

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN



"POR MI PATRIA Y POR MI BIEN"

TESIS

**“CORPUS DE LAS LENGUAS INDÍGENAS TÉNEK, NÁHUATL
Y XI’IUY PARA LA IDENTIFICACIÓN AUTOMÁTICA DEL
LENGUAJE HABLADO”**

Para obtener el Grado de:

Maestro en Ciencias de la Computación

Presenta:

ISC Carlos Arturo Hernández Zepeda

Director de Tesis:

Dr. Juan Javier González Barbosa

Codirector de Tesis:

Dr. Arturo Hernández Ramírez

(Esta página ha sido dejada en blanco intencionalmente)

"2013, Año de la Lealtad Institucional y Centenario del Ejército Mexicano"

Ciudad Madero, Tamps; a **25 de Septiembre de 2013.**

OFICIO No.: U5.251/13
AREA: DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN
ASUNTO: AUTORIZACIÓN DE IMPRESIÓN DE TESIS

ING. CARLOS ARTURO HERNÁNDEZ ZEPEDA
NO. DE CONTROL G06070532
PRESENTE

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

PRESIDENTE :	DR. JUAN JAVIER GONZÁLEZ BARBOSA
SECRETARIO :	M.C. JOSÉ APOLINAR RAMÍREZ SALDIVAR
VOCAL :	DR. ARTURO HERNÁNDEZ RAMÍREZ
SUPLENTE	DRA. MARÍA LUCILA MORALES RODRÍGUEZ
DIRECTOR DE TESIS :	DR. JUAN JAVIER GONZÁLEZ BARBOSA
CO-DIRECTOR:	DR. ARTURO HERNÁNDEZ RAMÍREZ

Se acordó autorizar la impresión de su tesis titulada:

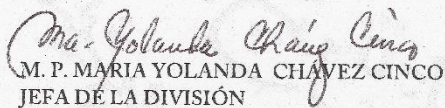
"CORPUS DE LENGUAS INDÍGENAS TÉNEK, NÁHUATL Y XI'UY PARA LA IDENTIFICACIÓN AUTOMÁTICA DEL LENGUAJE HABLADO"

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

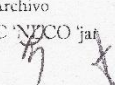
Espero que continúe con éxito su desarrollo profesional y dedique su experiencia e inteligencia en beneficio de México.

ATENTAMENTE

"Por mi patria y por mi bien"


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



c.c.p.- Minuta
Archivo
MYCHC 'NICO jat


S.E.P.
DIVISIÓN DE ESTUDIOS
DE POSGRADO E
INVESTIGACIÓN
ITCM



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

www.itcm.edu.mx



Declaración de Originalidad

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

Además, declaro que en 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.

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

16 de Octubre del 2013, Ciudad Madero, Tamaulipas.

ISC Carlos Arturo Hernández Zepeda

Entendámonos

Esta parte iba a tener como título “Agradecimientos”, pero decidí mejor dejarle el nombre del corpus que se elaboró gracias a la ayuda de aquellas personas que nombrare en esta hoja. Y es que, a opinión propia, este trabajo se pudo concluir gracias a una serie de causas y azares que normalmente sucedían para bien. Por ejemplo, y para empezar, fue una increíble suerte que para recolectar muestras del idioma *náhuatl* me tocara ir tantas veces al pueblo de *Matlapa*, lugar en donde pude volver a encontrar a mis amigos, *Don Moisés Pardiñas Melo* y la *Señora Adriana Osorno Vargas* (dos médicos muy excepcionales, por cierto), y a sus hijos, *Moi* y *Ami*, todos parte de una familia que me ayudó en muchísimas formas para que yo pudiera recolectar mis muestras y que la pasara de lo mejor en ese lugar.

Por otro lado, todo tiene un comienzo, y ese comienzo empezó cuando mi codirector y yo decidimos informarnos por primera vez acerca de las lenguas indígenas en México, particularmente, si existía algún corpus de habla de lenguas indígenas mexicanas que pudiésemos usar (que al final de ese mismo día descubrimos que no existía). Para esto, fuimos al *INALI*, ubicado en la *Ciudad de México*, en donde nos atendió una persona que tuvo un papel importante para que este proyecto comenzara, el *Director de Políticas Lingüísticas, Arnulfo Embriz Osorio*, quien no sólo nos ilustró un panorama muy completo acerca del tema, sino que nos ayudó con material bibliográfico y nos hizo el favor de organizarnos una cita para ir a un evento en *Tancanhuitz de Santos*, el cual fue el primer lugar en el que se recolectaron muestras de habla.

Para continuar el orden, después de haber concluido el primer viaje de recolección de muestras, decidimos buscar alguna institución en la *Huasteca Potosina* que nos pudiera ayudar para continuar la fase de recolección. Y fue entonces cuando encontramos en Matlapa al *IELIIP* y a su director, el *Prof. Juan Manuel González Vidales*, quien fue la persona que brindó y coordinó la mayor cantidad de ayuda al proyecto. Dicha ayuda brindada por el *IELIIP*, podría traducirse en 3 personas: el *Prof. Agustín Reyes Antonio*, la *Profra. María Josefa Hernández de la Cruz* y *Prof. Isaías Rosalino Martínez*, quienes fueron los lingüistas (de náhuatl, xi'iuy y tének, respectivamente) encargados de brindar ayuda y asistencia para la recolección de muestras y anotaciones. Especialmente usted, *Maestra María Josefa*, ya que sin su ayuda, conocimiento y recursos, la recolección de muestras del idioma xi'iuy jamás hubiera podido suceder.

También es importante mencionar a mis amigos, quienes con buenos momentos de risas, discusiones y alegría me apoyaron durante todo momento.

Personas de habla indígena que aceptaron donar su voz sin esperar recibir nada a cambio: Se los agradezco mucho. Sin ustedes yo no hubiese podido trabajar en algo tan genial como lo expresado en esta tesis.

Finalmente, *Pá, Má*, nunca se los hago saber porque nomás tantito les digo un cumplido y ya se creen (no es cierto), pero mejores padres no puede haber. Lo que soy y lo que seré no podría ser sin lo que me han enseñado (o de perdido intentado).

Resumen

El presente trabajo expone la investigación y desarrollo del corpus *Entendámonos*, el cual es un *corpus de habla* diseñado específicamente para tratar el problema de la *Identificación Automática del Lenguaje Hablado*. Dicho corpus está conformado por grabaciones de voz de las lenguas indígenas habladas en el estado de *San Luis Potosí*, las cuales son: *Tének*, *Náhuatl* y *Xi'iu*.

La Identificación Automática del Lenguaje Hablado (*LID*) es la tarea de reconocer automáticamente el lenguaje que se habla en una muestra de audio. Por ejemplo, la *LID* puede ser utilizada en cualquier sistema de comunicación humano-máquina o humano-humano, cuando inicialmente no se conoce el idioma que se está hablando o que debería ser utilizado. Así mismo, un corpus de habla (*Speech Corpus*) es una base de datos de archivos de audio de habla y/o transcripciones de texto. Estos corpus son utilizados, entre otras cosas, para crear modelos acústicos, los cuales son utilizados por sistemas *LID*, sistemas de reconocimiento de habla, entre otros.

El sistema *LID* implementado en esta tesis considera únicamente el análisis acústico, utilizando la *transformada wavelet* para diferenciar las altas y bajas frecuencias de la señal de habla, partiendo de la hipótesis en donde las bajas frecuencias poseen características para representar la prosodia. Posteriormente se realiza un análisis estadístico sobre las bajas frecuencias para construir los modelos acústicos.

La investigación realizada para el desarrollo del corpus *Entendámonos* abarca lo que en general se considera necesario para construir un corpus de habla, aplicado claro, para las lenguas indígenas mexicanas, tomando en cuenta diversos tipos de tópicos y temas, ya sean técnicos o sociales.

Los procesos de evaluación *LID* realizados, utilizando *Entendámonos*, consideran diferentes tipos de pruebas, realizando análisis de 10, 5 y 3 segundos sobre las muestras de audio. Los resultados del sistema *LID* muestran porcentajes de identificación de un 68%, tomando como base 40 muestras de habla por idioma, cuyo promedio de duración por cada muestra es de 50 segundos.

Finalmente, esta tesis representa un estudio pionero sobre las lenguas indígenas mexicanas, avalado por el *Director de Políticas Lingüísticas del Instituto Nacional de las Lenguas Indígenas (INALI)*, debido a que, en base al *Estado del Arte* expuesto, es la primera vez que se ven a las lenguas indígenas mexicanas desde un punto de vista más técnico en el sentido que se utiliza un corpus estandarizado, y cuya finalidad es abrir las puertas al desarrollo de aplicaciones de habla para la población multilingüista del estado de *San Luis Potosí* y, en un futuro, de *México* en general.

Abstract

This paper presents the research and development of the *Entendámonos* corpus, which is a *speech corpus* specifically designed to deal with the *Automatic Language Identification* problem. This corpus consists of voice recordings of indigenous languages spoken in the state of *San Luis Potosi*, which are: *Tének*, *Náhuatl* and *Xi'iu*.

The Automatic Language Identification (*LID*) is the task of automatically recognizing the language spoken in an audio sample. For example, the *LID* can be used in any human-machine or human-human communication system, when you initially do not know the language that someone is talking about or should be used. Also, a speech corpus is a database of speech audio files and/or text transcripts. These corpus are used, among other things, to create acoustic models, which are used by *LID* systems, speech recognition systems, etc.

The implemented *LID* system in this thesis considers only the acoustic analysis using the *wavelet transform* to differentiate the high and low frequencies of the speech signals, based on the hypothesis where low frequencies have features to represent prosody. Subsequently, a statistical analysis over the low frequencies is performed to build the acoustic models.

The research for the corpus development covers what is generally considered necessary to build a speech corpus, clearly applied to the mexican indigenous languages, taking into account various types of topics and issues, whether technical or social.

The *LID* evaluations performed using *Entendámonos*, consider different types of tests, doing analysis of 10, 5 and 3 seconds to the audio samples. The system results show average identification percentages of 68%, based on 40 samples by language, whose average duration per sample is 50 seconds.

Finally, this thesis represents a pioneer study of mexican indigenous languages, supported by the *Director de Políticas Lingüísticas* del *Instituto Nacional de las Lenguas Indígenas (INALI)*, because, based on the state of the art exhibit, this is the first time that anyone deals with the mexican indigenous languages from a technical point of view in the sense that a standardized corpus its used, and which purpose is to open the door to developing applications for the multilingual population of the state of *San Luis Potosi*, and in one future, of *Mexico* in general.

(Esta página ha sido dejada en blanco intencionalmente)

Índice

1. Introducción.....	1
1.1 Definición del Problema.....	3
1.2 Objetivo	3
1.2.1 Objetivos específicos.....	3
1.3 Justificación	3
1.4 Alcances, Limitaciones y Trabajo Realizado	4
1.4.1 Limitaciones	5
1.4.2 Alcances	6
1.4.3 Trabajo Realizado.....	6
1.5 Estructura de la Tesis.....	6
2. Praeludium.....	8
2.1 Preliminares	8
2.1.1 Corpus de Habla	8
2.1.2 Identificación Automática del Lenguaje Hablado (LID).....	10
2.2 Antecedentes.....	13
2.3 Lenguas Indígenas	17
2.3.1 Náhuatl	21
2.3.2 Tének	22
2.3.3 Xi'iuy.....	22
3. Estado del Arte	24
3.1 Florian Shiel et al.....	24
3.2 Lwazi	25
3.3 Jiří Navrátil.....	26
3.4 Ekaterina Timoshenko.....	27
3.5 Reyes Herrera y Vargas Martínez	29
3.6 Flores Paulín y Caballero Morales	31
4. Corpus Entendámonos.....	34
4.1 Especificación del corpus	35

4.2 Zonas, Localidades y Ubicaciones	36
4.2 Tipos de Habla y Cuestionarios.....	40
4.3 Perfil de los Hablantes y Número de Hablantes	43
4.3 Protocolo de Grabación	44
4.4 Especificaciones Técnicas	46
4.5 Post-procesamiento.....	47
4.5.1 Filtrado	48
4.5.2 Editado.....	49
4.5.3 Asignación de formato	50
4.5.4 Asignación de nombre.....	50
4.6 Anotaciones	51
4.7 Metadatos	55
4.8 Estructura del Corpus	57
5. Evaluación	58
5.1 Metodología.....	58
5.1.1 Señales de Habla de Entrenamiento/Prueba	59
5.1.2 Segmentación	59
5.1.3 Transformada Wavelet	59
5.1.4 Truncado por Fracción.....	61
5.1.5 Medidas Estadísticas	61
5.1.6 Entrenamiento y Modelos.....	61
5.1.7 Identificación.....	62
5.2 Pruebas y Resultados	63
6. Occāsus.....	65
6.1 Conclusiones.....	65
6.2 Aportaciones de la Investigación.....	66
6.3 Trabajo Futuro	67
6.3.1 Anotaciones	67
6.3.2 Número de Hablantes	67
6.3.3 Metadatos	68
6.3.4 Tecnologías del Habla	68
A. Marco Teórico	69

A.1	Desarrollo de un Corpus de Habla.....	69
A.1.1	Especificaciones del Corpus.....	70
A.1.1.1	Perfil de los Hablantes.....	71
A.1.1.2	Número de Hablantes.....	72
A.1.1.3	Contenido Hablado.....	72
A.1.1.3.1	Vocabulario.....	73
A.1.1.3.2	Dominio.....	73
A.1.1.3.3	Tarea.....	73
A.1.1.3.4	Distribución Fonológica.....	74
A.1.1.4	Estilos de Habla.....	74
A.1.1.4.1	Habla Leída.....	74
A.1.1.4.2	Habla en Respuesta a Información Específica.....	75
A.1.1.4.3	Habla de Control.....	75
A.1.1.4.4	Habla Descriptiva.....	75
A.1.1.4.5	Habla No-Apuntada.....	76
A.1.1.4.6	Habla Espontánea.....	76
A.1.1.4.7	Neutral vs. Emocional.....	77
A.1.1.5	Tipos de Grabación.....	77
A.1.1.5.1	Grabaciones telefónicas.....	79
A.1.1.5.2	Grabaciones de sitio.....	79
A.1.1.5.3	Grabaciones de campo.....	80
A.1.1.5.4	Grabaciones en un entorno Wizard-of-Oz.....	80
A.1.1.6	Anotaciones.....	81
A.1.1.7	Especificaciones Técnicas.....	81
A.1.1.7.1	Frecuencia de Muestreo.....	81
A.1.1.7.2	Tipo y Formato de Muestra.....	82
A.1.1.7.3	Número de Canales, Intercalado.....	82
A.1.1.7.4	Formatos de Archivo.....	82
A.1.1.7.4.1	Formatos de Señales.....	83
A.1.1.7.4.2	Formatos de Anotaciones.....	83
A.1.1.7.4.3	Formatos de Metadatos.....	84
A.1.1.7.4.4	Formatos de Lexicon.....	84

A.1.1.8	Estructura de un Corpus de Habla	85
A.1.1.8.1	Estructura.....	85
A.1.1.8.2	Nombres de Archivo.....	86
A.1.1.8.3	Medios de Distribución	86
A.1.1.9	Metadatos.....	87
A.1.1.9.1	Protocolos de Grabación.....	88
A.1.1.9.2	Perfiles de los Hablantes.....	89
A.1.1.9.3	Comentarios.....	90
A.1.1.11	Procesos de Validación.....	91
A.1.1.11.1	Validaciones Internas Vs. Externas	91
A.1.1.11.2	¿Cuándo Validar?	92
A.1.1.11.3	¿Qué Validar?.....	93
A.1.1.11	Documentación	93
A.1.2	Recolección.....	94
A.1.2.1	Instrucciones para el Hablante.....	95
A.1.2.2	Técnicas de Grabación.....	95
A.1.2.2.1	Grabaciones en Sitio.....	95
A.1.2.2.2	Grabaciones de Campo.....	96
A.1.2.3	Cuestionarios y Formularios.....	97
A.1.2.4	Aspectos Legales	97
A.1.2.5	Plan de Reclutamiento	98
A.1.2.5.1	Incentivos.....	98
A.1.3	Post-procesamiento.....	98
A.1.3.1	Transferencia de Archivos	99
A.1.3.2	Asignación de Nombre a los Archivos	99
A.1.3.2	Editado.....	99
A.1.3.3	Filtrado.....	100
A.1.3.4	Remuestreo	100
A.1.3.5	Conversión de Formato	101
A.1.3.6	Detección Automática de Errores	101
A.2	Teoría de Wavelets	101
A.2.1	Espacios de Hilbert.....	102

