales o ingresivas (no pulmónicas) (Colaboradores de Wikipedia, 2014a).*

Vocales

El AFI define una vocal como un sonido que ocurre en el núcleo de una sílaba (International Phonetic Association, 1999). En la ilustración 15 se encuentran representadas en una tabla las vocales con signo propio en el AFI. El AFI ubica las vocales en un gráfico bidimensional según la posición de la lengua. Las dos dimensiones según este gráfico son la anterioridad (vocal anterior/central/posterior) y la altura o abertura (vocal cerrada/semicerrada/semiabierta/abierta, etc.). Estas dos dimensiones se corresponden respectivamente con el segundo formante y el primer formante, encontrados en el espectrograma de

dichos sonidos:

- El eje vertical de la tabla está determinado por la altura de la vocal. Las vocales pronunciadas con la lengua baja están en la base, y las vocales pronunciadas con la lengua alzada están en la cima. Por ejemplo, [ɑ] está en la base porque la lengua está baja en esta posición. Sin embargo, [i] está en la cima porque el sonido es pronunciado con la lengua alzada hacia el techo de la boca.
- De manera similar, el eje horizontal está determinado por el fondo de la vocal. Las vocales con la lengua movida hacia el frente de la boca (como [ɛ]) están a la izquierda de la tabla, mientras que aquellas en que se mueve hacia atrás (como [ʌ]) son colocadas a la derecha de la tabla.

Donde las vocales están en pares, la derecha representa una vocal redondeada mientras que la izquierda es la equivalente no redondeada. Este redondeamiento se asocia al tercer formante (que es menos prominente en la mayoría de lenguas).

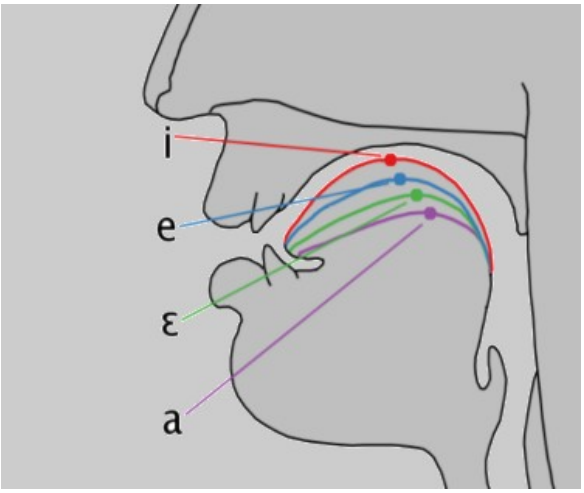


***Ilustración 14**. Posiciones de la lengua de vocales frontales cardinales con el punto más alto indicado. La posición del punto más alto es usado para determinar altura y fondo (D. Jones, 1972).*

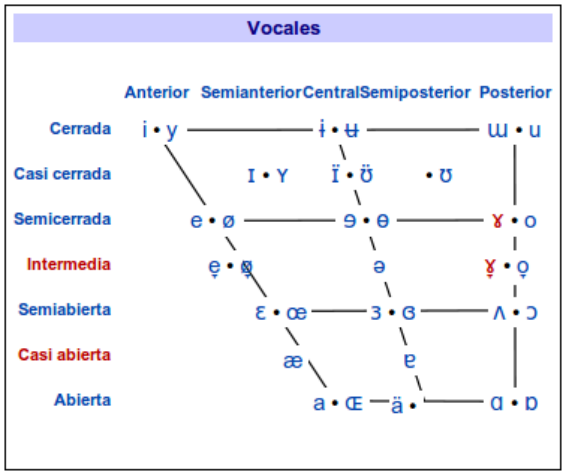


***Ilustración 15**. Vocales con signo propio en el AFI (Colaboradores de Wikipedia, 2014a).*

Diacríticos

Los diacríticos son pequeñas marcas que se colocan alrededor de la letra AFI para mostrar una cierta alteración o descripción más específica en la pronunciación de la letra (International Phonetic Association, 1999). Sub-diacríticos (marcas normalmente puestas bajo una letra o símbolo) pueden ponerse arriba de un símbolo con descendiente, v.g. n̊ŋ̊ (International Phonetic Association, 1999).

La i sin punto, <ɪ>, es usada cuando el punto interferiría con el diacrítico. Otros símbolos AFI pueden aparecer como diacríticos para representar detalle fonético: tˢ (salida fricativa), bʱ (voz murmurada), ˀa (comienzo glotal), ᵊ (epenthetic schwa), oᵘ (diptonguización). Diacríticos más avanzados fueron desarrollados en el AFI extendido para codificación de pronunciación más específica.

Diacríticos de silabicidad			
ɻ̩ n̩	Silábica	e̯ o̯	No silábica
Diacríticos de realización consonántica			
tʰ dʰ	Aspirada[n. 1]	d̚	Sin soltamiento perceptible
dⁿ	Soltamiento nasal	dˡ	Soltamiento lateral
Diacríticos de fonación			
n̥ d̥	Sorda	s̬ t̬	Sonora
b̤ a̤	Voz murmurada[n. 2]	b̰ a̰	Voz rechinada
Diacríticos de articulación			
t̪ d̪	Dental	t̼ d̼	Linguolabial
t̺ d̺	Apical	t̻ d̻	Laminal
u̟ t̟	Avanzada	i̠ t̠	Retraída
ë ä	Centralizada	e̽ ʊ̽	Medio centralizada
e̝ ɻ̝	Levantada (ɻ̝ = fricativa alveolar sonora no silbante)		
e̞ β̞	Bajada (β̞ = aproximante bilabial)		
Diacríticos de co-articulación			
ɔ̹ x̹	Más redondeada	ɔ̜ x̜ʷ	Menos redondeada
tʷ dʷ	Labializada	tʲ dʲ	Palatalizada
tˠ dˠ	Velarizada	tˤ dˤ	Faringealizada
ɬ̴ ɮ̴	Velarizada o faringealizada		
e̙ o̙	Raíz de la lengua avanzada	e̘ o̘	Raíz de la lengua retraída
ẽ z̃	Nasalizada	ɚ˞ ɝ˞	Rotización

***Ilustración 16**. Tabla de diacriticos (Colaboradores de Wikipedia, 2014a).*

Consideraciones:

- Con consonantes sonoras aspiradas, la aspiración también es sonora. Muchos lingüistas prefieren uno de los diacríticos dedicados a la voz murmurada.
- Algunos lingüistas restringen este diacríticos a sonorantes, y transcriben las obstruyentes como bʱ.

El estado del glotis puede ser bien transcrito con diacríticos. Una serie de oclusivas alveolares desde una fonación de glotis abierta a una cerrada son:

[t]	sorda	[t̪]	voz murmurada
[t̤]	voz floja	[d]	voz modal
[t̰]	voz dura	[d̰]	voz rechinada
[ʔ͡t]	cierre glotal		

***Tabla 6**. Serie de oclusivas alveolares desde una fonación de glotis abierta a una cerrada (Colaboradores de Wikipedia, 2014a).*

Suprasegmentales

Duración, acento, y ritmo			
ˈ	Acento primario (antes de sílaba acentuada)	ˌ	Acento secundario (antes de sílaba acentuada)
ː	Consonante geminada o vocal larga	ˑ	Semilarga
˘	Extra breve	.	Ruptura silábica
‿	Ausencia de ruptura		
Entonación			
|	Ruptura menor	‖	Ruptura mayor
↗	Subida global	↘	Bajada global
Tonos			
e̋ ő ˥	Extra alto	ê	Caída
é ó ˦	Alto	ě	Subida
ē ō ˧	Medio		
è ò ˨	Bajo	↓e (ꜜe)	Descendente
ȅ ȍ ˩	Extra bajo	↑e (ꜛe)	Ascendente

***Ilustración 17**. Suprasegmentales (Colaboradores de Wikipedia, 2014a).*

4.9. Naïve Bayes

Las redes bayesianas (Malagón Luque, 2003), 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 Naïve 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.
- Es sencillo asociar un porcentaje de confianza a las predicciones, y combinar predicciones en base a su confianza
- Una nueva instancia es clasificada como función de la predicción de múltiples hipótesis, ponderadas por sus probabilidades.
- Incluso en algunos casos en los que el uso de estos métodos se ha mostrado imposible, pueden darnos una aproximación de la solución óptima.