A.2.2	Ortogonalidad y Bases Ortonormales.....	102
A.2.3	Bases de la Función de Escala.....	103
A.2.4	Análisis Multiresolución.....	104
A.2.5	Bases Wavelet.....	104
A.2.5.1	Transformada Wavelet.....	105
A.2.5.2	Wavelets Ortonormales y Discretas	106
A.2.5.3	Relación dos-escala	107
A.2.5.4	Descomposición Wavelet (Algoritmo piramidal).....	108
A.2.6	Wavelet de Daubechies.....	111
A.3	Aprendizaje Automático.....	112
A.3.1	Aprendizaje Automático: Naive Bayes	112
A.3.2	Ganancia de Información.....	114
A.3.3	Validación Cruzada	114
B.	Equipo de Trabajo.....	116
B.1	Laptop: MackBook	116
B.2	Micrófono: SHURE PG42-USB	117
B.3	Laptop: Asus N56V	118
	Bibliografía.....	120

Índice de Figuras

Figura 1.1. Organismos públicos de México que trabajan para preservar...	3
Figura 2.1. Analogía: El mundo como un corpus de habla...	9
Figura 2.2. Niveles de abstracción de la señal por medio de un análisis...	11
Figura 2.3. Arquitectura general de un sistema LID usando...	12
Figura 2.4. Arquitectura general de un sistema LID.	13
Figura 2.5. Resumen de los resultados las pruebas LID hechas sobre...	17
Figura 2.6. Las 11 familias lingüísticas indoamericanas con presencia en...	19
Figura 2.7. 7 agrupaciones lingüísticas de las 11 familias. INALI.	20
Figura 2.8. 4 agrupaciones lingüísticas de las 11 familias. INALI.	21
Figura 3.1. Logo del proyecto Lwazi.	25
Figura 3.2. Diagrama del sistema LID fonotáctico.	28
Figura 3.3. Diagramas de los métodos para calcular las unidades <i>syllable-like</i> ...	28
Figura 3.4. Diagrama del sistema LID acústico.	29
Figura 3.5. Metodología propuesta en [8].	31
Figura 3.6. Resultados (en porcentaje de clasificación correcta) de todas las...	31
Figura 3.7. Repertorio de fonemas mixtecos.	33
Figura 4.1. Zona: La Huasteca Potosina (color verde). México.	34
Figura 4.2. Diagrama de las fases de desarrollo por las cuales paso el...	35
Figura 4.3. Área mexicana (color rojo) que el proyecto abarco.	37
Figura 4.4. Localidades (marcadores azules, amarillos y verdes) del estado...	37
Figura 4.5. Zona náhuatl abarcada por el proyecto (marcadores verdes)...	38
Figura 4.6. Zona tének abarcada por el proyecto (marcadores amarillos) y...	39
Figura 4.7. Zona xi'iuy norte abarcada por el proyecto (marcadores azules).	39
Figura 4.8. Zona xi'iuy sur abarcada por el proyecto (marcadores azules).	40
Figura 4.9. Ejemplo del nombre de una carpeta de un hablante, usando la...	45
Figura 4.10. Archivos resultantes de una grabación de un cuestionario...	46
Figura 4.11. Señal de 20s de habla. La parte seleccionada al principio con gris...	48
Figura 4.12. Para crear un <i>perfil de ruido</i> se le da clic en el botón 'Obtener'...	49

Figura 4.13. Esta figura se puede apreciar como el color gris oscuro se...	49
Figura 4.14. Señal de 11s. (1) Señal Original. (2) Selección del segmento...	50
Figura 4.15. Parte de la anotación de una muestra (correspondiente a la...	52
Figura 4.16. Parte de la anotación de una muestra (correspondiente a la...	53
Figura 4.17. Parte de la anotación de una muestra (correspondiente a la...	53
Figura 4.18. Parte de la anotación de una muestra (correspondiente a la...	54
Figura 4.19. Parte de la anotación de una muestra (correspondiente a la...	54
Figura 4.20. Metadatos de <i>Entendámonos</i> .	55
Figura 5.1. Estructura del sistema LID propuesto.	58
Figura 5.2. Señal de 10s, segmentada en bloques de 1s. Si la frecuencia...	59
Figura 5.4. Proceso de descomposición de la transformada wavelet con $\eta=3$.	60
Figura 5.5. Truncado por fracción del 1% aplicado a la tercera señal...	61
Figura 5.6. Grafica de una matriz A . La grafica de color azul, correspondiente...	62
Figura A.1. Proceso de desarrollo típico de un corpus de habla.	69
Figura A.2. Esquema de la descomposición en Series Wavelet.	111
Figura A.3. Función de aproximación ϕ_t (izquierda) y función de detalle ψ_t .	112
Figura B.1. Imágenes del MacBook utilizado.	116
Figura B.2. Micrófono SHURE PG42-USB (con su <i>araña</i> enroscada) y...	117
Figura B.3. Imágenes de la laptop Asus N56VJ utilizada.	118

Índice de Tablas

Tabla 2.1. Características sobresalientes del corpus OGI_HQ.....	14
Tabla 2.2. Distribución de las llamadas a través de los 10 idiomas del cor....	15
Tabla 2.3. Estadísticas de los hablantes del corpus OGI_TS.	16
Tabla 2.4. Matriz de confusión y promedio de identificación sobre la eval.....	16
Tabla 2.5. Las diez lenguas indígenas más habladas en México [15].	18
Tabla 2.6. <i>Machiopamitl</i> (abecedario) de la lengua náhuatl propuesto por.....	22
Tabla 2.7. <i>Nik'adh</i> (abecedario) de la lengua tének propuesto por el....	22
Tabla 2.8. Abecedario de la lengua xi'iuuy propuesto por el IELIIP.....	23
Tabla 3.1. Información general de 5 de los 11 idiomas del corpus Lwazi....	26
Tabla 3.2. Resultados (porcentaje de clasificación correcta) reportados por....	30
Tabla 4.1. Resumen general de <i>Entendámonos</i>	44
Tabla 5.1. Prueba de 3 segundos con $\eta = 1$	63
Tabla 5.2. Prueba de 3 segundos con $\eta = 2$	64
Tabla 5.3. Prueba de 10 segundos con $\eta = 1$	64
Tabla 5.4. Prueba de 10 segundos, con $\eta = 2$	64

1

Introducción

El termino *corpus*¹ (pl. *corpora*) es típicamente definido como un conjunto de datos recolectados y preparados para un uso específico, estos recursos son frecuentemente explotados para propósitos atípicos para los que fueron diseñados. En lingüística, un corpus es el estudio del lenguaje expresado en muestras y/o texto del mundo real [1]. Este estudio representa un acercamiento digestivo a la derivación de un conjunto de reglas abstractas por medio de las cuales, un lenguaje natural es gobernado o se relaciona con otro lenguaje.

Normalmente *hechas a mano*, las corpora están ampliamente relacionadas con diversos procesos automáticos. Por ejemplo, en el campo de las tecnologías del habla, el corpus es una base de datos de archivos de audio de habla y/o transcripciones de texto. Los *Corpus de Habla (Speech Corpus)* son utilizados, entre otras cosas, para crear modelos acústicos, los cuales pueden ser utilizados por un mecanismo de reconocimiento de habla. En lingüística, los corpora de habla son utilizados para hacer investigaciones en la fonética, análisis de conversación, dialectología, entre otras [1].

Así mismo, en el campo de las tecnologías del habla también se encuentra la investigación de los sistemas de reconocimiento multi-idioma, la cual es normalmente reforzada por su disponibilidad de dominio público [6].

Lo anterior, permite introducirnos al área de la *Identificación Automática del Lenguaje Hablado (Automatic Language Identification, LID)*, la cual es la tarea de determinar el idioma de un lenguaje hablado a partir de una muestra de audio. En vista de las tendencias actuales de globalización en las tecnologías de la comunicación, la LID juega un rol importante al proveer aplicaciones de habla a una gran comunidad multi-lingüista de usuarios [3].

La intención del corpus desarrollado en este proyecto, es que se utilice sobre las lenguas indígenas mexicanas, específicamente las del estado de *San Luis Potosí* (debido a la cercanía

¹ “(De or. latino). m. Conjunto lo más extenso y ordenado posible de datos ó textos científicos, literarios, etc., que pueden servir de base a una investigación”, Diccionario de la Real Academia de la lengua española.

geográfica), con el fin de estudiarlas e identificar, evaluar y comparar sus idiomas con algoritmos de reconocimiento.

En el caso particular de *México*, las condiciones actuales de los pueblos indígenas han provocado una gran migración de sus poblaciones a las ciudades industriales [4], observándose lamentables casos de injusticia hacia estos hablantes monolingües por carecer de un intérprete [5]. Estos casos han generado la necesidad de contar con aplicaciones que permitan a los usuarios, un medio para identificar el idioma de los habitantes indígenas monolingües, ya sea por alguno de los casos mencionados, alguna necesidad no contemplada o eventos de emergencia.

Desafortunadamente, el Estado del Arte revela que el desarrollo de corpora de habla en *México* es un tema bastante limitado, lo que vuelve a esta tesis una investigación pionera en el campo debido a que es la primera en desarrollar un corpus de habla para las lenguas indígenas mexicanas, el cual se llamó *Entendámonos*.

Siendo *Entendámonos* el primer corpus de habla estandarizado que trata las lenguas indígenas mexicanas. Es importante expresar que el trabajo realizado en esta investigación limita su uso a la tarea de la LID. Además, el método LID utilizado para evaluar este corpus tiene un enfoque acústico debido a que *Entendámonos* carece todavía de una gran cantidad de anotaciones.

Las lenguas indígenas contenidas en el corpus *Entendámonos* son el *tének*, *náhuatl* y *xi'iuy*, las cuales son habladas en el ya mencionado estado de San Luis Potosí. Las grabaciones de este corpus fueron recolectadas en los pueblos donde se hablan una o varias lenguas indígenas, lo que hace que *Entendámonos* posea únicamente grabaciones de hablantes nativos, por así llamarles (ver Sección 4.3), de su respectivo idioma. Otra característica que vale la pena mencionar es la calidad de las grabaciones, la cual es de *44.1 kHz*, y que puede ser remuestreada debido a que el corpus provee los archivos *raw* de cada grabación.

Como ya se mencionó, esta investigación se considera pionera debido al desarrollo de *Entendámonos*, sin embargo, no es la primera en aplicar la LID a las lenguas indígenas, siendo [11] la primera en realizarlo. El detalle es que [11] utiliza muestras del proyecto *AILLA* [12] el cual carece de hablantes por idioma, por lo que los resultados de [11] aplicados a las lenguas indígenas no se consideran fiables en esta investigación.

Para determinar la calidad de un corpus de habla, éste tiene que pasar por severos procesos de evaluación respecto a las especificaciones para las que fue desarrollado. El proceso de evaluación que se realizó sobre *Entendámonos* consiste en la implementación de un sistema LID con un enfoque acústico debido a la carencia de sus transcripciones. Este sistema está basado en la metodología de [8] y parte de la hipótesis en donde las bajas frecuencias poseen características para representar la prosodia². Para realizar lo anterior se utiliza la *transformada wavelet* para diferenciar las altas y bajas frecuencias de la señal de habla y

² En términos de acústica, la prosodia consiste en ritmo, entonación y acentuación del habla.

Sección 1.1 Definición del Problema

posteriormente, se realiza un análisis estadístico sobre las bajas frecuencias para construir modelos acústicos y realizar la tarea del LID.

1.1 Definición del Problema

Dado un sistema LID, se requiere un corpus de lenguas indígenas estandarizado, con el fin de distinguir sus idiomas, sintetizando la problemática en: ¿Cómo estructurar dicho corpus? y ¿Qué pruebas o evaluaciones hacerle?, para que el sistema LID pueda identificar el idioma de una señal de habla.

1.2 Objetivo

Desarrollar un corpus de lenguas indígenas, el cual esté conformado principalmente de muestras de audio, cuya finalidad es que sea evaluado por un sistema LID.

1.2.1 Objetivos específicos

- Definir el número de idiomas que conformarán al corpus.
- Establecer un protocolo con el cual se recolectarán las muestras del corpus.
- Realizar un proceso de estandarización para el corpus.
- Realizar pruebas con un sistema LID tomando como base el corpus de lenguas indígenas.

1.3 Justificación



Figura 1.1. Organismos públicos de México que trabajan para preservar la cultura indígena mexicana.

Cuando uno lee las noticias, es casi una certeza, al menos en México, que la mayoría de nosotros nos hemos encontrado con artículos periodísticos cuyo contenido redacta algún tipo

de trato con personas monolingües, cuyo idioma se caracteriza por formar parte de alguna de las variantes de lenguas indígenas mexicanas, algunos ejemplos de dichos artículos son: “*Hay 8 mil indígenas presos que carecen de un intérprete*” [4] o “*Indígenas en zonas metropolitanas*” [5]. Desafortunadamente, en muchos casos, uno de los problemas comunes en ese tipo de artículos es siempre el mismo: El conflicto que existe entre las personas de habla indígena y las demás para poder entenderse.

Existen organismos públicos (Figura 1.2) como el *INALI*³ y la *CDI*⁴, así como también centros de investigación como el *IELIIP*⁵, los cuales tienen como objetivo común, el promover el fortalecimiento, preservación y desarrollo de las lenguas indígenas que se hablan en el territorio nacional. Personas como el antropólogo *Arnulfo Embriz Osorio* (*Director de Políticas Lingüísticas* del INALI), menciona explícitamente la necesidad de un corpus de lenguas indígenas así como de un sistema LID para poder facilitar asistencia a los hablantes de las lenguas indígenas mexicanas. Además, un corpus de lenguas indígenas ayudaría a preservar dichas lenguas gracias al contenido lingüístico de cada una de sus muestras.

Un problema muy habitual que se presenta en las instituciones, hospitales o lugares que presten algún servicio a estos hablantes monolingües, no es el que los demás no entiendan lo que alguno de estos hablantes quiera decir, sino que no saben siquiera qué idioma es para poder buscar a una persona que ayude a traducir o interactuar. En estas situaciones el personal sabe que debe de llamar a organismos como el INALI que tienen personal o los medios para localizar a un intérprete que hable dicho idioma y ayude con la labor de traducción o interacción, el detalle es que este problema se resuelve a *prueba y error*, si es que se logra identificar el idioma.

1.4 Alcances, Limitaciones y Trabajo Realizado

A opinión propia, una de las ventajas en trabajar en un proyecto de las *ciencias de la computación* es la carencia de aspectos o factores sociales que la mayoría de estos proyectos gozan dentro de la rama.

Desafortunadamente para mí, el desarrollo de un corpus de habla no pertenece a esa mayoría. Debido a los aspectos sociales con los que uno se tiene que enfrentar al desarrollar un corpus de habla, el ritmo de trabajo varía en cada día o actividad debido a que uno no puede tener control sobre estos factores, por ejemplo, uno puede dedicar un día entero a recolectar muestras sin éxito debido a que nunca se encontró a una persona que quisiera colaborar. Son estos aspectos y otros factores lo que hace que uno plantee los alcances al comienzo de un proyecto. Y son esos mismos los que incomodaron el desarrollo del proyecto, perjudicando

³ Instituto Nacional de Lenguas Indígenas.

⁴ Comisión Nacional para el Desarrollo de los Pueblo Indígenas.

⁵ Instituto Estatal de las Lenguas Indígenas e Investigaciones Pedagógicas.

la consumación de los alcances planteados y permitiendo contrastar así el trabajo realizado, o bien, lo que se logró.

Así mismo, me es conveniente empezar listando las limitaciones con las que me tocó tratar, y que considero importantes de señalar a pesar de que varias de éstas carezcan de motivos meramente científicos.

1.4.1 Limitaciones

A continuación se exponen varias limitaciones que se presentaron en el desarrollo del proyecto. Sin embargo, uno también podría considerar a varias de éstas como recomendaciones a seguir para el desarrollo de un corpus de habla de lenguas indígenas.

- En la localidad (*Ciudad Madero, Tamaulipas*) es muy difícil encontrar a alguna persona que hable una lengua indígena, por lo que fue necesario trasladarse a diversos lugares del estado de *San Luis Potosí* para poder grabar a dichas personas.
- En muchas ocasiones, las personas no aceptaron donar su voz, por lo que es necesario tener suma paciencia.
- Como caso particular, la localidad más cercana donde se habla una lengua indígena se encuentra a mínimo *3 horas* de viaje.
- Se debe considerar el costo y tiempo de los viajes cada vez que uno vaya a grabar. Y más importante aún, saber cómo llegar.
- Saber viajar con lo mínimo.
- Se debe tener cierta facilidad para poder comunicarte por otros medios diferentes al uso del idioma español ya que en muchos casos uno se encuentra con gente que sólo puede hablar su lengua indígena y uno carece de un intérprete.
- Se debe contar con una condición física digna de poder caminar algunas horas sobre terrenos ásperos y con el peso extra del equipo de grabación.
- Contar con un carácter que sepa trabajar a pesar de encontrarse con situaciones de pobreza extrema.
- No siempre es obligatorio contar con alguna persona que sepa hablar una de las lenguas indígenas, sin embargo si se puede conseguir a alguien que sepa, mejor.
- En algunas comunidades es recomendable ir acompañado de alguna persona que sea conocida en el pueblo o comunidad para evitar situaciones adversas.
- Saber trabajar en lugares que carezcan con los servicios básicos.
- Se debe considerar el hecho de que en algunas comunidades no habrá forma de conseguir algo de comer.
- Hay comunidades a las que sólo se puede llegar si se cuenta con un vehículo que pueda transitar sobre cierto tipo de terreno (terracería).

1.4.2 Alcances

- Desarrollar un corpus de habla cuyas características sean:
 - ▶ 3 idiomas.
 - ▶ 50 muestras de audio por cada tipo de habla.
 - ▶ Que maneje el tipo de habla: Habla Espontanea (*Spontaneous/Free Speech*).
 - ▶ Que maneje un dictado característico.
 - ▶ Misma cantidad de hombres y mujeres entrevistados.
- Realizar una evaluación al corpus con la metodología de la LID usada en [8], [9] o [10].

1.4.3 Trabajo Realizado

La siguiente lista es un resumen de todas las actividades realizadas en el proyecto, en las Secciones 4 y 5 se explica con más detalle todas las características del corpus y las pruebas LID hechas.

- Corpus Entendámonos.
 - ▶ 3 idiomas.
 - ▶ Tipos de habla manejados: habla específica, habla descriptiva, habla libre.
 - ▶ 46 personas del idioma náhuatl, 50 del idioma tének y 47 del idioma xi'iuy.
 - ▶ 77 voces de mujer, 66 de hombre.
 - ▶ 143 relatos, 37 cuestionarios y dictados.
 - ▶ Juntando todas las grabaciones se sobrepasa las 3hrs de contenido hablado.
- Evaluaciones al corpus *Entendámonos* con un sistema lid basado en [8] y [10].

1.5 Estructura de la Tesis

Capítulo 2. Praeludium⁶. Este capítulo brinda una ligera introducción de los conocimientos fundamentales que se utilizan en este proyecto.

Capítulo 3. Estado del Arte. Este capítulo brinda un panorama actual de varias investigaciones relacionadas con el tema del desarrollo de corpora de habla, lenguas indígenas mexicanas e identificación automática del lenguaje hablado.

Capítulo 4. Corpus Entendámonos. Este capítulo explica de manera detallada los procesos de desarrollo por lo que paso el corpus de esta investigación.

⁶ Palabra en latín para *preludio*, que significa: *Aquello que precede y sirve de entrada, preparación o principio a algo.*

Sección 1.5 Estructura de la Tesis

Capítulo 5. Evaluaciones. Este capítulo presenta las pruebas que se realizaron con el sistema LID, utilizando el corpus *Entendámonos*.

Capítulo 6. Occāsus⁷. Este capítulo contiene las secciones finales de toda tesis: las conclusiones, las aportaciones de la investigación y el trabajo futuro.

Apéndice A. Marco Teórico. Este apéndice contiene las bases teóricas relacionadas al presente trabajo.

Apéndice B. Equipo. Este apéndice contiene las especificaciones técnicas de los ordenadores usados en el proyecto, así como del micrófono con el que se desarrolló *Entendámonos*.

Bibliografía.

⁷ Palabra en latín para *ocaso*.

2

Praeludium

Este capítulo aduce los conceptos y nociones fundamentales que el proyecto abarca, ya que debido a su naturaleza son varios. Los conceptos y tópicos de este capítulo se manejan en un modo ligero, evitando ser ambicioso con su descripción, pero siendo lo suficientemente justo para que éstos asistan a los capítulos siguientes de esta tesis. Algunos de los tópicos de este capítulo son tratados de una manera más profunda en el Apéndice A.

Primero se empieza listando varios de los conceptos básicos manejados a lo largo de esta tesis, seguido se los antecedentes de los *corpus de habla* y de la *identificación automática del lenguaje hablado*. Así mismo se exponen temas sobre las lenguas indígenas de México y se proporciona información de las 3 lenguas abarcadas en este proyecto.

2.1 Preliminares

El objetivo de *Entendámonos*, hasta el momento, es para que se utilice en la LID. Pero ¿Qué es eso y de dónde viene? Además, ¿Corpus de habla? ¿Lenguas indígenas?

Hasta el momento se han mencionado términos como corpus de habla, LID, corpus de habla estandarizado, modelos acústicos, entre otros. Sin embargo, al adentrarse al proyecto uno encuentra varios tópicos que son ampliamente manejados en diferentes áreas. Por ejemplo, la transformada wavelet se maneja en temas que van desde el procesamiento digital de la voz, hasta la ingeniería de temblores, vientos y océanos [13], por mencionar algunos.

2.1.1 Corpus de Habla

En México, los corpus de habla son un tema esencial en las áreas del *Procesamiento Digital de la Voz (Speech Processing)* y *Lingüística Computacional*.

Sección 2.1 Preliminares

El término *corpus de habla* hace referencia a colecciones de grabaciones de habla digitales junto con (aunque no forzosamente) anotaciones, metadatos y documentación. Los corpora 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 (Figura 2.1), por ejemplo, Reconocimiento del habla (*Speech recognition*), Conversores texto-voz (*Text-to-Speech*), Síntesis de habla (*Speech synthesis*), Reconocimiento de locutores (*Speaker Recognition*), etcétera.



Figura 2.1. Analogía: El mundo como un corpus de habla dividiéndose en las diferentes áreas del procesamiento digital de la voz.

Entre los aspectos más comunes para el desarrollo de los corpora de habla se encuentran:

- Especificación,
- Preparación de la colección,
- Colección, (que en la mayoría de los casos superpuesta por)
- Post-procesamiento,
- Anotaciones,
- Documentación,
- Validación.

Asimismo, otro aspecto importante a mencionar, aunque no forma parte de la fase de desarrollo, es la evaluación, la cual se encarga de determinar la calidad del corpus y consiste en el uso del corpus para una tarea específica, por ejemplo, las pruebas LID hechas con el corpus *Entendámonos*.

Finalmente, la siguiente lista define algunos de los términos técnicos considerados importantes en el desarrollo de los corpora de habla.

1. Corpus de habla (*Speech Corpus*).- 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. A partir de ahora, en este documento la palabras corpus o corpora se refieren específicamente a esta definición.
2. Validación (*Validation*).- El control/inspección (formal) de un corpus de habla con respecto a sus especificaciones predefinidas.
3. Especificación (*Specification*).- La descripción técnica fija de un corpus de habla con respecto a sus características (incluyendo anotaciones, metadatos y documentación).
4. Formato de archivo (*File format*).- Formato estandarizado o específico de una señal digital y datos simbólicos (anotaciones, metadatos).
5. Anotación (*Annotation*).- Descripción discreta (categorizada) asociada con una señal física (codificada). Usualmente consiste de un conjunto cerrado de símbolos y un esquema para enlazar estos símbolos a cualquier punto en el tiempo o segmentos de tiempo.
6. Dominio (*Domain*).- Tópicos de comunicación verbal o situación en la cual una comunicación verbal toma lugar.
7. Guion (*Prompt*).- Documento de habla (palabra, frase u oración) presentada al hablante. Una lista de guiones o un corpus de guiones es una colección de guiones que define el contenido hablado del corpus.
8. Contenido hablado (*Spoken content*).- Qué es lo que fue dicho en un corpus de habla.
9. Metadatos (*Metadata*).- Información acerca de datos. Normalmente el término está restringido a tres tipos: protocolos de grabación, comentarios y perfiles de los hablantes.
10. Evaluación (*Evaluation*).- Una evaluación cualitativa de un corpus respecto a su usabilidad en cierta tarea o escenario de desarrollo.

2.1.2 Identificación Automática del Lenguaje Hablado (LID)

La LID es la tarea de identificar automáticamente el lenguaje, idioma o dialecto que es hablado por un hablante humano a partir de muestra de habla [14]. Con respecto a la actual necesidad global de interfaces humano-computadora multilingües [3], la LID juega un rol esencial en proveer aplicaciones de habla. Durante las últimas 3 décadas, la LID se ha vuelto una tecnología clave en áreas del procesamiento del lenguaje hablado tales como los sistemas de reconocimiento de habla-entendimiento, sistemas de diálogo hablado, sistemas de comunicación humano-a-humano, recuperación de documentos hablados, y sistemas de minería multimedia [14].

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.

Sección 2.1 Preliminares

Hay una variedad de características que los humanos y las computadoras usan para discriminar un idioma de otro. En la Figura 2.2 se pueden apreciar varios niveles de abstracción por medio de un análisis acústico acompañado con componentes de información [3].

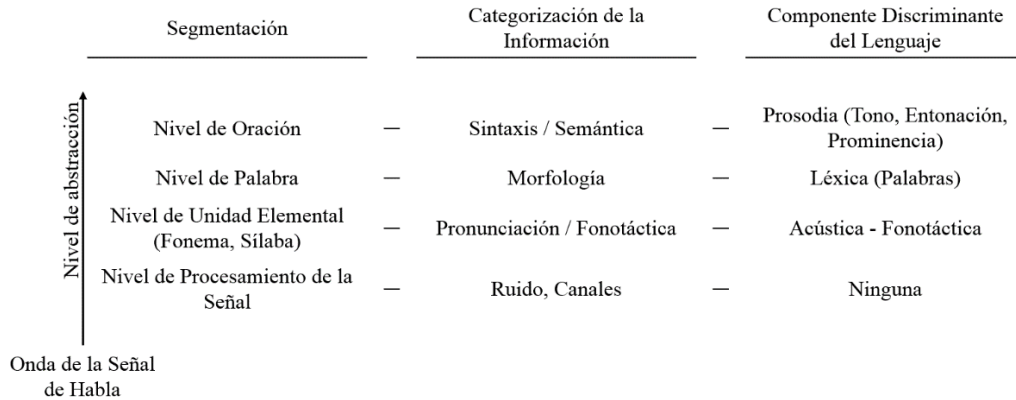


Figura 2.2. Niveles de abstracción de la señal por medio de un análisis acústico acompañado con componentes de información.

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, y a pesar de que varios idiomas comparten algunos fonemas, esta es la característica más popular como método LID.

Otra característica importante utilizada 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 uno 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 la cual establece que en las bajas frecuencias hay información sobre las características suprasegmentales⁸ que utilizamos al hablar. Esta hipótesis fue presentada en [11] (Ver Sección 3.5) y ha sido causa de varios trabajos de investigación [8-10].

Los sistemas LID están basados en las propiedades lingüísticas extraídas de la o las señales de habla entrantes. El rendimiento de un sistema LID recae en la cantidad y confiabilidad de la información discriminativa y en cómo está incorporada para que dicha información sea eficiente.

⁸ También llamadas *prosódicas*, son aquellas características que afectan a un segmento más largo que el fonema, tales como el acento, la entonación, el ritmo, la duración, entre otras. El término suprasegmental implica la existencia de elementos que recaen sobre más de un segmento a la vez.

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 (normalmente referidas en inglés como *acoustic-phonetics*) y sus secuencias (normalmente referidas en inglés como *phonotactics*). En adición a la información espectral, algunos sistemas LID también incorporan información prosódica. La Figura 2.3 muestra la arquitectura general de un sistema LID basado en diferentes enfoques.

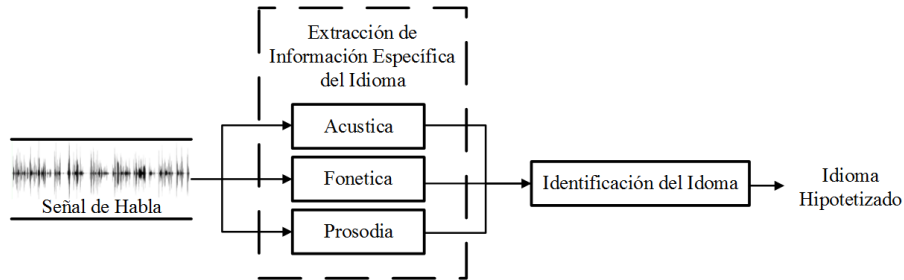


Figura 2.3. Arquitectura general de un sistema LID usando diferente información discriminativa.

Cada tipo de información usada por el sistema LID no goza únicamente de ventajas, sino también desventajas. Por ejemplo, la información espectral, *acoustic-phonetics* y *phonotactics*, son fáciles de obtener ya que usualmente vienen predefinidas en muchas corpora y por ende se vuelve un buen trato entre la complejidad computacional y el rendimiento, ya que podemos elegir que unidades usar o cantidad de información que vayamos a utilizar, pero al mismo tiempo, el rendimiento del sistema LID basado en la información espectral normalmente se degrada debido al ruido o a las condiciones acústicas de las muestras. La propiedades prosódicas son menos afectadas por el ruido pero la confiabilidad de esta información todavía se está investigando. La prosodia de una muestra está influenciada por las características específicas del hablante, tales como su tipo de voz, ritmo de habla, estado emocional, etcétera, así como también por el contexto sintáctico de lo dicho en la muestra, por ejemplo, que sea una declaración, una pregunta, una exclamación, etc. Para explotar satisfactoriamente la información prosódica, las características dependientes del idioma deben ser separadas de los componentes independientes del lenguaje.

A manera de usar diferentes tipos de información para discriminar los idiomas, se debe considerar un balance entre la eficiencia y precisión del sistema. Diseñar sistemas LID depende de la tarea concreta de la LID, de los métodos que se vayan a utilizar y de la disponibilidad de una cantidad suficiente de información para la fase de entrenamiento. Algunos sistemas requieren solo habla digitalizada y sus correspondientes etiquetas sobre el idioma hablado. Otros sistemas requieren de la existencia de las transcripciones de los fonemas y palabras de lo hablado de cada muestra para la fase de entrenamiento. En la Figura 2.4 se puede ver la arquitectura general de un sistema LID.

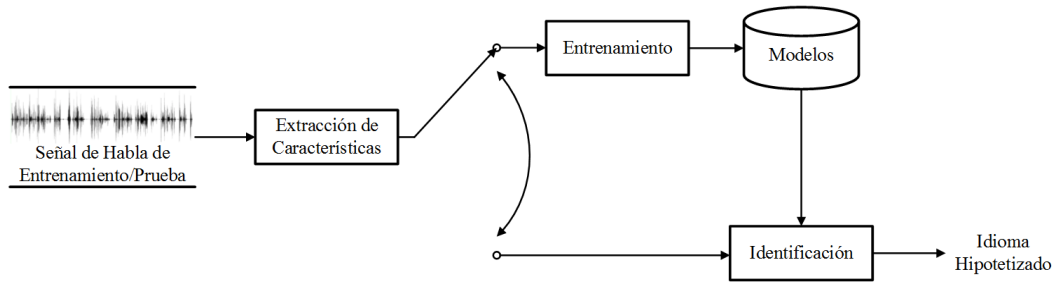


Figura 2.4. Arquitectura general de un sistema LID.

De cualquier forma, lo ideal en un sistema LID es que tenga una alta precisión en la identificación del idioma y al mismo tiempo ser:

- Computacionalmente eficiente.
- Robusto ante diferentes casos de prueba como el ruido, hablante, vocabulario del hablante, etc.
- Simple en el sentido de su requerimiento de información específica de los idiomas.
- Actualizable, para que sea relativamente sencillo poder añadir algún nuevo idioma.

2.2 Antecedentes

El primer trabajo que trata el desarrollo de un corpus de habla se le atribuye a *Yeshwant Kumar Muthusamy* [6], cabe mencionar que existían *colecciones de información de habla* previas al trabajo de Muthusamy, sin embargo, éstas no contaban con las características de un corpus de habla. Por ejemplo, en [6] se expone un resumen del trabajo de investigación desarrollado por el laboratorio de *Texas Instruments* entre 1973 y 1980, éste ya utilizaba grabaciones de habla (específicamente grabaciones de sólo hombres adultos leyendo algún texto) para extraer características de 7 idiomas, sin embargo, jamás se reportaron las características de dichas grabaciones, ni los lenguajes que fueron analizados, ya que en dicha investigación los lenguajes estaban identificados como L_1, L_2, \dots, L_7 .