El aprendizaje se puede ver como el proceso de encontrar la hipótesis más probable, dado un conjunto de ejemplos de entrenamiento D y un conocimiento a priori sobre la probabilidad de cada hipótesis (Prieto Izquierdo & Casillas Díaz, 2011).

4.10. Teorema de Bayes

Una forma de describir el teorema de Bayes es la propuesta por (Prieto Izquierdo & Casillas Díaz, 2011):

$$P(h|D)=\frac{P(D|h)P(h)}{P(D)}$$

Donde:

- $P(h)$ es la probabilidad a priori de la hipótesis *h*.
- $P(D)$ es la probabilidad de observar el conjunto de entrenamiento *D*.
- $P(D|h)$ es la probabilidad de observar el conjunto de entrenamiento *D* en un universo donde se verifica la hipótesis *h*.
- $P(h|D)$ es la probabilidad a *posteriori* de *h*, cuando se ha observado el conjunto de entrenamiento *D*.

4.11. Clasificación

De acuerdo con (Prieto Izquierdo & Casillas Díaz, 2011), el procedimiento adecuado para realizar una clasificación mediante Naïve Bayes es el siguiente:

- Cada ejemplo x se describe con la conjunción de los valores de sus atributos: $\langle a_{1,} a_{2,}\ldots, a_n\rangle$.
- La función objetivo $f(x)$ puede tomar cualquier valor de un conjunto finito V .
- La clasificación viene dada por el valor de máxima probabilidad a *posteriori*: v_{MAP} .

  $$v_{MAP}=\underset{v_j\in V}{arg\,max}\, P(v_j|a_1,a_2,\ldots,a_n)$$

  $$v_{MAP}=\underset{v_j\in V}{arg\,max}\,\frac{P(a_1,a_2,\ldots,a_n|v_j)P(v_j)}{P(a_1,a_{2,}\ldots,a_n)}$$

  $$v_{MAP}=\underset{v_j\in V}{arg\,max}\, P(a_1,a_2,\ldots,a_n|v_j)P(v_j)$$

- Los términos se han de estimar basándose en los ejemplos de entrenamiento.
  - $P(v_j)$ contando la frecuencia con la que ocurre cada valor v_j .
  - Hay demasiados términos de la forma $P(a_1,a_2,\ldots,a_n|v_j)$. Harían falta muchísimos ejemplos de entrenamiento para obtener una buena estimación.
- La suposición del clasificador *naive* es que los atributos son independientes entre sí con respecto al concepto objetivo y, por lo tanto: $P(a_1,a_2,\ldots,a_n|v_j)=\prod_i P(a_i|v_j)$

- La aproximación del clasificador bayesiano *naive* es: $v_{nb} = \underset{v_j \in V}{arg\,max}\, P(v_j) = \prod_i P(a_i|v_j)$
- Las probabilidades $P(a_i|v_j)$ resultan mucho más fácil de estimar que las $P(a_1, a_2, \ldots, a_n)$.

4.12. Algoritmo

```
Aprendizaje_Bayesiano_Naive(ejemplos)
    Para cada posible valor del resultado vⱼ
        Obtener estimación P'(vⱼ) de la probabilidad P(vⱼ)
            Para cada valor aᵢ de cada atributo a
                Obtener una estimación P'(aᵢ|vⱼ) de la
    probabilidad P'(aᵢ|vⱼ)

Clasificar_instancia(x)
    devolver  v_nb = arg max_{v_j ∈ V} P(v_j) = ∏_i P(a_i|v_j)
```

Capítulo 5. Experimentación y resultados

A lo largo de este capítulo se explicarán las circunstancias, procedimientos y resultados que hicieron del Identificador Automático del Lenguaje aquí desarrollado, y que tiene como nombre «Kibo», una realidad. Lo que a continuación se relata, es entonces, lo que fue necesario para realizar uno de los primeros LID que utiliza un RAH —el cual realiza las transcripciones de datos de voz al Alfabeto Fonético Internacional— para realizar las identificaciones de datos de voz tanto para el idioma téenek como para el español.

5.1. Zonas, localidades y ubicaciones

Una de las más grandes ventajas conseguidas con el enfoque con el que se desarrolló el *corpus* fue que no se requirió desplazarse a las zona donde el idioma en cuestión es hablado de forma nativa. Por lo mismo, toda la investigación fue desarrollada en la zona metropolitana que componen las ciudades de Tampico, Madero y Altamira.

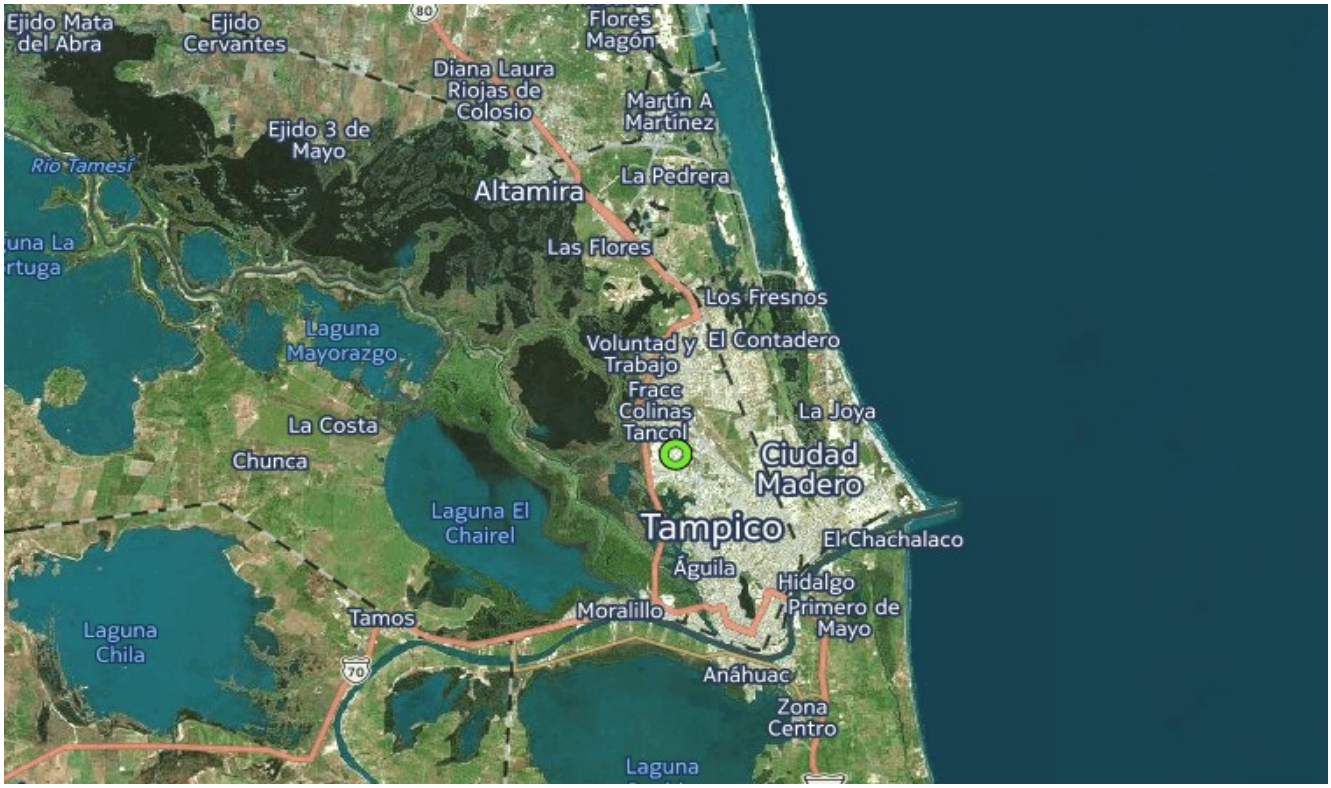


***Ilustración 18**: Ubicación geográfica donde la investigación fue realizada. Mapa creado mediante la tecnología de HERE (©2015 HERE)*

5.2. Composición del *corpus*

El *corpus* de Kibo, es sin duda, uno de los más interesantes componentes de este LID debido al diseño alternativo propuesto para la arquitectura del mismo (léase el capítulo 1.3). Así pues, es importante hacer notar una de sus mejores características, la cual es la de permitir su reutilización independientemente del idioma que se esté tratando de identificar, puesto que no está atado a ningún idioma en particular sino que las transcripciones se hacen a un lenguaje fonético estandarizado universal. Esto lo hace extremadamente flexible aunque también hace que requiera una robustez mucho mayor de la que se contempló en un principio. Lamentablemente, esta robustez no fue posible conseguirla en el transcurso de esta investigación; mucho tiene que ver que este sea un primer acercamiento a un problema sin apenas precedentes.