La investigación en el área de LID requiere de enormes cantidades de información de habla de varios idiomas. Desafortunadamente, cuando Muthusamy empezó su investigación, no existía algún corpus de habla con tales características, por lo que él tuvo que dedicar una cantidad considerable de tiempo y esfuerzo para recolectar y desarrollar los primeros dos corpus de habla, los cuales fueron el *OGI_HQ* y el *OGI_TS*.

El corpus *OGI_HQ* (*Four-language High-quality Speech Corpus*) consistió en un corpus con muestras de alta calidad, desarrollado en un ambiente controlado y sus idiomas fueron 4. Su objetivo era examinar los detalles y problemas relacionados a la LID sobre un conjunto pequeño de muestras, antes de que empezara a trabajar con un conjunto más grande de idiomas. Los idiomas incluidos en este corpus fueron el inglés, japonés, mandarín y tamil.

Para realizar las grabaciones, Muthusamy utilizó un micrófono *Sennheiser HMD 224* y las muestras tuvieron una frecuencia de muestreo de *16 kHz*. Además, a las personas que aportaron su voz se les pidió decir *20* diferentes tipos de sentencias y se les pagó *5 U.S. dollars* por su participación. La Tabla 2.1 muestra las características sobresalientes de este corpus.

El motivo para desarrollar el corpus OGI_TS (*Ten-language Telephone Speech Corpus*) se debió a varios problemas con la adquisición de las muestras del corpus OGI_HQ: 1) Su proceso de recolección de muestras fue demasiado lento y requería una cantidad considerable de supervisión al momento de grabar. 2) La adición de nuevos idiomas al corpus estaba altamente sujeta a la disponibilidad de los hablantes nativos de su respectivo idioma respecto y a la cercanía de cada uno de ellos al laboratorio de grabación. 3) El uso de muestras grabadas en un ambiente controlado para el entrenamiento de un sistema LID hace que éste sea limitado.

Tabla 2.1. Características sobresalientes del corpus OGI_HQ.

Idioma	# de Hablantes	#♂ / #♀	# de Muestras	Rango del # de S/M	Media del # de S/M	Duración de la Muestra
Inglés	20	10 / 10	400	15 – 151	59.1	4.6
Japonés	25	12 / 13	453	13 – 153	68.2	6.1
Mandarín	24	14 / 10	470	11 - 130	51.1	5.0
Tamil	20	11 / 09	398	16 - 155	71.4	5.8

Para que un sistema LID funcione adecuadamente, tiene que estar entrenado con muestras cuyas características sean similares a las de la muestra cuyo idioma se quiere identificar. Es por esto que el corpus OGI_HQ no se desarrolló más. En aquel entonces, la probabilidad más alta en donde un sistema LID se pudiese utilizar residía en medios de comunicación más comunes, por ejemplo, la línea telefónica, la cual estaba caracterizada por un bajo ancho de banda, distorsión en los canales, variabilidad en el micrófono y baja SNR⁹. Fue entonces así que Muthusamy decidió recolectar muestras de habla de varios lenguajes utilizando las líneas de teléfono comerciales.

Muthusamy reporta diversas ventajas en utilizar colecciones de información de habla telefónica sobre habla de alta calidad.

- El proceso de recolección puede ser fácilmente automatizado y necesita el mínimo de supervisión humana.

⁹ La relación señal/ruido (en inglés *Signal to Noise Ratio*, *SNR* o *S/N*) se define como el margen que hay entre la potencia de la señal que se transmite y la potencia del ruido que la corrompe. Este margen es medido en decibelios.

Sección 2.2 Antecedentes

- Las largas distancias de las redes telefónicas proveen acceso a hablantes nativos de diferentes idiomas sobre una amplia área geográfica.
- La adición de nuevos idiomas requiere el contacto de sólo dos hablantes nativos para cada idioma (uno para pregrabar las instrucciones y guiones, y otro para verificarlas) y una pequeña campaña publicitaria antes de empezar la recolección de muestras para cada idioma.

El corpus OGI_TS consistió de 10 idiomas: el Inglés, Farsi (Persa), Francés, Alemán, Coreano, Japonés, Chino (Mandarín), Español, Tamil y Vietnamita. Estos idiomas fueron seleccionados basados en consideraciones lingüísticas y disponibilidad de hablantes nativos en USA.

La adquisición de los datos del corpus, el formato de las llamadas, equipo de grabación, etc., se explicará más adelante en la sección de *Estado del Arte*. La Tabla 2.3 muestra las características estadísticas de los hablantes que conforman el corpus OGI_TS y la Tabla 2.2 nos muestra las distribuciones de sus llamadas originales.

Tabla 2.2. Distribución de las llamadas a través de los 10 idiomas del corpus OGI_TS.

Idioma	Llamadas Originales (Raw)	Llamadas Utilizables	# de M (Muestras)	Media del #M/Llamadas	Media del #Seg/Llamadas
Inglés	1044	868	7991	9.2	89.7
Farsi	154	115	993	8.6	79.6
Francés	149	122	1082	8.9	85.4
Alemán	157	118	1059	9.0	87.7
Japonés	147	107	930	8.7	79.5
Coreano	149	112	905	8.1	71.1
Mandarín	186	141	1103	7.8	66.4
Español	150	128	1150	9.0	86.0
Tamil	194	149	1189	8.0	67.8
Vietnamita	159	127	1023	8.1	69.8

Aparte de haber sido la primer persona que formalmente desarrolló un corpus de habla, Muthusamy también desarrolló un sistema LID, el cual utilizó para evaluar sus dos corpus. Las primeras evaluaciones fueron realizadas sobre el corpus OGI_HQ, cuyos resultados se pueden ver en la Tabla 2.4, promediando un 89.5% de identificación.

Basándose en los resultados obtenidos, Muthusamy realizo evaluaciones sobre el corpus OGI_TS. Estos últimos experimentos los hizo utilizando categorías fonéticas en grupos de 2

Capítulo 2. Praeludium

y 3 fonemas, usando características espectrales (PLP) y características basadas en el tono. Además exploró la posibilidad de utilizar segmentos largos de habla para la identificación del idioma. Todas las pruebas que él realizó fueron hechas sobre muestras cortas de habla y sobre las muestras tipo *story*, las cuales son muestras de habla que relatan una historia cualquiera. Su sistema LID entrenado con todas las características mencionadas anteriormente obtuvo rangos de precisión que variaban entre el 47.3% y 91.9% para cada idioma, siendo el inglés el idioma mejor identificado. Estos resultados se pueden observar en la Figura 2.5.

Tabla 2.3. Estadísticas de los hablantes del corpus OGI_TS.

Idioma	Llamadas	Hombres	Mujeres	Género Desconocido	Adultos	Niños
Inglés	868	595	269	4	853	15
Farsi	115	92	23	0	114	0
Francés	122	85	37	0	122	0
Alemán	118	74	44	0	118	0
Japonés	107	69	38	0	107	0
Coreano	112	84	28	0	111	1
Mandarín	141	89	42	0	141	0
Español	128	83	44	1	125	3
Tamil	149	127	21	1	149	0
Vietnamita	127	81	46	0	127	0
Total	1987	1379	602	6	1968	19

Tabla 2.4. Matriz de confusión y promedio de identificación sobre la evaluación del corpus OGI_HQ.

Idioma	Inglés	Japonés	Mandarín	Tamil	
Inglés	40	0	1	1	95.2%
Japonés	6	28	0	0	82.4%
Mandarín	3	3	35	1	83.3%
Tamil	0	1	0	34	97.1%

A pesar de todo el trabajo realizado por Muthusamy, él no fue la primer persona que realizo trabajo de experimentación en el área del LID. En [2] se reportan una serie de trabajos con el

Sección 2.3 Lenguas Indígenas

fin de identificar el idioma, sin embargo, varios de estos trabajos no lo hacen basándose en grabaciones de voz o archivos de audio como hoy en día.

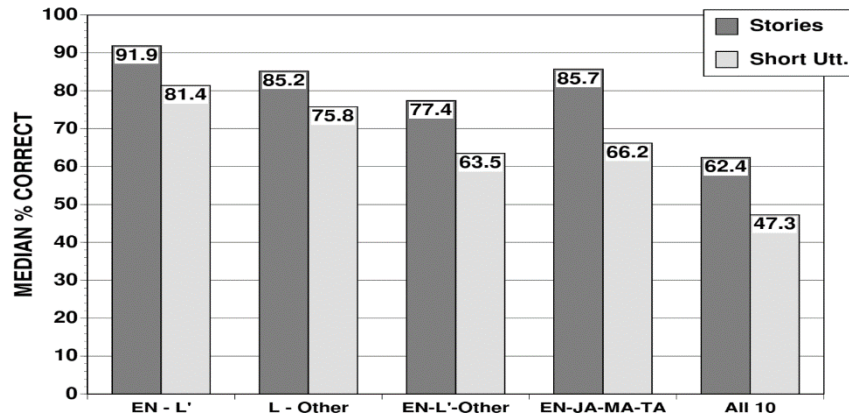


Figura 2.5. Resumen de los resultados las pruebas LID hechas sobre el corpus OGI_TS.

Entre los trabajos mencionados en [2], destacan los trabajos de investigación hechos por el laboratorio de Texas Instruments (mencionado al principio de esta sección) cuya investigación de la LID expuso un promedio del 80% de identificación usando 5 lenguajes, además es importante señalar que estos estudios dieron una importante noción acerca del uso de fonemas como característica distintiva para cada idioma.

Otra investigación que vale la pena mencionar es la que realizaron en [7] en donde se desarrolló un sistema LID que no estaba basado en pequeños segmentos que representaban fonemas o sílabas. Este enfoque fue aplicado a técnicas de análisis de patrones en base a característica acústicas extraídas de una señal de habla. La evaluación de este sistema fue aplicada a 8 lenguajes los cuales fueron: inglés, checo, farsi, alemán, coreano, mandarín, ruso y vietnamita. Los resultados de identificación tenían un promedio de 84%, siendo el idioma mejor identificado el Coreano (93.4%) y el peor el Inglés (76.8%). El problema de este sistema fue el escaso número de hablantes (sólo 5 hombres adultos) que formaron parte del conjunto de entrenamiento, por lo que años después en [2] se concluyó que el sistema no era independiente si se utilizaba fuera de las muestras utilizadas para el conjunto de entrenamiento. Sin embargo esta investigación fue la primera en mostrar que existía la posibilidad de poder crear un sistema LID usando únicamente características acústicas.

2.3 Lenguas Indígenas

Las poblaciones indígenas tienen usos y costumbres propias. Poseen formas particulares de comprender el mundo y de interactuar con él. Visten, comen, celebran sus festividades, conviven y nombran a sus propias autoridades de acuerdo a esa concepción que tienen de la

vida. Un elemento muy importante que los distingue y les da identidad, es la lengua con la que se comunican.

En México, 6,695,228 personas de 5 años y más hablan alguna lengua indígena, las más habladas son: *Náhuatl, Maya y lenguas mixtecas* (Tabla 2.5). A nivel nacional, 6 de cada 100 habitantes, de 5 años o más, hablan alguna lengua indígena, de las cuales existen 89. Se considera únicamente a los habitantes de 5 años y más, porque a partir de esta edad una persona puede tener dominio de una lengua para comunicarse [15].

Tabla 2.5. Las diez lenguas indígenas más habladas en México [15].

Lengua	Número de Hablantes
Náhuatl	1,659,029
Maya	892,723
Mixteco	510,801
Zapoteco	505,992
Tzotzil	356,349
Tzeltal	336,448
Mazahua	151,897
Purépecha	136,388
Mixe	135,316
Mayo	34,770

Gran parte de la población indígena de México, además de hablar alguna lengua indígena, también habla español. Sin embargo, hay cerca de un millón de indígenas monolingües que no hablan español.

Además, investigaciones realizadas hasta el presente, así como consultas y estudios propios realizados por el INALI, han demostrado que la realidad lingüística del país es mucho más compleja de lo que en términos generales se ha creído hasta ahora.

El uso que se le ha dado al concepto lengua en torno a la diversidad lingüística mexicana ha resultado impreciso; por ejemplo, a partir de la época virreinal (quizá desde antes), se difunde la creencia de que los pueblos indígenas hablan "una sola lengua" (altamente uniforme en todos sus componentes), sin advertir, la mayoría 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. Ante este panorama, el INALI catalogó la diversidad lingüística de los pueblos indígenas en México a partir de las siguientes tres categorías, relacionadas de mayor a menor grado de inclusión:

Sección 2.3 Lenguas Indígenas

- Familia lingüística
 - ▶ Agrupación lingüística
 - ▶ Variante lingüística

Considerando información tanto de las estructuras lingüísticas como de carácter sociolingüística, se definieron las categorías rectoras para el proceso de elaboración del “*Catálogo de las Lenguas Indígenas Nacionales, Variantes Lingüísticas de México con sus autodenominaciones y referencias geoestadísticas*” [16], el cual es representa un trabajo exhaustivo que enlista, de acuerdo a las 3 categorías mencionadas, todas las lenguas indígenas habladas en México. En total, [16] enlista 364 lenguas considerando las variantes lingüísticas y 68 lenguas considerando la agrupación lingüística, ambas perteneciendo a una de las 11 familias lingüísticas.

La *Familia lingüística* es la categoría más inclusiva de los niveles de catalogación aplicados por el INALI. Se define como un conjunto de lenguas cuyas semejanzas estructurales y léxicas se deben a un origen histórico común. Once son las familias lingüísticas indoamericanas consideradas en razón de que cada una de ellas se encuentra representada en México con al menos una de sus lenguas. En la Figura 2.6 se aprecia cada una de ellas.



Figura 2.6. Las 11 familias lingüísticas indoamericanas con presencia en México. INALI.

La *Agrupación lingüística* ocupa el lugar intermedio en los niveles de catalogación aplicados por el INALI. Se define como el conjunto de variantes lingüísticas comprendidas bajo el nombre dado históricamente a un pueblo indígena. De acuerdo con esta definición, las

Capítulo 2. Praeludium

agrupaciones lingüísticas aquí catalogadas se encuentran relacionadas, respectivamente, con un pueblo indígena y pueden estar conformadas por conjuntos de una o más variantes lingüísticas. En las Figuras 2.7 y 2.8 se pueden apreciar las 68 lenguas según su agrupación.

La *Variante lingüística* es la categoría que alcanza el mayor grado de detalle de los niveles de catalogación aplicados por el INALI. Se define como una forma de habla que presenta diferencias estructurales y léxicas en comparación con otras variantes de la misma agrupación lingüística; e implica para sus usuarios una determinada identidad sociolingüística, que se diferencia de la identidad sociolingüística de los usuarios de otras variantes. Esta categoría es comúnmente empleada por la población hablante de lengua indígena, en particular por la que es bilingüe lengua indígena-español, para hacer referencia, precisamente, a formas de hablar que contrastan, en mayor o menor medida, en los planos estructural, léxico y/o sociolingüístico, entre comunidades o regiones asociadas con un mismo pueblo indígena.



Figura 2.7. 7 agrupaciones lingüísticas de las 11 familias. INALI.

Las variantes lingüísticas representan uno de los indicadores más fehacientes de la enorme diversidad lingüística y cultural de México. De conformidad con el estado que guardan los estudios sobre la realidad lingüística de nuestro país y con el propósito de evitar la discriminación lingüística, el INALI considera que las variantes lingüísticas deben ser tratadas como lenguas, al menos en las áreas educativas, de la impartición y la administración

Sección 2.3 Lenguas Indígenas

de justicia, de la salud, así como en los asuntos o trámites de carácter público y en el acceso pleno a la gestión, servicios e información pública.



Figura 2.8. 4 agrupaciones lingüísticas de las 11 familias. INALI.

De cada variante han sido consignados dos elementos:

- Su autodenominación, es decir, la expresión con la cual los hablantes de lenguas indígenas nombran a éstas en su propia variante lingüística.
- Su referencia geostadística, esto es, las localidades, municipios y entidades federativas en donde se habla cada una de ellas.

Las siguientes 3 secciones muestran información muy general de las lenguas indígenas náhuatl, tének y xi'iuy. La información mostrada fue recabada del IELIIP y del trabajo [16] del INALI.

2.3.1 Náhuatl

- Familia lingüística: Yuto-Nahua
- Variante lingüística: Náhuatl de la Huasteca Potosina [nawalλ]
- Forma de escritura según el IELIIP: Nawatl

Capítulo 2. Praeludium

Tabla 2.6. *Machiopamitl* (abecedario) de la lengua náhuatl propuesto por el IELIIP. *Wexmachiotl* (mayúsculas) a la izquierda y *silmachiotl* (minúsculas) a la derecha en cada recuadro.

A Ayojtli	a	Ch Chapolin	ch	E Elotl	e	I Ixteyolli	i
J Ijwitl	j	K Konetl	k	KU Kuawtli	ku	L Lalax	l
M Miston	m	N Nanakatl	n	P Papalotl	p	O Okuilin	o
S Sitlalin	s	T Tepetl	t	TI Tlankonchtli	tl	Ts Tsitsitl	ts
W Wexolotl	w	X Xochitl	x	Y Yolotl	y		

2.3.2 Tének

- Familia lingüística: Maya
- Variante lingüística: Teenek/Tenek (Huasteco) del occidente [teːnek, tenek]
- Forma de escritura según el IELIIP: Tének

Tabla 2.7. *Nik'adh* (abecedario) de la lengua tének propuesto por el IELIIP. *Ok'lab* (mayúsculas) a la izquierda y *ts'uts'lab* (minúsculas) a la derecha en cada recuadro.

A Ajin	a	B Bichim	b	C Chuch	c	CH' Ch'uri'	ch'
DH Dhiniy	dh	E Edhem	e	I Its'amal	i	J Jajnek	j
K Koxte'	k	K' K'wadhap	k'	KW Kwix-tót	kw	K'W K'wa'	k'w
L Lem	l	M Mitsu'	m	N Ni'ni'	n	O Olom	o
P Pik'o'	p	R Ráw	r	S <small>Esta letra sólo existe en palabras que se refieren a sonidos (onomatoyecas)</small>	s	T To'ol	t
T' T'íw	t'	Ts Tsan	Ts	TS' Ts'ili'	ts'	U Udhu'	u
W Wa'juts	w	X Xa'wix	x	Y Yoy	y	' (glotal) <small>Se usa para dar énfasis y entrecortar la palabra.</small>	

2.3.3 Xi'iuy

- Familia lingüística: Oto-Mangue
- Variante lingüística: Pame del centro [ʃiʔoi] y Pame del norte [ʃiʔiwi]
- Forma de escritura según el IELIIP: Xi'iuy Sur y Xi'iuy Norte

Sección 2.3 Lenguas Indígenas

Tabla 2.8. Abecedario de la lengua xi'iuypropuesto por el IELIIP. Mayúsculas a la izquierda y minúsculas a la derecha en cada recuadro.

A Nava'a	a	B Bung	b	CH Chikil'	ch	ND Ndutsu'ul	nd
E Meng	e	NG Ngututs	ng	GY Gyajaut	gy	I Xixi	i
J Jueut	j	K Kase'e	k	L L'jua	l	Ly Lyii	ly
M Manjua	m	N Nilyjañg	n	Ñ Ñuut	ñ	P Pakas	p
R Rimjie	r	S Stilyjañg	s	ST stumje	st	TS tsutue	ts
U kunee	u	V Vareik	v	X Xilyjua	x	Y Kasaily	Y

3

Estado del Arte

A manera de reseña, este capítulo relata algunos de los trabajos relevantes revisados durante el transcurso de este proyecto, en donde la mayoría de ellos están relacionados al desarrollo de corpora, o bien, a la tarea del LID.

Desafortunadamente, el casi nulo uso de las lenguas indígenas mexicanas en las áreas del procesamiento digital de la voz es una lamentable falta por parte de muchos organismos, instituciones o centros de investigación, ya que se desperdicia una buena forma de fomentar su desarrollo con el uso de las tecnologías del habla.

En adición, debido a lo anterior, el encontrar proyectos relacionados a lo que se trabaja en esta tesis fue imposible, lo que indujo al autor a confirmar que realmente estaba desarrollando algo no hecho antes para las lenguas indígenas mexicanas.

3.1 Florian Shiel et al.

Muy a menudo, grandes cantidades de dinero y esfuerzo son gastados sobre corpora de habla de mala calidad, por ejemplo, corpora que sólo sirve para un propósito en particular y que jamás fue pensada en ser compartida. Estas corpora no pueden ser reutilizadas para otros propósitos que aquellos para los cuales fueron diseñadas; además de ser difíciles de actualizar o mantener. Como consecuencia, este tipo de corpora descuida totalmente su valor, ya sea de investigación, comercial, entre otros.

The Production of Speech Corpora es lo que se conoce como un libro tipo *cookbook* desarrollado en el BITS (*BAS Infrastructures for Technical Speech Processing*), ubicado en la universidad de Múnich, Alemania. El libro describe a detalle el desarrollo y producción de un corpus de habla; menciona a las corpora de habla como la fuente principal de información para la investigación en el área de la Comunicación del Lenguaje Hablado (*Spoken Language*

Sección 3.2 Lwazi

Communication), y para el desarrollo de tecnología en el área del Procesamiento Digital de la Voz.

En general, provee prospectiva a los usuarios de las comunidades científicas e ingenieras con consejos sobre cómo producir corpora reusable, consistente y de alta calidad para sus respectivas necesidades y así evitar las corpora no reusables, la cual, en la mayoría de los casos, se origina debido al pobre proceso de producción que no fue monitoreado debidamente. Asimismo, brinda un resumen general de las mejores prácticas en el campo y presenta modelos para algunos casos estándar.

Debido a la calidad del trabajo, este documento fue la principal base teórica para el desarrollo de *Entendámonos* y cuyos conceptos forman parte del Apéndice A; algunos de ellos ya mencionados en la sección 2.1.1.

3.2 Lwazi

El proyecto *Lwazi*¹⁰ [19,20] se enfoca al desarrollo de un sistema de información basado en muestras de habla telefónica. Éste fue desarrollado por el *Meraka Institute* y comisionado por el *Departamento de Arte y Cultura* de la *República de Sudáfrica*.



Figura 3.1. Logo del proyecto Lwazi.

El proyecto intenta proveer a los sudafricanos una oportunidad de acceder a la información y servicios de su gobierno en cualquiera de los once lenguajes oficiales de Sudáfrica, usando ya sea teléfonos fijos o móviles, todo sin costo alguno para el usuario.

Este proyecto comenzó a desarrollarse porque más y más sudafricanos tienen acceso y uso de teléfonos como manera de comunicarse, y al mismo tiempo, no todos los sudafricanos

¹⁰ Palabra que significa *conocimiento* en el idioma *Xhosa*.

tienen acceso a Internet (desde un ordenador fijo o portátil) o a los medios impresos, lo cual hace a los teléfonos una alternativa muy usada para acceder a información importante. Cabe aclarar que los beneficiados de este proyecto son los departamentos del gobierno Sudafricano y sus instituciones relacionadas. Ellos usaron tecnologías desarrolladas durante el proyecto para proveer información y optimizar la "entrega del servicio" a la comunidad Sudafricana.

Los usuarios de Lwazi son todos aquellos que buscan acceder a la información del gobierno de la República de Sudáfrica. Una de las tareas de este proyecto, es la creación de su corpora para poder realizar pruebas en las diferentes áreas de investigación que el proyecto abarca hasta la fecha, las cuales son: *Text-to-Speech* (TTS) y *Speech Recognition*.

La Lwazi corpus abarca los 11 idiomas de la República de Sudáfrica, los cuales son: *IsiZulu*, *IsiXhosa*, *Afrikaans*, *Setswana*, *Sepedi*, *Inglés*, *Sesotho*, *Siswati*, *Xitsonga*, *Tshivenda* e *IsiNdebele*. En la Tabla 3.1 se muestra un resumen de la información general del corpus para Speech Recognition.

Tabla 3.1. Información general de 5 de los 11 idiomas del corpus Lwazi para Speech Recognition [20].

Idioma	Porcentaje de Transcripción	Número de Hablantes	Número de Muestras	Número de Palabras
IsiZulu	97.75%	195	294	1780
IsiXhosa	95.61%	198	302	1801
Afrikaans	99.72%	200	298	1776
Setswana	99.62%	199	297	2609
Sepedi	98.16%	190	295	3149

A pesar de no haber sido desarrollada para la LID, la información expuesta en las publicaciones de Lwazi brindó un panorama muy ameno para el desarrollo de *Entendámonos*. Por ejemplo, en [20] se explica detalladamente el desarrollo de los corpus de Lwazi, empezando por la selección de su vocabulario, sus procesos de verificación, y su protocolo de grabación, además de exponer algunas de las dificultades por las que el desarrollo del corpus tuvo que pasar.

3.3 Jiří Navrátil

El *Dr. Jiří Navrátil* contempla en [3] un acercamiento muy completo al campo de la LID, analizando diferentes soluciones en términos de sus aplicaciones prácticas y concluyendo

con un panorama sobre las direcciones de los temas actuales e investigaciones futuras del problema.

El menciona que el principal problema en la tarea de la LID, es el encontrar una manera de reducir la complejidad del lenguaje humano, de tal forma que un algoritmo pueda determinar la identidad del lenguaje relativamente desde una muestra de audio corta.

A su vez, también hace mención de varios niveles de abstracción desde la señal pura de habla y los tipos de información que típicamente se encuentran en esos niveles (Figura 2.2, expuesta en la sección 2.1.2). Donde en nivel más bajo incluye solamente la variación pura de la señal acústica, determinada por el canal y el ruido de fondo, mientras que los otros niveles proveen información relevante para distinguir a los lenguajes uno de otro.

Finalmente, menciona y detalla algunas de las corpora de habla más populares, las cuales son: OGI-TS (después transformada en la OGI-11L), OGI-22L, CALLHOME, CALLFRIEND, y como éstas generaron un resultado notorio en cuanto al volumen de publicaciones técnicas en el campo de la LID.

3.4 Ekaterina Timoshenko

El trabajo de Ekaterina Timoshenko [14] trata el problema del LID y presenta resultados de investigación del ritmo del habla con el propósito de usarlo como un tipo de información adicional para mejorar el rendimiento de los sistemas LID.

La idea principal de su trabajo es explorar la duración de las unidades “*neighboring syllable-like*” como una característica discriminadora de los idiomas. Para el tratamiento apropiado del ritmo del habla, ella propone un algoritmo para la segmentación del habla basado en unidades rítmicas y desarrolla un enfoque independiente de los idiomas para modelar el ritmo. Como resultado, su sistema LID basado en el ritmo no requiere información transliterada del conjunto de entrenamiento y puede extenderse hacia nuevos idiomas.

Para explorar la influencia de las características rítmicas sobre el rendimiento de los sistemas LID, ella implementa un sistema LID fonotáctico y dos sistemas LID acústicos, los cuales fueron toma como base para realizar su trabajo. Sus sistemas LID son entrenados y probados con el corpus SpeechDat II, el cual fue diseñado para entrenar reconocedores de habla comerciales. Asimismo, cada sistema LID fue primero evaluado separadamente y después fueron fusionados en diferentes combinaciones.

Su sistema LID fonotáctico (Figura 3.2) es implementado utilizando el enfoque *PRLM*. En estos sistemas varios reconocedores de fonemas por idioma son utilizados independientemente para crear una secuencia de *tokens* con las muestras de entrada. Dicha secuencia es analizada por medio de un modelo de lenguaje estadístico para cada idioma del

conjunto. El número de reconocedores varía dependiendo el número de lenguajes en el conjunto de entrenamiento.

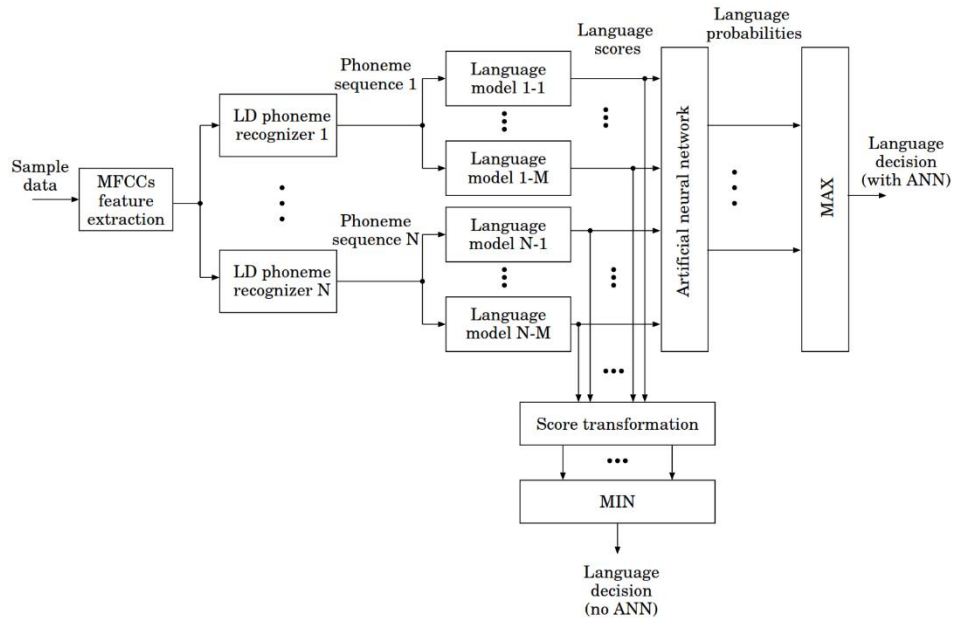


Figura 3.2. Diagrama del sistema LID fonotáctico.

Por otra parte, el sistema LID acústico que ella implementa trabaja dependiendo del tipo de segmentación del habla para generar las unidades rítmicas. Es aquí donde utiliza las unidades *syllable-like* para caracterizar el ritmo, las cuales son calculadas como se muestra en los diagramas de la Figura 3.3.

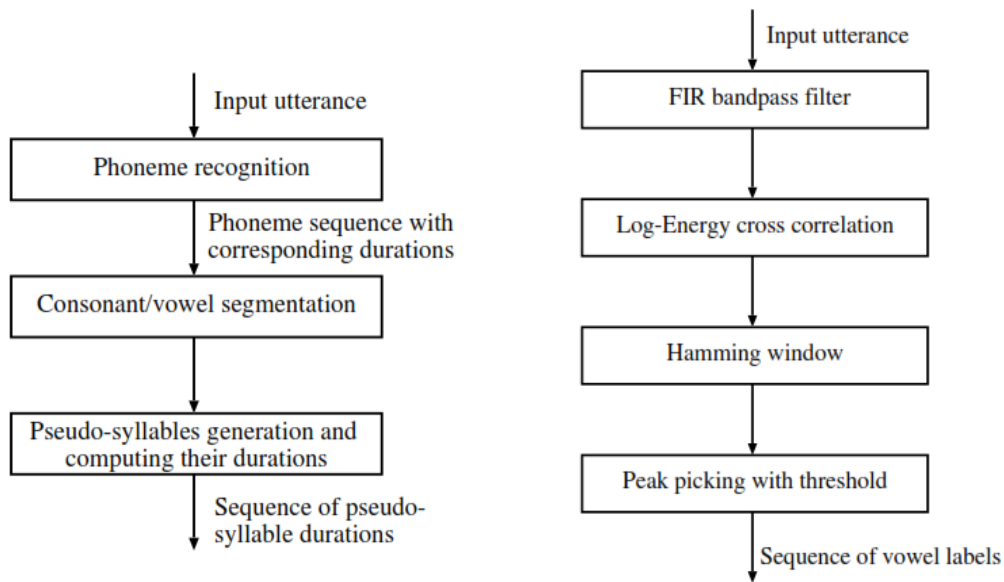


Figura 3.3. Diagramas de los métodos para calcular las unidades *syllable-like*.

Una vez calculadas las unidades *syllable-like*, el sistema crea una secuencia de estas unidades representando a cada muestra del conjunto de entrenamiento, creando a su vez los modelos acústicos para cada lenguaje. Una vez realizado esto el sistema LID funcionará de acuerdo a su representación de la Figura 3.4.

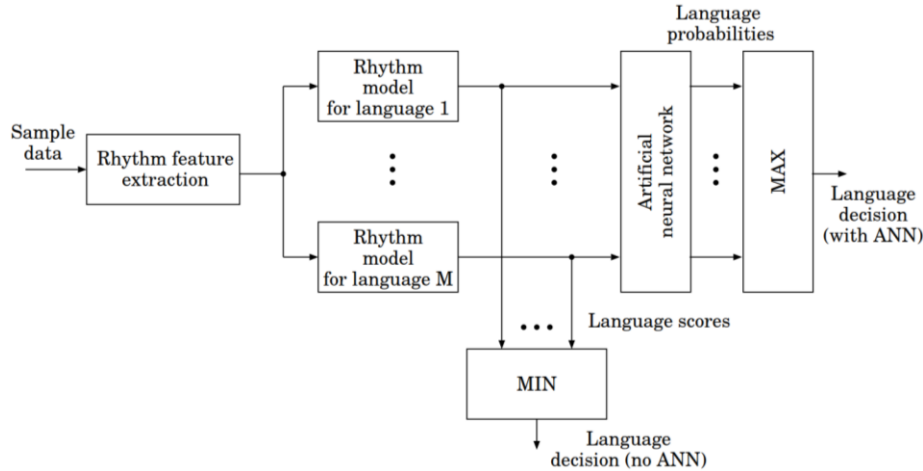


Figura 3.4. Diagrama del sistema LID acústico.