Este corpus está compuesto por un total de 4585 conjuntos de atributos procesados a partir de 588 muestras de audio, las cuales fueron obtenidas mediante el tratamiento de 130 grabaciones realizadas de la siguiente manera:

1. Se le pidió a un conjunto de 20 personas hablantes del idioma español —10 mujeres, 10 hombres— que pronunciarán las palabras listadas a continuación:
    1. beca (*beka*)
    2. bicho (*biʧo*)
    3. butaca (*butaka*)
    4. chacal (ʧ*akal*)
    5. chequeo (ʧ*ekeo*)
2. Posteriormente, se aprovecharon las grabaciones que realizó (Alviso Vargas, 2014) (que obtuvo de 6 personas hablantes de la lengua téenek —3 mujeres, 3 hombres—) y que contienen el siguiente conjunto de palabras:
    1. akak (*akak*)
    2. bo' (*boo*)
    3. buuk (buuk)
    4. chab (ʧ*aab*)
    5. che' (ʧ*ee*)
3. Cada elemento de este conjunto de grabaciones, se segmentó de tal forma que cada fonema quedó registrado en un archivo de audio individual; dando como resultado las 588 grabaciones mencionadas anteriormente.

4. Por último, las 588 grabaciones fueron procesadas para obtener sus MFCC; construyendo con éstos el conjunto de entrenamiento, el cual es usando en procesos posteriores por el LID.

De todo esto, hay algo que sobresale de forma insoslayable: las grabaciones no necesitan ser de palabras pronunciadas en el mismo idioma. Esto es posible gracias al diseño del mismo que recaba cada fonema que compone cada palabra y lo etiqueta haciendo uso de un alfabeto fonético estandarizado.

5.3. Grabación

Para la etapa de grabación de cada muestra se consideraron los lineamentos propuestos por (Hernández Zepeda, 2013) en su estrategia de muestreo:

- Las condiciones de ruido deben ser lo más cercanas a nulas. Es decir, intentar en la medida de lo que sea posible, que en la grabación de los archivos de audio sólo hubiesen sonidos útiles.
- Frecuencia de muestreo de 44100 Hz.
- Formato de muestra PCM.
- 16 bits como medida de resolución/cuantificación de bits.
- Formato de archivo WAV (sin compresión de datos).
- Sonido Monoaural.

Para realizar la grabación se utilizó el micrófono Logitech H390 con las siguientes características técnicas:

- USB compatible (1.1 y 2.0).
- Respuesta en frecuencia: 100 Hz - 10 KHz.
- Sensibilidad de entrada: -62 dBV/µbar, -42 dBV/Pa +/- 3 dB.
- Tecnología noise-canceling integrada.

5.4. Eliminación de ruido

Aunque se mencionó que las grabaciones debían haberse realizado intentando evitar sonidos no deseados, lo cierto que en la práctica esto resulta algo muy difícil. Una forma sencilla de solucionar esto es usar un algoritmo de eliminación de ruido. Para fines prácticos, se usó el algoritmo proporcionado por el editor de audio Audacity (Audacity development team, Mazzoni, & Dannenberg, 2015); éste incorpora un método de reducción de ruido por identificación de sus componentes armónicos. En la documentación proporcionada por el mismo software se describen adecuadamente los pasos para usar tal función.

5.5. Segmentación

Una vez eliminado el ruido de las grabaciones, éstas debían ser fragmentadas para conseguir que cada fonema estuviese almacenado en un archivo de audio individual. Para esta tarea, se requirió hacer uso del software para análisis científico del habla llamado Praat (Boersma, Weenink, & University of Amsterdam, 2014). Mediante el mismo, fue posible realizar la segmentación de una manera más cómoda que con Audacity, puesto que al poderse etiquetar fragmentos de audio y utilizar scripts para guardar cada uno en un archivo es algo en lo que Praat está especializado. Además, Praat cuenta con algunos algoritmos (utterance segmentation y *phonetizer*) que si bien no son exactos, aproximan lo suficientemente la alineación de cada segmento de audio como para que sea más sencillo ajustar de forma manual cada uno.

El algoritmo utterance segmentation calcula una heurística basada en la duración de la señal de audio y la longitud de la transcripción para estimar la duración del enunciado, incluidas las pausas.

El algoritmo *phonetizer* hace uso del modulo grafema-fonema de voz a texto llamado eLite desarrollado en la Faculté Polytechnique de Mons, Belgium para hacer un análisis lingüístico de la transcripción ortográfica y así producir una transcripción fonética basada en reglas fonéticas de diccionario y de pronunciación.

Un ejemplo de la segmentación mediante el software Praat se puede apreciar en la ilustración 19.

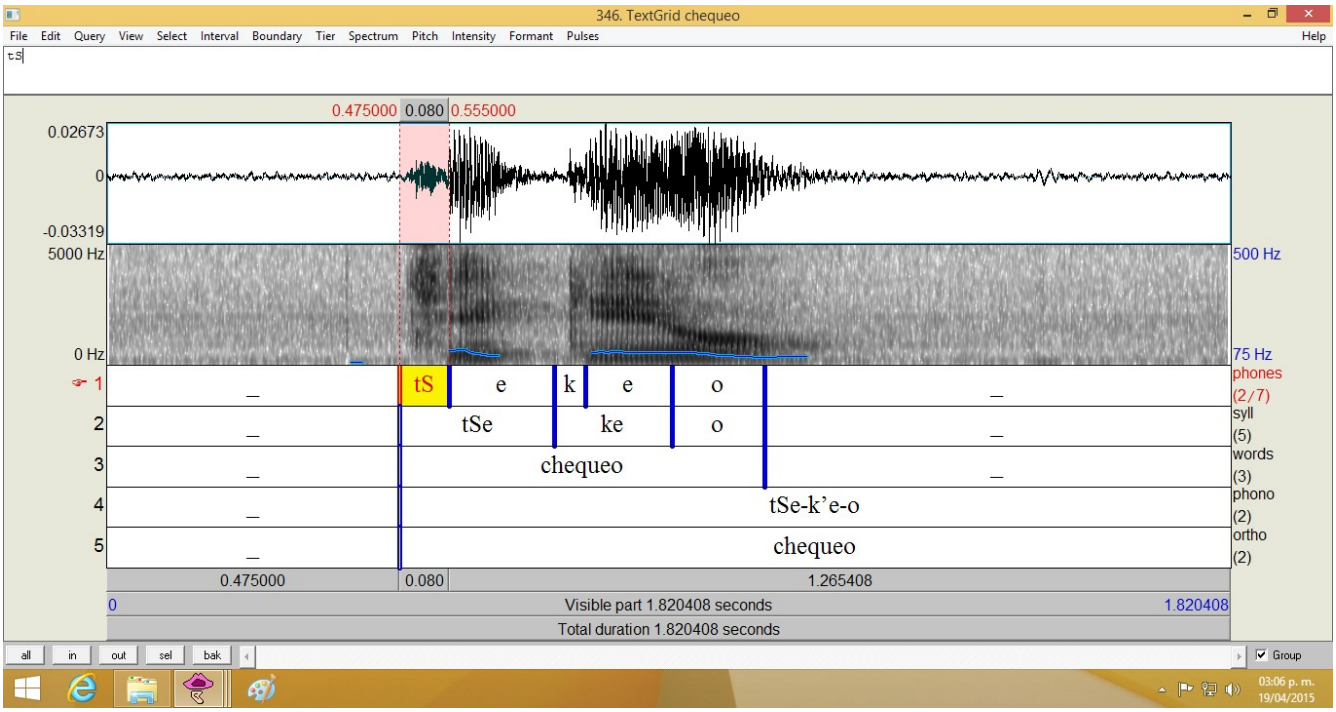


***Ilustración 19**: Segmentación y etiquetado temporal de un archivo de audio en Praat.*

Una vez definido que el segmento de *x* segundo a *y* segundo correspondía a determinado fonema, este era guardado en un archivo de audio independiente con la convención de estilos *fonema-número de persona-conteo de fonema.wav*, tal como es posible observar en la ilustración 20.

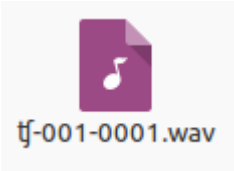


***Ilustración 20**: Archivo de audio correspondiente al fonema ʧ correctamente nombrado usando la convención de estilos propuesta.*

Lo anterior se trabaja de esa forma porque, aunque se cuenta con algoritmos de segmentación y alineación automática, éstos sólo pueden usarse para un idioma en particular. Hasta el momento no se conoce un algoritmo de segmentación y alineación automática general, o lo que aquí se determinó llamar: *Language Independent Automatic Speech Segmentation and Labeling Problem*.

5.6. Interpretación de fonemas a MFCC

El proceso de obtención de los MFCC, *grosso modo*, es el siguiente:

1. Dividir la señal en ventanas pequeñas. De preferencia usar una ventana de Hamming.
2. Para cada ventana calcular el periodograma que representa la energía por cada valor de frecuencia.
3. Filtrar las frecuencias y energías del periodograma usando un banco de filtros Mel.
4. Calcular el logaritmo de la energía por cada uno de los filtros.
5. Calcular la transformada coseno discreta de las energías.
6. Conservar los valores 2 a 13 de esta transformada, descartar el resto.

Siguiendo los pasos anteriormente descritos, se obtuvieron los MFCC correspondientes a cada fonema. Una vez obtenidos éstos, se agregaron a un archivo final estructurado de forma conveniente para el sistema; el llamado conjunto de entrenamiento. Dado que este archivo es usado por el entorno para análisis del conocimiento de la Universidad de Waikato (Hall et al., 2009), la estructura del mismo es la de un archivo .ARFF, propio de WEKA.

Para dar más luz a esta descripción, se mostrará una parte de este archivo (cabecera y dos líneas de atributos con sus etiquetas) a continuación:

```
% 1. Title: Mel-frequency cepstral coefficients obtained from phoneme speech file(s).
%
% 2. Sources:
%    (a) Creator: Manuel Alejandro Jiménez Quintero
%    (b) Date: May, 2015
@relation training_set
@attribute mfcc_1 numeric
@attribute mfcc_2 numeric
@attribute mfcc_3 numeric
@attribute mfcc_4 numeric
@attribute mfcc_5 numeric
@attribute mfcc_6 numeric
@attribute mfcc_7 numeric
@attribute mfcc_8 numeric
@attribute mfcc_9 numeric
@attribute mfcc_10 numeric
@attribute mfcc_11 numeric
```

```
@attribute mfcc_12 numeric

@attribute classes {a,ð,b,l,u,p,k,i,o,e,ʧ,m,t,n,r}


@data

-5.66393801107,     -18.229756071,     3.2110609716,     8.37452685724,     -18.4459623065,
-2.90902185146,    -4.91464053707,    -4.27261480316,    -1.21914589257,    0.595414433511,
0.0305514854904, 1.97599715347, a

0.773613509776,     -42.3707185425,     18.2358473371,     8.30219329528,     13.0188466428,
14.6709272798,     -27.2793402457,     -25.6946666162,     -3.72877677925,     9.80804871973,
-8.3812906408, -12.0994135271, ʧ
```

Este archivo (junto con las 4583 líneas restantes) está listo para ser usado en el sistema.

5.7. Clasificación

WEKA cuenta en su entorno de trabajo con un gran conjunto de técnicas de clasificación disponibles. El sistema hace uso de un clasificador de éste conjunto, y este a su vez, hace uso del conjunto de entrenamiento creado antes; todo esto con la finalidad de poder realizar una inferencia sobre un caso de prueba. Éste caso de prueba, no es otro mas que un archivo de datos de voz desconocido. Para esta primera aproximación se decidió usar el clasificador Naïve Bayes porque 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, siguiendo el procedimiento probabilístico definido con anterioridad (Zhang, 2004).

Entonces, primero se realiza el entrenamiento del sistema. La forma de realizar éste proceso se describe de forma detallada en la documentación para las funciones CLI publicada por la Universidad de Waikato, aunque en Kibo este proceso —como todos a continuación— es transparente para el usuario final.

Una vez realizado el anterior procedimiento, se procede a evaluar un caso de prueba usando el clasificador antes mencionado. Un caso de prueba es un archivo de datos de voz correspondiente a un fonema en particular. Con el fin de sistematizar lo más posible, un conjunto de casos de prueba correspondientes a una palabra desconocida es almacenado en un archivo .KIB.

La estructura de un archivo .KIB es la siguiente:

```
unknown_0001/
---description.txt
---phoneme/
------phoneme_0001.wav
------phoneme_0002.wav
.
.
.
-----phoneme_n.wav
---word.wav
```

Donde cada uno de estos casos de prueba (archivos de audio) son preparados de la misma forma en la que cada fonema contenido en el archivo de entrenamiento fue realizado. Así pues, de cada archivo de audio se extraen los MFCC y se depositan en un archivo .ARFF, el cual es usado como argumento para el clasificador Naïve Bayes.

Una vez que se obtiene una inferencia por cada fonema desconocido, estos se concatenan en una sola cadena. Por ejemplo, un archivo de audio con la palabra chequeo:

phoneme_0001.wav → phoneme_0001.arff → proceso de clasificación → ʧ

phoneme_0002.wav → phoneme_0002.arff → proceso de clasificación → *e*

phoneme_0003.wav → phoneme_0003.arff → proceso de clasificación → *k*

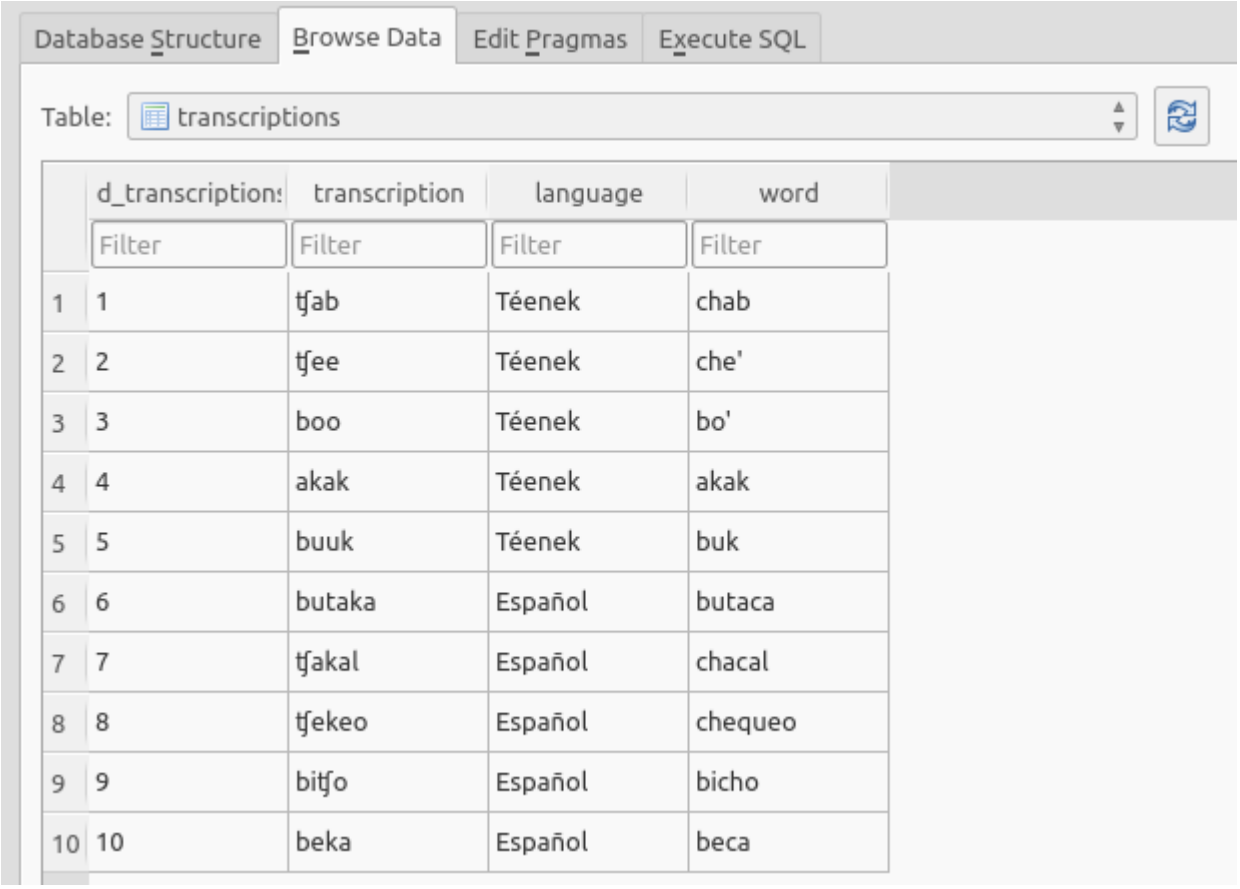
phoneme_0004.wav → phoneme_0004.arff → proceso de clasificación → *e*