El añadir ritmo a cada sistema LID individual mejora la habilidad de identificación de cada uno de los sistemas, mejorándolos aproximadamente un 50%. Esto corresponde a la reducción aproximada de más del 20% para el escenario de identificación, lo que muestra confiabilidad en la decisión del sistema. Cada adición de un sistema nuevo a una combinación brinda una mejora menor. Por lo tanto, el añadir ritmo a una combinación de 3 sistemas LID puede mejorar el rendimiento de sus resultados por casi 3% y 7% para la identificación y detección de escenarios respectivamente. Lo que confirma que el ritmo del habla definido en su trabajo puede usarse satisfactoriamente en cualquier sistema LID.

3.5 Reyes Herrera y Vargas Martínez

La línea de investigación de *Reyes Herrera* [11] es la que por primera vez aporta en el campo de la LID la transformada wavelet como forma de caracterizar la señal de habla. Tiene como precedente la utilización en: el reconocimiento del habla, la detección de voz activa, el reconocimiento del locutor, entre otras.

La metodología usada por Reyes en ese trabajo usa la transformada *Wavelet Daubechies db2*. El método se basa en la idea de que los coeficientes de mayor magnitud corresponden a una buena representación de la señal de voz, y los coeficientes con magnitudes pequeñas corresponden a una mala representación de la señal. Entonces el siguiente paso es hacer un truncado de los coeficientes para solamente obtener los coeficientes de mayor magnitud, de esta forma se truncan de acuerdo a su magnitud con un umbral del 1%, eliminando

coeficientes que no representan bien la señal. Por ejemplo, para una muestra de 10 segundos donde obtuvo 131,072 coeficientes, se hizo una reducción a 1312.

Finalmente, usando el corpus de la OGI-TS y discriminando entre pares de idiomas con los mismos idiomas que [21] usó en su experimentación, con excepción del francés, ella compara sus resultados con los de él. Por ejemplo, para una muestra de 50 segundos (Tabla 3.2) sus resultados muestran que el uso de la transformada wavelet es muy pertinente en esta clase de estudios, mejorando los resultados obtenidos en [21].

Tabla 3.2. Resultados (porcentaje de clasificación correcta) reportados por [11] sobre muestras de 50s.

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	97	97	93	94	96	95	99	96
Alemán		93	94	93	98	98	94	91
Español			91	86	92	98	91	94
Mandarín				95	95	93	89	94
Vietnamita					93	96	95	95
Japonés						93	89	94
Coreano							95	91
Tamil								90

Vargas Martínez [8] sigue la misma línea de investigación que [11] usando la transformada Wavelet para caracterizar la señal de habla, su aporte a la investigación de Reyes, es que él contempla cuestiones no vistas en el trabajo de Reyes, como lo es el tiempo necesario para una aplicación real. Es decir, una aplicación real necesita una cantidad corta de habla para identificar el lenguaje en un tiempo razonable (en [11] sólo se reportan análisis de 50 segundos sobre habla libre continua más el procesamiento). También se utilizó un módulo de detección de voz activa, para la eliminación de pausas largas.

De esta forma la metodología está diseñada para trabajar con muestras pequeñas de habla, además de eliminar pausas largas en las señales de habla, usa medidas estadísticas simples (media, desviación estándar, máximo, mínimo) sobre los coeficientes wavelet obtenidos a partir de la señal segmentada. Asimismo, también realiza un truncado por fracción para extraer los coeficientes más representativos de la señal de habla, después se aplica la ganancia de información para posteriormente terminar en el módulo de identificación de lenguaje, donde utiliza un clasificador *Naive Bayes* (Figura 3.5).

Para las pruebas y experimentación, Vargas Martínez usó las muestras de habla espontánea del corpus OGI-TS, tomó 50 muestras por idioma de una duración de 50 segundos de los siguientes 9 idiomas: inglés, alemán, español, mandarín, vietnamita, japonés, coreano, tamil y farsi. Él realizó pruebas con diferente duración en las muestras de habla, tomando 30s, 10s, 5s y 4s de diferentes muestras. Dando como mejores resultados las de menor tiempo (Figura 3.6).

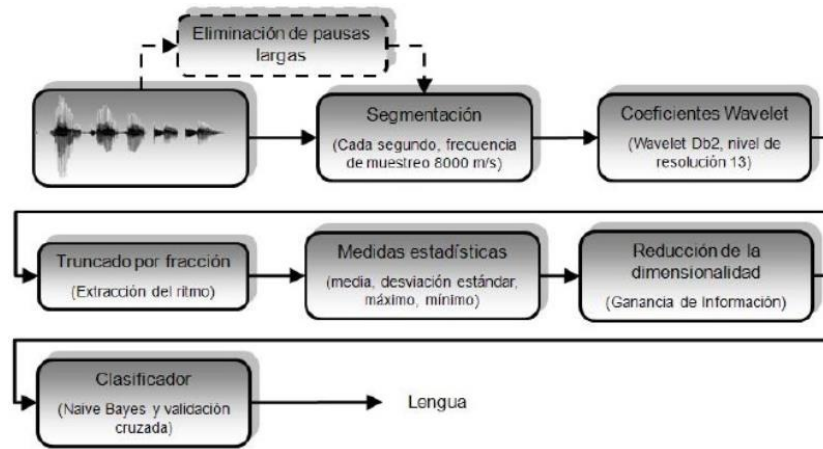


Figura 3.5. Metodología propuesta en [8].

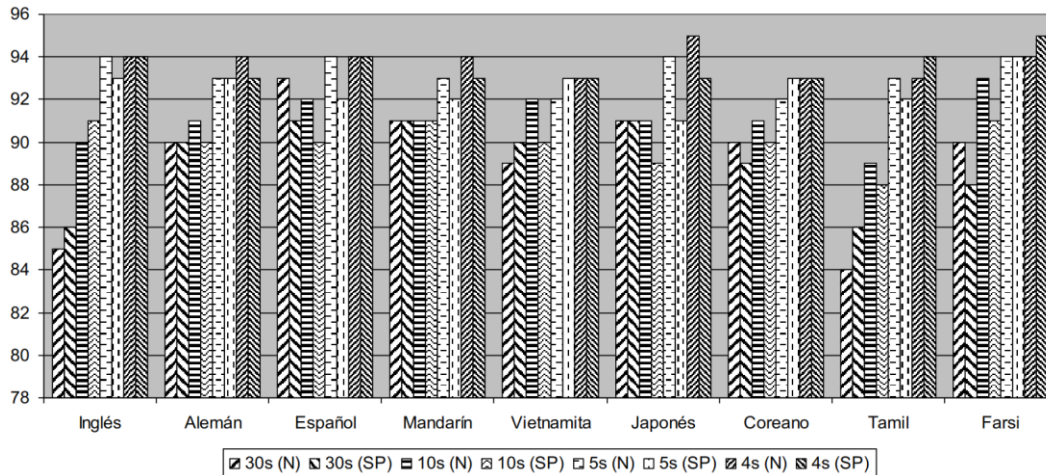


Figura 3.6. Resultados (en porcentaje de clasificación correcta) de todas las pruebas (Normal y Sin Pausas) documentadas en [8].

3.6 Flores Paulín y Caballero Morales

Encontrar trabajos que utilicen las lenguas indígenas mexicanas para crear sistemas o aplicaciones de habla no es algo muy común. Sin embargo, los trabajos [17] y [18] son dos trabajos actuales que utilizan las lenguas indígenas como base para desarrollar aplicaciones de habla.

El trabajo de Flores Paulín [17] está enfocado al reconocimiento de habla para palabras de la lengua náhuatl, específicamente las palabras correspondientes a los números del 1 al 10. Entre los aspectos que el consideró para poder realizar su trabajo, se encuentran: el corpus de habla en lengua náhuatl, las características a extraer y los modelos de reconocimiento.

El corpus utilizado en [17] consta de 20 grabaciones por dígito obtenidas de 3 personas adultas, la variante del idioma náhuatl no se especifica. Las grabaciones son de un solo canal (Mono), guardadas en el formato de archivo *wav*, con una frecuencia de muestreo de 22050 Hz y un *bit depth* de 16 bits. El dispositivo utilizado para realizar las grabaciones fue una grabadora digital modelo *Olympus VN-3100 PC*. Las grabaciones fueron hechas en lugares que tenían bajos niveles de ruido ambiental, cada persona decía 20 veces la palabra correspondiente a cada número, obteniendo al final 400 grabaciones por cada persona (1200 grabaciones en total). De cada conjunto de grabaciones por número, Flores Paulín reporta que de cada 20, 10 las utilizó para el conjunto de entrenamiento y las otras 10 para el conjunto de prueba.

Las características discriminatorias que [17] utiliza para el reconocimiento del habla son los coeficientes *LPC* y *MFCC*. Posteriormente, utiliza dichas características para entrenar los modelos acústicos utilizando *Modelos Ocultos de Markov*. En general, obtiene un promedio de 89% de acierto sobre los conjuntos de entrenamiento que probó.

El trabajo de Caballero Morales [17] presenta el desarrollo de recursos para el idioma indígena *mixteco*, los cuales son: un corpus de habla de relatos tradicionales de la cultura mixteca hablados por un hablante nativo del idioma (con el etiquetado a niveles fonético y ortográfico) y un sistema *ASR* (entrenado con el corpus del trabajo mismo) integrado con un traductor de texto *Mixteco-Español/Español-Mixteco*. El corpus de habla, aparte de estar limitado a una sola variante, fue lo suficientemente bueno para implementar la aplicación de habla multiusuario la cual presenta una media de rendimiento del 94.36% respecto a su reconocimiento/traducción en relación con los experimentos realizados con hablantes no nativos (los usuarios finales).

El vocabulario y texto representativo para el corpus de habla fue tomado de un material de educación del *Centro Cultural* de la ciudad de *Huajuapán de León, Oaxaca*. La variante del idioma mixteco usado en el corpus es la *San Juan Diquiyú*, la cual está localizada en el sur de *Huajuapán de León*, en el municipio de *Tezoatlán de Segura y Luna*. Dado que la variante *San Juan Diquiyú* comparte aspectos similares con otras variantes en *Oaxaca*, fue que hubo confianza sobre el uso de dicho material de habla.

El materia educación mencionado consiste en una colección de 15 relatos tradicionales de la cultura mixteca. Para este trabajo, se seleccionaron 7 relatos, los primeros relatos fueron usados para los usuarios principiantes y los últimos para los usuarios de más nivel. Cada relato fue dicho un cierto número de veces por un hablante nativo. Estas repeticiones fueron grabadas en el *Media Lab* de la Universidad de la Mixteca, en formato *wav*, con una frecuencia de muestreo de 44100 Hz y un canal (Mono). Aproximadamente 45 minutos de habla mixteca narrativa fue grabada. Esas grabaciones fueron transcritas a niveles fonético y de palabras (de acuerdo al estándar del corpus de habla *TIMIT*) usando la lista de fonemas de la Figura 3.7, la ayuda de una hablante nativo y el software *WaveSurfer*.

Description	IPA	Mixtec	Mexbet
Voiceless bilabial stop	p	p	p
Voiceless dental stop	t	t	t
Voiceless velar stop	k	k	k
Voiced bilabial stop	b	b	b
Voiced dental stop	d	d	d
Voiced velar stop	g	g	g
Voiceless palatal affricate	tʃ	ch	tS
Voiceless palatoalveolar fricative	ʃ	sh	
Voiceless labiodental fricative	f		f
Voiceless alveolar sibilant	s	s, dj	s
Voiceless velar fricative	x	j	x
Voiced palatal fricative	ʒ	y	Z
Bilabial nasal	m	m	m
Palatal nasal	ɲ	ñ	ñ
Alveolar nasal	n	n	n
Alveolar lateral	l	l	l
Alveolar trill	r	r	r
Alveolar flap	ɾ		r(
Close front unrounded vowel	i	i, í, ì	i
Close-mid front unrounded vowel	e	e, é	e
Open front unrounded vowel	a	a, á, à	a
Close-mid back rounded vowel	o	o, ó, ò	o
Close back rounded vowel	u	u, ú, ù	u
Glottal stop	ʔ	'	
Additional		nd, ng	ks _D, _G _N, _R
		sil	sil

Figura 3.7. Repertorio de fonemas mixtecos.

En total, este corpus de habla Mixteco consta de 931 palabras con un vocabulario de 192 palabras únicas. Usando todas esas palabras, calcularon la frecuencia de cada fonema, por lo que pudieron sustentar el uso de este corpus para que pudiese ser usado por su sistema ASR.

4

Corpus Entendámonos

Este capítulo intenta narrar todos los aspectos que se consideraron en el diseño y desarrollo del corpus *Entendámonos*. Las muestras del corpus de habla fueron recolectadas únicamente en el estado de San Luis Potosí, sobre algunas localidades del área conocida como *La Huasteca Potosina* (Figura 4.1). Este corpus contiene grabaciones de personas adultas, de habla nativa de los idiomas náhuatl, tének y xi'iuuy. Sin embargo se espera que más idiomas se incorporen en un futuro (ver Capítulo 6).



Figura 4.1. Zona: La Huasteca Potosina (color verde). México.

Cabe mencionar que a principios de su desarrollo, este proyecto resultó ser muy ambicioso en el sentido de querer recabar más muestras de lo que la realidad del problema permite; desarrollar a gran escala un corpus de habla de lenguas indígenas es una tarea que podría llevar años debido a la gran cantidad de diversidad cultural, al difícil acceso de algunas

Sección 4.1 Especificación del Corpus

variantes de las lenguas y la falta de interés o cooperación que las instituciones o personas de habla indígena pueden mostrar.

Sin embargo, la idea se encuentra ahí, la necesidad de un corpus de lenguas indígenas existe y las ventajas de tener este tipo de recursos de habla podría ayudar al desarrollo de diversas aplicaciones de habla que podrían ser la clave para solucionar diversos problemas de hoy en día, ya sean los mencionados en el Capítulo 1 o aquellos que fueron mencionados en los trabajos [17] y [18].

4.1 Especificación del Corpus

El corpus *Entendámonos* es un corpus de habla que, hasta el momento, consiste en una colección de grabaciones de personas adultas de los idiomas indígenas náhuatl, tének y xi'iuy.

Entendámonos fue diseñado con la idea de tratar únicamente el problema de la Identificación Automática del Lenguaje Hablado debido al propósito con el que fue diseñado, ya que al comienzo, el proyecto careció de lingüísticas que pudieran colaborar con las transcripciones de las muestras.

Además, al parecer este trabajo es el primero en su clase debido a que nunca antes se había decidido tratar a las lenguas indígenas desde esta perspectiva, cuya finalidad es el crear recursos de habla (*speech resources*) reutilizables para otras áreas y dejar una primera propuesta para que pueda ser considerada como base o referencia hacia trabajos futuros de esta índole. La Figura 4.2 se pueden apreciar las fases y fechas por las cuales paso *Entendámonos*.

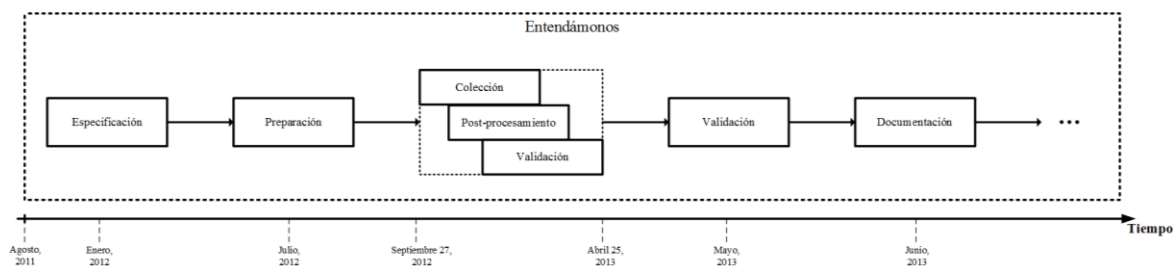


Figura 4.2. Diagrama de las fases de desarrollo por las cuales paso el corpus *Entendámonos*.

■ ■ Nota: Los últimos 3 puntos mostrados en la Figura 4.2 indican que el proyecto aún sigue en curso en la *División de Estudios de Posgrado e Investigación del Instituto Tecnológico de Ciudad Madero*. □ □

4.2 Zonas, Localidades y Ubicaciones

Las muestras de los idiomas indígenas recabados en este corpus fueron recolectadas en diferentes comunidades y pueblos de la zona conocida como *La Huasteca Potosina* (Figura 4.3). Las localidades que el proyecto abarcó en dicha zona están ubicadas en la Figura 4.4.