phoneme_0005.wav → phoneme_0005.arff → proceso de clasificación → *o*

ʧ + e + k + e + o = ʧekeo

5.8. Proceso de identificación del idioma de la muestra

La parte de la transcripción de los datos de voz de la palabra desconocida al AFI del anterior paso es realmente lo más complicado del proceso. Una vez obtenida la posible cadena que conforma la muestra se procede a buscar una coincidencia en una base de datos preparada para este caso y con la que se relaciona determinadas transcripciones con el idioma del cual procede la palabra. Esta puede observarse en la ilustración 21.



	d_transcription	transcription	language	word
	Filter	Filter	Filter	Filter
1	1	ʧab	Téenek	chab
2	2	ʧee	Téenek	che'
3	3	boo	Téenek	bo'
4	4	akak	Téenek	akak
5	5	buuk	Téenek	buk
6	6	butaka	Español	butaca
7	7	ʧakal	Español	chacal
8	8	ʧekeo	Español	chequeo
9	9	biʧo	Español	bicho
10	10	beka	Español	beca

Ilustración 21*: Base de datos que contiene la relación transcripción, el idioma y la representación de la palabra en el idioma del que procede.*

Las transcripciones al AFI de las palabras en español que se usaron en esta investigación pertenecen a (Butterfield, González, & Breslin, 1995). En cambio, las transcripciones de las palabras en téenek al AFI, fueron conseguidas vía entrevistas con la lingüista (Pison Alcaraz, 2014) y la Licenciada en Educación Preescolar Indígena (Flores Hernández, 2014).

Ahora bien, si el sistema obtuvo la transcripción ʧ*ekeo*, ésta se busca en la base de datos en el campo transcription. Al darse una coincidencia, entonces se puede saber el idioma mediante el campo language; e incluso la palabra original mediante el campo word. Dicho esto, ahora es posible conocer que la muestra desconocida en realidad se trata de la palabra *chequeo* del idioma español.

La base de datos en cuestión debe ser lo suficientemente robusta como para contener la mayor cantidad de transcripciones. Sin embargo y para efectos de demostración, aquí sólo contiene las palabras que han sido objeto de prueba a lo largo de esta investigación.

Quizás se pensaría lo anterior como un posible inconveniente por cuestiones de rendimiento, pero dado que tanto el lenguaje SQL como los sistemas manejadores de base de datos actuales están preparados para soportar colosales cantidades de datos en tiempo real, este proceso sencillamente está bien sustentado tecnológicamente.

Un verdadero inconveniente es que la transcripción inferida debe ser exacta a la transcripción conocida. Lamentablemente, no es posible usar métodos probabilísticos adicionales como los HMM (Modelos Ocultos de Markov) debido a que pudieran tergiversar los resultados a causa de existirán una gran cantidad de transcripciones semejantes conforme se vayan agregando nuevos idiomas.

5.9. Desarrollo y estructura de Kibo

Kibo, como ya se mencionó, es el nombre del sistema LID resultado de esta tesis. Kibo en realidad es la romanización de la palabra japonesa 希望 y en español significa *esperanza*. Kibo es también el nombre del módulo de experimentación japonés, el más grande de la ISS (Estación Espacial Internacional). Es debido a estar razones —personales— por las que se determinó nombrar al sistema así.

En un principio se tenía previsto desarrollar un sistema completamente automatizado, sin embargo, debido a la complejidad del problema y dado que se trata de un primer acercamiento, esto no fue posible. Pero pese a que el proceso de segmentación y etiquetado se realiza de forma manual, todo el resto está sistematizado. Esto logra que, a pesar de que hay que preparar cada archivo .KIB de forma manual (para referencia, léase el capítulo 5.7), una vez obtenido éste sólo hace falta indicarle la ubicación al sistema para que éste intente identificar el idioma de la muestra.

Kibo se estructura de la siguiente forma:

- *Procesador de archivos de audio*: obtiene los MFCC de los archivos de audio, etiqueta los conjuntos de atributos y los estructura en un archivo .ARFF compatible con los estándares de WEKA.
- *Reconocedor automático de los datos de voz*: utilizando las librerías de WEKA mediante una conexión vía socket, se realiza un entrenamiento del sistema y se procede a realizar una inferencia de cada uno de los archivos de audio del paquete .KIB suministrado, para posteriormente, concatenar cada fonema obtenido en una sola palabra.
- *Identificador automático del idioma*: usando la transcripción obtenida por el *reconocedor automático de los datos de voz* consigue identificar el idioma de la muestra de voz al encontrar una coincidencia en la *base de datos de transcripciones* (refiérase al capítulo 5.8).

La estructura de archivos del sistema se puede apreciar en la ilustración 22.

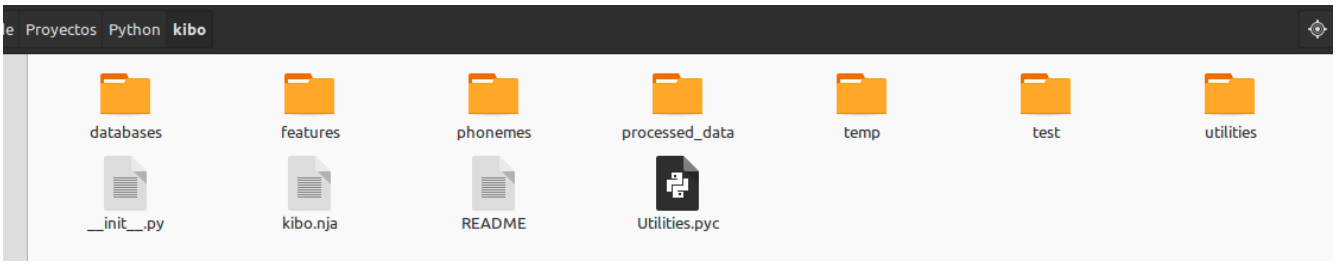


***Ilustración 22**: Estructura de archivos del sistema LID.*

Procedamos a describir cada directorio:

- *Databases*: contiene la base de datos de transcripciones.
- *Features*: contiene las clases (conjunto de variables, estados y métodos apropiados para operar con dichos datos) encargadas de procesar los archivos de audio para obtener los MFCC.
- *Phonemes*: contiene los archivos de los fonemas soportados por el *corpus* y que son el conjunto de entrenamiento del sistema.
- *Processed_data*: en esta carpeta se conservan los archivos .ARFF obtenidos en la iteración actual.
- *Temp*: es un directorio que sirve de repositorio para archivos transitorios.
- *Test*: aquí se colocan los contenedores .KIB de casos de prueba.
- *Utilities*: contiene las clases que sirven de soporte para tareas extras, tales como formateo de cadenas o conexiones.
- *__init__.py*: archivo principal del sistemas, en él se encuentra toda la lógica del sistema.
- *Kibo.nja*: archivo con metadatos del IDE utilizado.
- *README*: archivo con información general sobre el sistema, su funcionamiento y estructura.
- *Utilities.pyc*: archivos compilados auto-generados por el intérprete de Python.

Kibo fue escrito en el lenguaje de programación Python en un computador con las características descritas en la ilustración 23.

***Ilustración 23**: Características del computador en el que fue desarrollado el sistema.*

5.10. Pruebas

En esta sección se describe las pruebas realizadas al sistema.

Se procedió a efectuar el sistema para la obtención de la transcripción al AFI correspondiente a 10 palabras diferentes, 5 en téenek y 5 en español. Éstos archivos de audio son independientes y algunos forman parte del trabajo realizado previamente por (Alviso Vargas, 2014). Estas muestras de audio no son incluidas en el corpus del cual hace uso este sistema.

Las pruebas que se realizaron para este trabajo fueron pruebas de funcionamiento, donde se observó la eficiencia del resultado proporcionado por el sistema para cada instancia. El estudio del resultado arrojado por la interfaz consistió en revisar que la transcripción fuese una aproximación lo más exacta de la transcripción correcta conocida.

Los resultados por cada ejecución se describen a continuación:

Bicho

biʧo	100.00%
biʧu	75.00%
biʧo	100.00%
biʧo	100.00%
tiʧu	75.00%
biʧo	100.00%
biʧo	100.00%
biʧo	100.00%
biʧo	100.00%
biʧo	100.00%
	95.00%

Akak

okak	75.00%
okak	75.00%
kkkk	50.00%
kkkk	50.00%
obak	50.00%
kʧiʧ	0.00%
okae	50.00%
okak	75.00%
okak	75.00%
okak	75.00%
	57.50%

Beca

beka	100.00%
beka	100.00%
beka	100.00%
beka	100.00%
bebe	50.00%
beka	100.00%
ueʧa	50.00%
beka	100.00%
beka	100.00%
beka	100.00%
	90.00%

Bo'

boo	100.00%
too	66.66%
boo	100.00%
uua	0.00%
boa	33.33%
boo	100.00%
boo	100.00%
boo	100.00%
boo	100.00%
boo	100.00%
	79.99%

Butaca

bubaka	83.33%
butaaa	83.33%
botaka	83.33%
bbtaka	83.33%
buaaka	83.33%
butaaa	83.33%
buaaka	83.33%
kuaaka	66.67%
tuaata	50.00%
tuuaka	66.67%
	76.66%

Buuk

buuk	100.00%
book	50.00%
buuu	75.00%
uaak	25.00%
buuo	75.00%
buuk	100.00%
buuo	75.00%
buuk	100.00%
buuk	100.00%
buuo	75.00%
	77.50%

Chaab

ʧaaa	75.00%
ʧaob	75.00%
ʧoku	25.00%
ʧoku	25.00%
ʧaao	75.00%
ʧooo	25.00%
ʧooo	25.00%
ʧaao	75.00%
ʧaaa	75.00%
ʧaob	75.00%
	55.00%

Chacal

ʧakaa	80.00%
ʧakaa	80.00%
ʧaaae	60.00%
ʧakab	80.00%
ʧakao	80.00%
ʧakal	100.00%
ʧatal	80.00%
ʧaaaa	60.00%
ʧakaa	80.00%
ʧakal	100.00%
	80.00%

Che'

ʧee	100.00%
ʧie	66.66%
ʧee	100.00%
ʧea	66.66%
ʧee	100.00%
ʧee	100.00%
ʧee	100.00%
ʧee	100.00%
ʧea	66.66%
ʧie	66.66%
	86.66%

Chequeo

ʧeteo	80.00%
ʧeeeo	80.00%
ʧeʧeo	80.00%
ʧeteo	80.00%
ʧeueo	80.00%
ʧeʧeo	80.00%
ʧeʧeo	80.00%
ʧeeeo	80.00%
ʧekeo	100.00%
ʧeteo	80.00%
	82.00%

Resumiendo,

Palabra analizada	Idioma	Eficiencia
akak	Téenek	57.50%
boo	Téenek	79.99%
buuk	Téenek	77.50%
chaab	Téenek	55.00%
chee	Téenek	86.66%
beca	Español	90.00%
bicho	Español	95.00%
butaca	Español	76.66%
chacal	Español	80.00%
chequeo	Español	82.00%
	Total	**78.03%**

***Tabla 7**: Porcentajes de eficiencia obtenidos en promedio por cada palabra.*

Los resultados obtenidos en esta experimentación dejan en evidencia que el sistema tiene, en promedio, al menos un 78.03% de efectividad.

Como se mencionó anteriormente, la única investigación que ha trabajado en la identificación de palabras en el idioma téenek es la desarrollada por (Alviso Vargas, 2014). Los resultados para la versión que hace uso de la combinación de MFCC, el clasificador Naive Bayes equiprobable y HMM con error en su entrenamiento (sucio), diseñada para uso en instancias obtenidas de locutores no entrenados —la cual es la combinación más cercana a la implementada en esta investigación—, se muestran en la Tabla 8.

Palabra analizada	Idioma	Eficiencia
akak	Téenek	100.00%
boo	Téenek	60.00%
buuk	Téenek	90.00%
chaab	Téenek	80.00%
chee	Téenek	100.00%
juun	Téenek	40.00%
oox	Téenek	40.00%
waxik	Téenek	100.00%
beleeu	Téenek	100.00%
laaju	Téenek	20.00%
	Total	**73.00%**

***Tabla 8**: Resultados de la implementación realizada por (Alviso Vargas, 2014) que hace uso de la combinación de MFCC, Naive Bayes (equiprobable) y HMM.*

Capítulo 6. Conclusiones

6.1. Objetivos cumplidos

6.1.1. Objetivo general

Se desarrolló un sistema LID que hace uso de la transcripción de una muestra de audio —que sólo contiene una única palabra—, proveniente de un hablante de la lengua téenek a un subconjunto de fonemas del AFI y que en base a ésta realiza la identificación del idioma.

6.1.2. Objetivos específicos

- Se definió la estrategia con la cual se logro abordar el problema de transcribir los datos de voz a texto, encontrando tanto los métodos de representación adecuados como los métodos de detección que mejor se ajustaron a la identificación de fonemas.
- Se obtuvieron algunas muestras de audio con palabras utilizadas en la lengua téenek y en el idioma español, con el fin de crear el modelo acústico del sistema.
- Se realizó el entrenamiento del sistema.
- Se ejecutaron un conjunto de pruebas con el fin de medir la calidad de la transcripción realizada.
- Se comprobó el grado de eficiencia para reconocer los sonidos acústicos correctamente.
- Se realizaron pruebas en el sistema para encontrar puntos de mejora.

6.2. Conclusiones y comentarios finales

La investigación reveló ciertos problemas que no habían sido contemplados con anterioridad. Es conocido que aunque el téenek es un idioma bien establecido, las variantes del mismo suelen tener, a su vez, aún más variantes —como cualquier otro idioma—. La información anteriormente recolectada con respecto a la que posteriormente fue obtenida se resaltan los siguientes puntos:

- Los números 4, 5, 9 y 10 cambian de pronunciación en la variante hablada en Tantoyuca.
- La variante de Tantoyuca, por lo visto, no hace el uso extensivo de la doble vocal como si lo

hacen las demás variantes.

- La pronunciación de la letra «j» en la variante hablada en la región de Tantoyuca se pronuncia como [h] (aspirada) y no [x] (Pison Alcaraz, 2014), como se hace en San Francisco (Kondic, 2012).
- La forma de pronunciar las palabras y de escribirlas varía en cada variante.
- No todos los hablantes de téenek saben escribir en su idioma.

El sistema se había planteado para que fuera pleno en sus funciones y con total automatización, sin embargo, esto no será posible en un futuro cercano. Dado que cada parte del sistema es, en esencia, un nuevo descubrimiento, muchos procedimientos se tendrán que realizar de una forma manual. Por lo que se dio prioridad a comprobar la factibilidad del RAH para realizar transcripciones al AFI y a la identificación del idioma.