■ ■ Nota: A manera de brindar una mejor precisión de la ubicación de estas localidades, se hará uso de los mapas de la tecnología de *Google Maps*. □□

Todas las voces incorporadas en el corpus provienen de 20 localidades en total, ya sea porque se realizó una visita a la comunidad/pueblo específico o porque se encontró a una persona que hablase un idioma distinto al de la zona en concreto.

Las comunidades o pueblos abarcados por el proyecto fueron:

- | | |
|--------------------------|------------------|
| 1. Jalpilla | 11. Coxcatlán |
| 2. Picholco | 12. Agua Puerca |
| 3. Matlapa | 13. Agua Buena |
| 4. Tamazunchale | 14. Tamasopo |
| 5. Tanlajás | 15. Rayón |
| 6. Aquismón | 16. La Cuchilla |
| 7. Tancanhuitz de Santos | 17. La Parada |
| 8. Tanquián de Escobedo | 18. San Pedro |
| 9. Tampamolón Corona | 19. El Mezquital |
| 10. Huehuetlán | 20. Santa María |

De las 20 localidades, formalmente 4 pertenecen a la zona náhuatl, 7 a la zona tének y 9 a la zona xi'iuy. Aunque, en las zonas náhuatl y tének, la realidad es que uno puede encontrar gente que hable alguno de esos dos idiomas independientemente del lugar, gracias al transporte público y a la comunicación que tienen esas localidades. Además es importante señalar que estas zonas se refieren únicamente a las áreas que el proyecto abarcó, ya que formalmente la zonas designadas a los náhuatl, tének y xi'iuy de la Huasteca Potosina abarcan una mayor cantidad de lugares.

Las localidades que forman parte de la zona náhuatl son: Jalpilla, Picholco, Matlapa, Tamazunchale. La ubicación de ellas se puede observar en la Figura 4.5, junto con la del *IELIIP*.

Sección 4.2 Zonas, Localidades y Ubicaciones

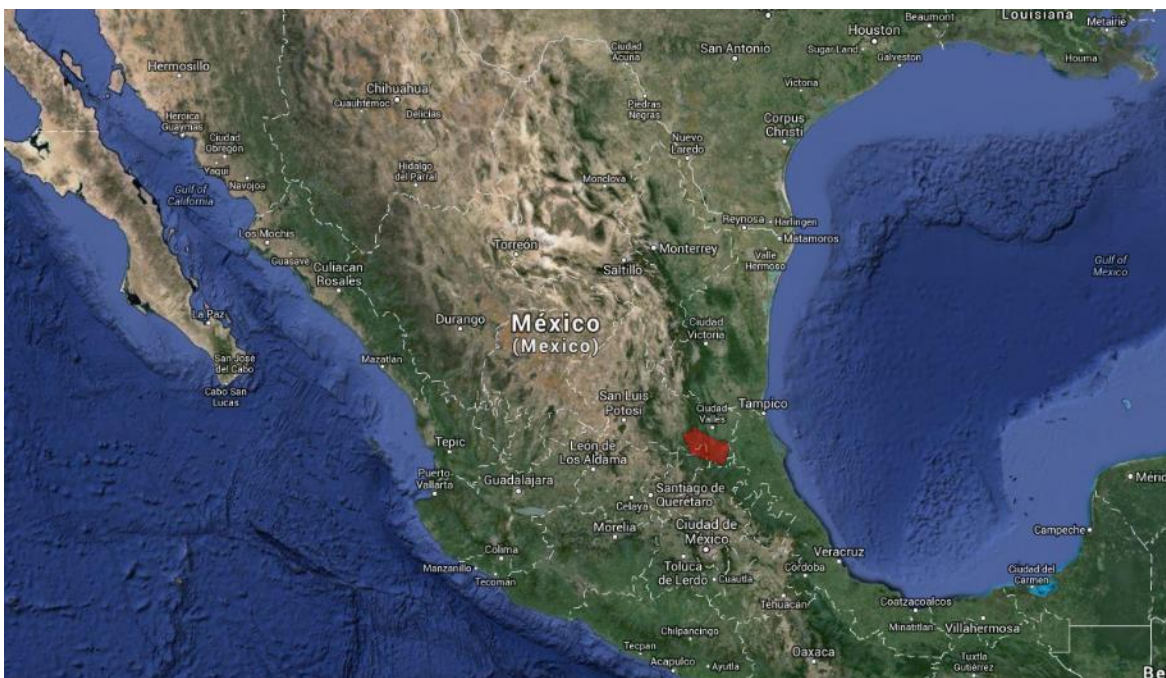


Figura 4.3. Área mexicana (color rojo) que el proyecto abarco.

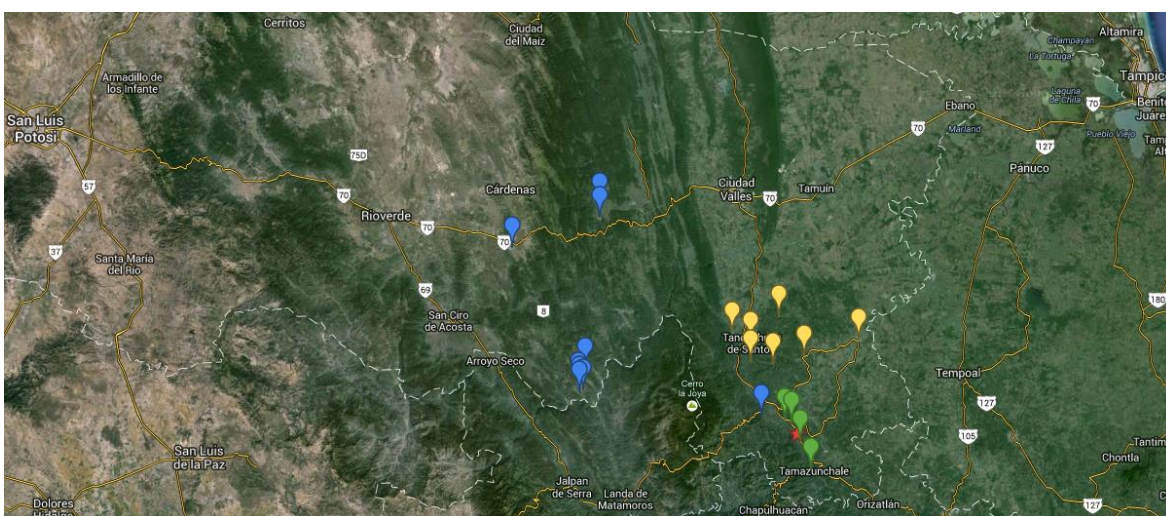


Figura 4.4. Localidades (marcadores azules, amarillos y verdes) del estado de San Luis Potosí abarcadas por el corpus.

■ ■ Nota: En la Figura 4.4 y Figura 4.5 se puede apreciar un marcador estrella de color rojo, este representa la ubicación del *Instituto Estatal de las Lenguas Indígenas e Investigaciones Pedagógicas (IELIIP)* el cual tuvo un importante papel en el desarrollo del corpus, brindando información del campo y personal para poder ir a las comunidades para grabar.

Capítulo 4. Corpus Entendámonos

Asimismo, en la Figura 4.6 (también en la Figura 4.4 aunque no se alcanza a distinguir) se puede apreciar un marcador estrella de color naranja, el cual representa la ubicación de la *Radio Difusora XEANT “La Voz de las Huastecas”*. □□

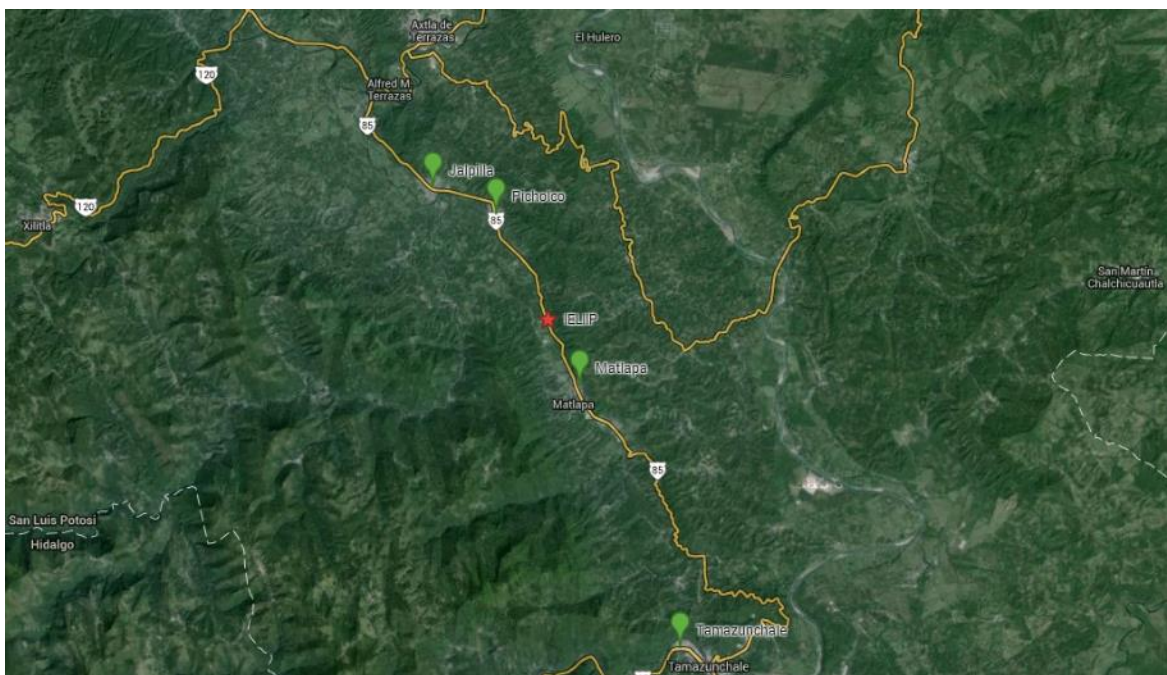


Figura 4.5. Zona náhuatl abarcada por el proyecto (marcadores verdes) y el IELIIP (marcador rojo).

Las localidades que forman parte de la zona tének son: Tanlajás, Aquismón, Tancanhuitz de Santos, Tanquián de Escobedo, Tampamolón Corona, Huehuetlán, Coxcatlán. La ubicación de ellas se puede observar en la Figura 4.6. Además, es importante señalar que en la ciudad de Tancanhuitz de Santos se encuentra la Radio Difusora XEANT, la cual fue la primer visita por parte del proyecto para la recolección de muestras, ya que cuando se visitó se estaba celebrando un evento llamado “*Fiesta de la Radiodifusora Indígena*”.

Finalmente, las localidades que forman parte de la zona xi’iuy son: Agua Puerca, Agua Buena, Tamasopo, Rayón, La Cuchilla, La Parada, San Pedro, El Mezquital, Santa María. Debido a la variante lingüística del idioma xi’iuy, la zona xi’iuy se divide en dos partes: *xi’iuy norte* y *xi’iuy sur*. A diferencia de las comunidades y pueblos de las zonas náhuatl y tének, estos lugares fueron los más difíciles de acceder por varias razones, por ejemplo que no hay transporte público o si quiera caminos pavimentados para poder llegar a la mayoría de estos lugares, entre otras. La ubicación de las comunidades o pueblos de la zona xi’iuy se puede apreciar en la Figura 4.6, Figura 4.7 y Figura 4.8.

Sección 4.2 Zonas, Localidades y Ubicaciones

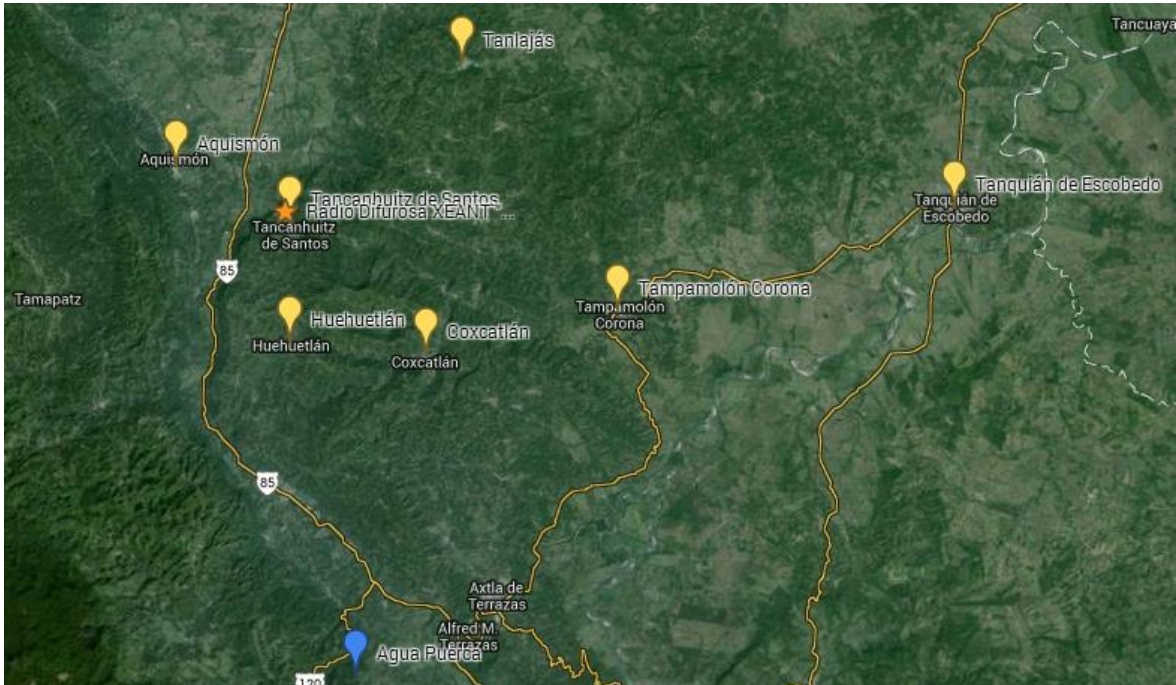


Figura 4.6. Zona tének abarcada por el proyecto (marcadores amarillos) y Radio Difusora XEANT (marcador naranja). Parte de la Zona xi'iuy (marcador azul).

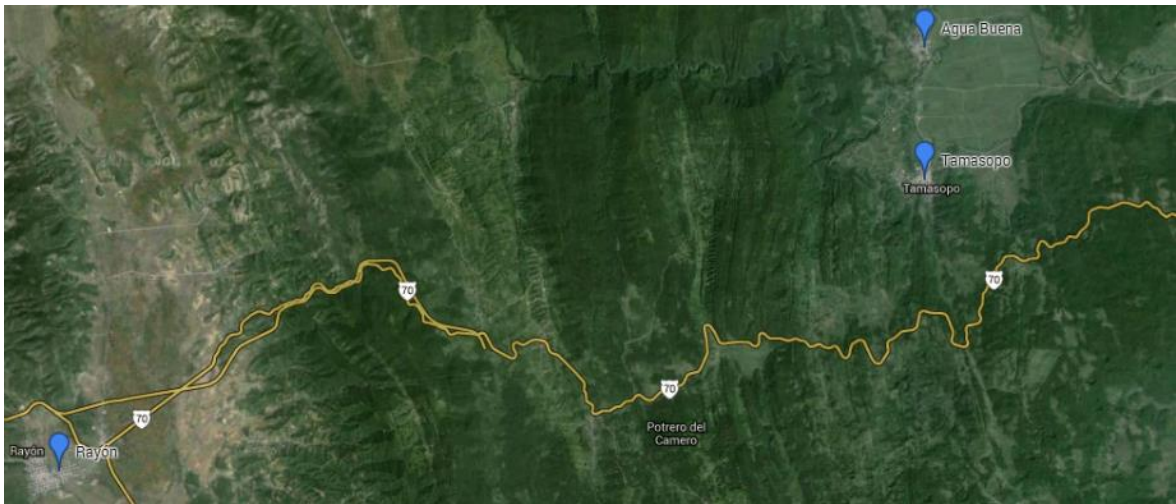


Figura 4.7. Zona xi'iuy norte abarcada por el proyecto (marcadores azules).