Adicionalmente se hacen las siguientes observaciones:

- Los lingüistas entrevistados no concuerdan totalmente en algunas de las transcripciones al AFI de cada palabra, por lo que es mejor posponer el trabajo en las palabras que presenten tales inconsistencias.
- Las pruebas con distintos clasificadores desde el explorador de WEKA, así como las realizadas desde el mismo sistema, comprueban la viabilidad de predicción de fonemas desconocidos.
- Los resultados obtenidos hasta el momento demuestran que hace falta un *corpus* mucho más robusto de lo que se pensaba para lograr una predicción más acertada que la obtenida en los sistemas RAH convencionales.
- Será necesario realizar de forma manual el etiquetado de los fonemas desconocidos en la muestra a identificar, dejando para trabajo futuro la solución al *Language Independent Automatic Speech Segmentation and Labeling Problem*.

Cabe aclarar que aunque es una primera aproximación, los resultados obtenidos en esta tesis, en general, son positivos.

Es posible afirmar abiertamente que el sistema comprueba la hipótesis principal y que lo propuesto cumple las expectativas.

En el desarrollo de esta investigación, desde la perspectiva social, se comprobó tristemente la marginación de la cultura téenek. La modernización, globalización, clasismo, racismo y otros graves

problemas éticos y morales han provocado que los hablantes del huasteco lleguen incluso a ocultar que tienen dominio completo en ambos idioma —téenek y español—; y en las regiones donde se habla de forma nativa, los jóvenes llegan a preferir sólo aprender español, siendo esta la razón por lo que la mayoría no sabe escribir su idioma —dado que sólo lo hablan debido a que lo aprendieron de sus madres de cuando niños—. Un par de ocasiones, de manera extraoficial y no comprobable, algunas personas comentaron que algunos miembros de primaria y secundaria prohíben que se hable en téenek; quizás porque los docentes no saben hablarlo o por decisión de ellos.

6.3. Aportaciones de la investigación

Esta investigación continuó la línea de pensamiento de los trabajos desarrollados por (Hernández Zepeda, 2013) y (Alviso Vargas, 2014) —por nombrar algunos—, y gracias a esto, ahora es posible encontrar más trabajos relacionados con el idioma téenek.

Como resultado del trabajo descrito en toda esta tesis, se logró crear un Identificador Automático del Lenguaje, que no sólo es capaz de referir el lenguaje de algunas palabras en téenek sino que también palabras en español. Y dado el enfoque con el que fue creado, es posible agregar fácilmente más idiomas agregando palabras a *la base de datos de transcripciones*.

Otro hito fue haber creado un transcriptor de datos de voz al Alfabeto Fonético Internacional. Es en realidad gracias a esto que el LID consigue las cualidades antes descritas.

Por último, se han publicado 2 artículos respecto al LID y al transcriptor desarrollados en base a la investigación realizada para la escritura de esta tesis.

6.4. Trabajo futuro

Debido a que esta tesis fue un primer acercamiento fonotáctico tanto a la lengua Téenek como al uso del Alfabeto Fonético Internacional de una forma tan particular —en un transcriptor—, y a que el tiempo para desarrollar la idea fue un tanto breve, cada parte del sistema LID tiene una gran variedad de áreas de oportunidad para mejorar.

De entre todos los componentes con puntos de mejora de Kibo se puede destacar el *corpus*. Como en todos los desarrollos computacionales de este tipo de problemas, un *corpus* robusto puede representar una mejora sustancial para el sistema en general. Conseguir un *corpus* con esta característica debe ser una prioridad para cualquier versión posterior ya que esto repercutirá directamente en la eficiencia al reconocer el idioma de las muestras.

Finalmente, también es necesario trabajar en conseguir la total automatización del sistema con el fin de facilitar una posible implementación de este sistema como aplicación para usuarios finales.

Capítulo 7. Bibliografía

Alviso Vargas, J. (2014). *Reconocimiento de habla en palabras aisladas en lenguas de San Luis Potosí*. Instituto Tecnológico de Ciudad Madero, Ciudad Madero, Tamaulipas, México.

Anderson, O., Dalsgaard, P., & Barry, W. (1994). On the use of data-driven clustering technique for identification of poly- and mono-phonemes for four European languages (Vol. i, p. I/121-I/124). IEEE. http://doi.org/10.1109/ICASSP.1994.389340

Apple Inc. (2014). Siri (Versión 8) [IOS]. Cupertino, California, USA: Apple Inc. Recuperado a partir de https://www.apple.com/ios/siri/

Attanayake, D. (2012, noviembre 27). Sphinx3 Phoneme set and the IPA? *Speech Recognition*. Recuperado a partir de http://sourceforge.net/p/cmusphinx/discussion/speech-recognition/thread/0910a220/#5d21

Audacity development team, Mazzoni, D., & Dannenberg, R. (2015). Audacity (Versión 2.0.6) [GNU/Linux]. Pittsburgh, Pennsylvania, USA. Recuperado a partir de http://audacityteam.org/

Boersma, P., Weenink, D., & University of Amsterdam. (2014). Praat (Versión 5.4) [Microsoft Windows 7]. Amsterdam, Netherlands. Recuperado a partir de http://www.fon.hum.uva.nl/praat/

Butterfield, J., González, M., & Breslin, G. (1995). *Collins compact diccionario inglés: español-inglés, English-Spanish* (Vol. 1). Glasgow; Barcelona: HarperCollins ; Grijalbo.

Caballero-Morales, S.-O. (2013). On the Development of Speech Resources for the Mixtec Language. *The Scientific World Journal*, *2013*, 1-19. http://doi.org/10.1155/2013/170649

Carnegie Mellon University. (2014). *CMU Sphinx - Open Source Toolkit For Speech Recognition*. Pittsburgh, Pennsylvania, USA: Carnegie Mellon University. Recuperado a partir de http://cmusphinx.sourceforge.net/

Colaboradores de Wikipedia. (2014a). Alfabeto Fonético Internacional — Wikipedia, La enciclopedia libre. Recuperado a partir de https://es.wikipedia.org/wiki/Alfabeto_Fon%C3%A9tico_Internacional

Colaboradores de Wikipedia. (2014b). Arpabet — Wikipedia, La enciclopedia libre. Recuperado a partir de https://en.wikipedia.org/wiki/Arpabet

Colás Pasamontes, J. (2001). Estrategias de incorporación de conocimiento sintáctico y semántico en sistemas de comprensión de habla continua en español, *12*. Recuperado a partir de http://elies.rediris.es/elies12/index.html

Consejo Nacional del Instituto Nacional de Lenguas Indígenas, & Gaxiola Moraila, F. (2007). *Catálogo de las Lenguas Indígenas Nacionales. Variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas* (p. 256). México, D.F.: Instituto Nacional de Lenguas Indígenas. Recuperado a partir de http://www.purepecha.mx/files/CLIN_completo.pdf

Contreras Morales, G. (2007). *Digitalizador de voz* (p. 13). Guadalajara, Jalisco, México: Universidad de Guadalajara. Recuperado a partir de http://proton.ucting.udg.mx/materias/ET201/modulo_11/2007A/Digitalizador_de_Voz.pdf

Daniels, P. T., & Bright, W. (Eds.). (1996). *The world's writing systems*. New York: Oxford University Press.

Flores Hernández, Y. (2014). Transcripciones al AFI del téenek de Tantoyuca [Comunicación por correo electrónico].

Flores Paulín, J. C. (2009, diciembre). *Técnicas para el reconocimiento de voz en palabras aisladas en la lengua náhuatl*. Instituto Politécnico Nacional, México, D.F. Recuperado a partir de http://www.saber.cic.ipn.mx/cake/SABERsvn/trunk/Repositorios/webVerArchivo/5331/2

Google Inc. (2014). Google Now (Versión 3.6) [Android]. Googleplex, Mountain View, California, USA: Google Inc. Recuperado a partir de https://www.google.com/landing/now/

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *II*(1).

Hauenstein, A. (1996). *The Syllable Re-revisited*. ICSI. Recuperado a partir de https://books.google.com.mx/books?id=SBvsmgEACAAJ

Hernández Zepeda, C. A. (2013, agosto). *Corpus de las Lenguas Indígenas Tének, Náhuatl y Xi'iuy para la Identificación Automática del Lenguaje Hablado*. Instituto Tecnológico de Ciudad

Madero, Ciudad Madero, Tamaulipas, México.

House, A. S. (1977). Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *The Journal of the Acoustical Society of America*, *62*(3), 708. http://doi.org/10.1121/1.381582

Hunt, A., & Speech Applications Group, Sun Microsystems Laboratories. (1997, septiembre 5). comp.speech Frequently Asked Questions WWW site [FAQ]. Recuperado a partir de http://www.speech.cs.cmu.edu/comp.speech/

Instituto Nacional de Estadística Geografía e Informática (México). (2011). *Panorama sociodemográfico de México*. Aguascalientes: Instituto Nacional de Estadística Geografía e Informática.

International Phonetic Association (Ed.). (1999). *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge, U.K. ; New York, NY: Cambridge University Press.

John Christopher Wells, & Department of Phonetics and Linguistics, University College London. (2000, mayo 3). Computer-coding the IPA: a proposed extension of SAMPA. Recuperado 19 de mayo de 2014, a partir de http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm

Jones, D. (1972). *An outline of English phonetics* (9th ed). Cambridge: Heffer.

Jones, R. J., Downey, S., & Mason, J. S. (1997). Continuous speech recognition using syllables. En *In Proc. Eurospeech '97* (pp. 1171–1174).

Kawahara Lab., Kyoto University, Information-technology Promotion Agency, Japan, Shikano Lab., Nara Institute of Science and Technology, & Julius project team, Nagoya Institute of Technology. (s. f.). Open-Source Large Vocabulary CSR Engine Julius (Versión 4.3.1). Pittsburgh, Pennsylvania, USA. Recuperado a partir de http://julius.osdn.jp/en_index.php

Kondic, A. (2012). Narraciones en huasteco de San Francisco. *Tlalocan*, *XVIII*, 35–78.

Kraiss, K.-F. (Ed.). (2006). *Advanced man-machine interaction: fundamentals and implementation*. Berlin ; New York: Springer.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford, OX, UK ;

Cambridge, Mass., USA: Blackwell Publishers.

Larsen, R., & Instituto Lingüístico de Verano, A.C. (1997). *Vocabulario huasteco del estado de San Luis Potosí*. D.F., México: Instituto Lingüístico de Verano en cooperación con la Dirección General de Asuntos Indígenas de la Secretaría de Educación Pública. Recuperado a partir de http://www-01.sil.org/mexico/maya/huasteco-sanluispotosi/G046b-VocHuastecoFacs-hus.pdf

Li, K., & Edwards, T. (1980). Statistical models for automatic language identification (Vol. 5, pp. 884-887). Institute of Electrical and Electronics Engineers. http://doi.org/10.1109/ICASSP.1980.1170832

Lleida, E., & Rose, R. C. (2000). Utterance verification in continuous speech recognition: decoding and training procedures. *IEEE Transactions on Speech and Audio Processing*, *8*(2), 126-139. http://doi.org/10.1109/89.824697

Malagón Luque, C. (2003, mayo 14). *Clasificadores bayesianos. El algoritmo Naïve Bayes.* Madrid, España. Recuperado a partir de http://www.nebrija.es/~cmalagon/inco/Apuntes/bayesian_learning.pdf

Manikandan, J., Venkataramani, B., Preeti, P., Sananda, G., & Sadhana, K. V. (2009). Implementation of a phoneme recognition system using zero-crossing and magnitude sum function (pp. 1-5). IEEE. http://doi.org/10.1109/TENCON.2009.5395954

Meza Ruiz, I. V. (2013, marzo 28). MFCCs. Recuperado a partir de http://turing.iimas.unam.mx/~ivanvladimir/es/post/MFCC/

Navarro Mesa, J. L. (2005). *Procesador Acústico: El Bloque de Extracción de Características*. Las Palmas de Gran Canaria, Las Palmas, España: Universidad de Las Palmas de Gran Canaria (ULPGC). Recuperado a partir de http://www2.ulpgc.es/hege/almacen/download/25/25296/apuntesextraccioncaracterisitcas.pdf

Nuance Communications, Inc. (2014). Dragon NaturallySpeaking (Versión 13). Burlington, Massachusetts, USA: Nuance Communications, Inc. Recuperado a partir de http://www.nuance.com/for-individuals/index.htm

Oropeza Rodríguez, J. L., & Suárez Guerra, S. (2009). Algoritmos y Métodos para el Reconocimiento

de Voz en Español Mediante Sílabas. *Computación y Sistemas*, *9*(003), 270 - 286.

Pison Alcaraz, P. (2014). Transcripciones al AFI del téenek de Tantoyuca [Comunicación por correo electrónico].

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., … Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. En *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society.

Prieto Izquierdo, O. J., & Casillas Díaz, R. (2011, septiembre 29). *Aprendizaje Bayesiano*. Valladolid, España. Recuperado a partir de http://www.infor.uva.es/~isaac/doctorado/AprendizajeBayesiano.pdf

Runstein, F., & Violaro, F. (1996). An isolated-word speech recognition system using neural networks (Vol. 1, pp. 550-553). IEEE. http://doi.org/10.1109/MWSCAS.1995.504498

Schultz, T., & Kirchhoff, K. (Eds.). (2006). *Multilingual speech processing*. Amsterdam ; Boston: Elsevier Academic Press.

Su-Lin Wu, Kingsbury, E. D., Morgan, N., & Greenberg, S. (1998). Incorporating information from syllable-length time scales into automatic speech recognition (Vol. 2, pp. 721-724). IEEE. http://doi.org/10.1109/ICASSP.1998.675366

Su-Lin Wu, Shire, M. L., Greenberg, S., & Morgan, N. (1997). Integrating syllable boundary information into speech recognition (Vol. 2, pp. 987-990). IEEE Comput. Soc. Press. http://doi.org/10.1109/ICASSP.1997.596105

Timoshenko, E. (2012, julio 2). *Rhythm information for automated spoken language identification*. LMU München, München, Deutschland. Recuperado a partir de https://mediatum.ub.tum.de/doc/1063301/1063301.pdf

UCL Division of Psychology & Language Sciences, & Wells, J. (2005, octubre 25). Handbook of Standards and Resources for Spoken Language Systems. Recuperado 19 de mayo de 2014, a partir de http://www.phon.ucl.ac.uk/home/sampa/index.html

Unión Editorialista, S.A. de C.V. (2013, junio 21). A la baja, cantidad de hablantes de lenguas indígenas

en México. *El Informador*. Guadalajara, Jalisco, México. Recuperado a partir de http://www.informador.com.mx/jalisco/2013/473192/6/a-la-baja-cantidad-de-hablantes-de-lenguas-indigenas-en-mexico.htm

United Nations, & Department of Public Information. (2009). *Las Naciones Unidas hoy*. Nueva York: Naciones Unidas. Recuperado a partir de http://www.un.org/es/aboutun/untoday/index.shtml

United Nations, & Permanent Forum on Indigenous Issues (Eds.). (2009). *State of the World's Indigenous Peoples*. New York: United Nations.

Yusnita, M. A., Paulraj, M. P., Yaacob, S., Abu Bakar, S., Saidatul, A., & Ahmad Nazri Abdullah. (2011). Phoneme-based or isolated-word modeling speech recognition system? An overview (pp. 304-309). IEEE. http://doi.org/10.1109/CSPA.2011.5759892

Zhang, H. (2004). The Optimality of Naive Bayes. En *The Florida AI Research Society Conference* (Vol. 2).

Capítulo 8. Anexo A: Algoritmo propuesto para el sistema LID «Kibo»

A continuación, se muestra el algoritmo utilizado por Kibo para identificar una muestra de audio desconocida, previamente segmentada (refiérase al capítulo 5.7).

```
LID(conjunto-entrenamiento, caso-prueba-kib, base-datos-transcripciones)
    Si conjunto-entrenamiento es nulo detener la ejecución
    Si caso-prueba-kib es nulo detener la ejecución
    Si base-datos-transcripciones es nulo detener la ejecución
    En otro caso
        Para cada archivo en caso-prueba-kib
            Guardar un archivo .ARFF con el resultado de Obtener-MFCC(archivo)
        Entrenar el sistema con conjunto-entrenamiento
        Para cada archivo .ARFF generado de caso-prueba-kib
            Usar el clasificador Naive Bayes de Weka para inferir el fonema
        Concatenar en cadena los fonemas inferidos
        Conectar a base-datos-transcripciones
        Buscar una coincidencia de cadena en base-datos-transcripciones
        Mostrar el idioma de la coincidencia
```