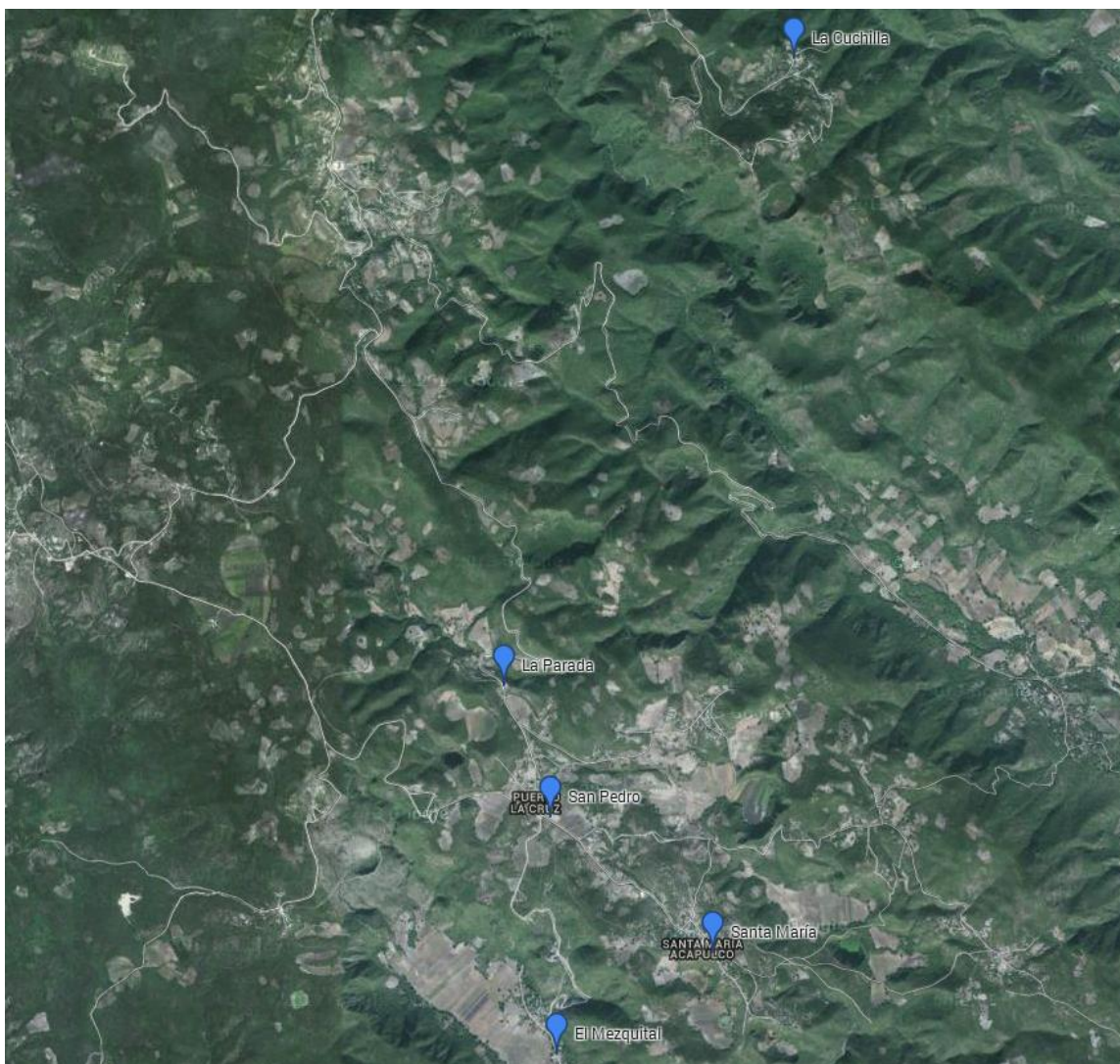


Figura 4.8. Zona xi'iuy sur abarcada por el proyecto (marcadores azules).

4.2 Tipos de Habla y Cuestionarios

Los *tipos de habla* es una característica clave en todo corpus ya que define los posibles usos de un corpus. Los corpus de habla que contienen un solo tipo de habla están reprimidos respecto su reuso hacia distintas aplicaciones. Para no caer en tal pena, el diseño de *Entendámonos* contempla 4 tipos de habla, los cuales son:

- Habla en respuesta a información específica (*Answering Speech*).
- Habla descriptiva (*Descriptive Speech*).
- Habla espontánea/libre (*Spontaneous/Free Speech*).
- Habla de dictado (*Dictation speech*).

Sección 4.2 Tipos de Habla y Cuestionarios

Es necesario hacer notar que el corpus también contiene un quinto tipo de habla, el cual es *habla emocional (emotional speech)*. Aunque las muestras que contienen dicho tipo de habla son escasas, aquellas con este tipo de habla no serán distribuidas de ninguna forma. Estos archivos faltantes están solamente notificados en los metadatos del corpus.

Para lograr dichos tipos de habla, a los hablantes se les facilitó un cuestionario. Este cuestionario cuenta con dos versiones; sin embargo la segunda versión fue la única que se utilizó debido a que se tuvo la oportunidad de mejorarlo antes de empezar la fase de recolección de muestras. Esto se logró con la ayuda de las personas (todas hablantes nativas de cada uno de los 3 idiomas) que ocupan un puesto de locutor en la *radio difusora XEANT* (su ubicación se puede apreciar en la Figura 4.6) así como también del director de dicha radio difusora.

La primer versión del cuestionario consistía de la siguiente forma:

1. ¿Cómo te llamas?
2. ¿Cuál es tu lenguaje nativo?
3. ¿Cuál es el lenguaje que más hablas en tu hogar?
4. ¿Cuántos años tienes?
5. ¿Cuál es tu fecha de nacimiento (día/mes/año)?
6. ¿Eres hombre o mujer?
7. ¿Cómo se llama la ciudad donde naciste?
8. ¿Cuál es tu día favorito, cuál es el día que más te gusta?
9. ¿Qué hora es?
10. Describe lo mejor que puedas la ruta que tomas de tu casa a la escuela.
11. Menciona algo que te guste de tu ciudad.
12. Menciona algo que te guste de tu escuela.
13. Menciona cómo es el clima de tu ciudad.
14. Describe tu escuela.
15. ¿Qué fue lo último que comiste? Descríbelo lo mejor que puedas.
16. Menciona los días de la semana.
17. Di los números del 1 al 10.
18. Ahora, hable del tema que usted quiera y/o desee por un periodo de 1 minuto, es importante hable continuamente sin realizar alguna pausa en su relato.
19. Como último, requerimos que realices la lectura de un dictado, el cual está conformado de palabras correspondientes a las partes del cuerpo. Por favor, haga una pausa entre cada palabra.

- | | | | |
|---------------|-------------------|--------------------|-----------------|
| 1) arteria | 17) cráneo | 33) ingle | 47) muslo |
| 2) apéndice | 18) coxis | 34) intestino | 48) nariz |
| 3) boca | 19) cuello | grueso | 49) nuca |
| 4) brazo | 20) dedo | 35) intestino | 50) oídos |
| 5) bronquios | 21) dientes | delgado | 51) ojos |
| 6) cabello | 22) esófago | 36) iris (del ojo) | 52) ombligo |
| 7) caderas | 23) espalda | 37) joroba | 53) orejas |
| 8) cara | 24) esqueleto | 38) laringe | 54) páncreas |
| 9) ceja | 25) esternón | 39) lengua | 55) pantorrilla |
| 10) cerebro | 26) estómago | 40) mano | 56) párpado |
| 11) cintura | 27) faringe | 41) mandíbula | 57) pecho |
| 12) clavícula | 28) fosas nasales | 42) mejilla | 58) pelos |
| 13) codo | 29) fémur | 43) mentón | 59) peroné |
| 14) colmillo | 30) ganglio | 44) muelas | 60) pestañas |
| 15) corazón | 31) hígado | 45) muñeca | 61) piel |
| 16) costilla | 32) índice (dedo) | 46) músculo | 62) pierna |

Capítulo 4. Corpus Entendámonos

63) pies	69) rodilla	75) tórax	81) vesícula
64) pómulo	70) rostro	76) tráquea	seminal
65) pulmones	71) rótula	77) uña	82) vientre
66) quijada	72) sien	78) venas	83) yugular
67) rabadilla	73) sangre	79) vértebra	
68) riñones	74) tobillo	80) vesícula biliar	

El primer cuestionario fue realizado con los ejemplos que se encuentran en los trabajos [1], [6] y [19]. El motivo de su apariencia un tanto absurda fue a causa de nuestra inexperiencia en los 3 idiomas, además de en aquel entonces no contaba el apoyo del *IELIIP* (no se sabía siquiera de su existencia), y al final de cuentas, no queríamos que se nos pasara preguntar algún dato sólo por la excusa de no saber.

A manera de elaborar un cuestionario que fuera parejo para los 3 idiomas, las modificaciones que se le hicieron siguieron una única regla: Si la pregunta o palabra del dictado era difícil de responder o decir, se removía o cambiaba.

Las recomendaciones que dieron lugar a las modificaciones del cuestionario (de la primera versión respecto a la segunda) fueron las siguientes:

- El cambio de varias palabras con que se utilizaban en la primer versión, por ejemplo el referirse a una *comunidad* o *pueblo* como *ciudad* para los hablantes es confuso. Estos cambios se realizaron en varias preguntas.
- Las preguntas 8 y 16 se removieron debido a que la mayoría de los hablantes no saben decir los días de la semana en su idioma.
- Las preguntas 10, 12 y 14 se tuvieron que remover debido a que desafortunadamente son pocos los jóvenes que van a una escuela (y que tengan mayoría de edad) en dichas regiones.
- El número de palabras del dictado de las partes del cuerpo se redujo de 83 a 37 palabras.

Una vez que se recibió la ayuda del personal de la radio difusora, la segunda versión del cuestionario quedó de la siguiente forma:

1. ¿Cómo te llamas?
2. ¿Cuál es tu lenguaje nativo?
3. ¿Cuál es el lenguaje que más hablas en tu hogar?
4. ¿Cuántos años tienes?
5. ¿Cuál es tu fecha de nacimiento (día/mes/año)?
6. ¿Eres hombre o mujer?
7. ¿Cómo se llama la comunidad donde naciste?
8. ¿Qué hora es?
9. Menciona algo que te guste de tu pueblo.
10. Menciona cómo es el clima de tu pueblo.
11. ¿Qué fue lo último que comiste? Descríbelo lo mejor que puedas.
12. Di los números del 1 al 10.
13. Ahora, hable del tema que usted quiera y/o desee por un periodo de 1 minuto, es importante hable continuamente sin realizar alguna pausa en su relato.

Sección 4.3 Perfil de los Hablantes y Número de Hablantes

14. Como último, requerimos que realices la lectura de un dictado, el cual está conformado de palabras correspondientes a las partes del cuerpo. Por favor, haga una pausa entre cada palabra.

- | | | | |
|--------------|---------------|--------------|-------------|
| 1) boca | 11) corazón | 21) mano | 31) pierna |
| 2) brazo | 12) costilla | 22) nariz | 32) pies |
| 3) cabello | 13) cuello | 23) nuca | 33) rodilla |
| 4) caderas | 14) dedo | 24) oídos | 34) rostro |
| 5) cara | 15) dientes | 25) ojos | 35) sangre |
| 6) ceja | 16) espalda | 26) ombligo | 36) uña |
| 7) cerebro | 17) esqueleto | 27) orejas | 37) venas |
| 8) cintura | 18) estómago | 28) pelos | |
| 9) codo | 19) intestino | 29) pestañas | |
| 10) colmillo | 20) lengua | 30) piel | |

La segunda versión del cuestionario fue la que se utilizó para realizar toda la fase de recolección de muestras del corpus. Sin embargo, en la Sección 6.3 se proponen algunas mejoras que ayudarían a realizar una mejor entrevista.

4.3 Perfil de los Hablantes y Número de Hablantes

Los hablantes que formaron parte de *Entendámonos* tenían que cumplir con dos requisitos en su perfil. El primero era que la persona debía tener 18 años de edad o más. El segundo era que la persona debía poseer un gran dominio de habla de uno de los 3 idiomas del corpus.

Sin embargo, en un principio, el segundo requisito era que las personas debían ser estrictamente *hablantes nativos* de uno de los 3 idiomas del corpus. Pero, conforme avanzó la investigación y el desarrollo del corpus, se descubrió que el término *hablante nativo* es todavía un concepto muy ambiguo.

Unos definen a un *hablante nativo* como una persona que utiliza cierto idioma como su *primer idioma o lengua madre*, o bien, una persona que ha hablado cierto idioma desde la infancia. Otros afirman que un *hablante nativo* es una persona que obtuvo los procesos de adquisición del o de los idiomas desde niño (o bien, a muy temprana edad), no antes de la pubertad, ya que después de la pubertad, es difícil (no imposible, sin embargo muy difícil) volverse un hablante nativo de cierto idioma [22].

Lo anterior tiene en cierta manera su justificación. Por ejemplo, conforme se fue desarrollando el corpus, nos encontramos con casos de jóvenes (entre 18 y 26 años aproximadamente) que empezaron a aprender a hablar el idioma indígena de sus padres poco antes de tener la mayoría de edad, y sin embargo, el manejo que poseen del idioma indígena demostró ser indudablemente aceptable debido al uso cotidiano que le dan. Asimismo, también es muy común encontrar hablantes, más que nada jóvenes, que hablan de manera fluida su idioma, pero mezclándolo mucho con palabras del español. Esto no quiere decir que posean o no un dominio en el idioma, es tan sólo un cambio que se ha venido dando en los

Capítulo 4. Corpus Entendámonos

idiomas (más que nada náhuatl y tének) indígenas debido a la comunicación que hay por parte de los pueblos y comunidades con las ciudades.

Entendámonos cuenta hasta el momento con 143 diferentes hablantes de los 3 idiomas, concretamente: 46 hablantes del idioma náhuatl, 50 hablantes del idioma tének y 47 hablantes del idioma xi'iuy. En donde cada hablante aporta mínimo 50 segundos de *habla libre*. En la Tabla 4.1 se muestra algunas estadísticas generales de los hablantes.

Tabla 4.1. Resumen general de *Entendámonos*.

	Náhuatl			Tének			Xi'iuy		
N° de Mujeres	13			26			38		
N° de Hombres	33			24			9		
Edad Mínima [Mujeres, Hombres]	19	19		21	24		16	19	
Edad Máxima [Mujeres, Hombres]	78	83		66	80		77	55	
Edad Promedio [General, Mujeres, Hombres]	47	44	48	46	40	53	32	32	34
N° de Cuestionarios Completos	25			9			3		
N° de Relatos	46			50			47		

El *N° de Relatos* de la tabla anterior se refiere a los relatos que corresponden a la respuesta de la *pregunta 13* del corpus.

4.3 Protocolo de Grabación

La manera de conseguir hablantes para el corpus, era hacer visitas a los pueblos y comunidades indígenas para buscar personas que aceptaran colaborar con el proyecto. Una vez que se llegaba al pueblo/comunidad, lo que seguía era buscar a la persona con el cargo político del lugar. Normalmente estas personas ocupaban los cargos de juez, comisariado, suplente de juez o suplente de comisariado. Y eran esas personas las que nos guiaban a los hogares de distintos individuos para preguntarles si querían participar en el proyecto.

La mecánica para empezar a grabar a una persona era que dicha persona primero aceptara colaborar, de lo contrario, la grabación jamás tenía lugar. La muestras de *Entendámonos* son sólo de personas que aceptaron donar su voz.

Una vez que la persona aceptaba colaborar con el proyecto, se buscaba un lugar que careciera de ruido. Aquí es importante señalar lo que se mencionó en la Sección 4.3, todas las muestras

Sección 4.3 Protocolo de Grabación

fueron grabaciones de campo. Estas normalmente se hacían en alguna parte del hogar del hablante, o bien, en alguna parte del lugar en donde la persona que nos guiaba reunía a la población de la comunidad. Siempre se intentó buscar el mejor lugar para grabar, sin embargo la presencia del *ruido ambiental* es algo con lo que siempre se tuvo que lidiar.

Antes de empezar a grabar, se le entregaba una hoja con el cuestionario a cada hablante y se les explicaba que la idea era que respondieran cada una de las preguntas, sin leer las preguntas. Enseguida se les comentaba que no era necesario contestarlas todas, o bien, si no querían contestar alguna, se les pedía que mínimo contestaran la *pregunta 13*. Se tuvo que hacer de esta forma debido a varias razones:

- A muchas personas les daba pena hablar en el micrófono.
- Hubo una gran cantidad de personas que no sabían leer español ni el idioma indígena que manejan, sin embargo eran persona que querían colaborar, por lo que el pedirles que sólo respondieran la *pregunta 13* resultaba lo más viable.
- Hubo otras personas que no mostraban mucho interés en el proyecto, pero que con ayuda del lingüista o de personas que conocí en el pueblo/comunidad, accedían a colaborar en una forma mínima, por lo que sólo les pedía responder la *pregunta 13*.

Enseguida, se creaba un directorio, en la carpeta del respectivo idioma que hablase, con un nombre que seguía la siguiente nomenclatura:

[id del hablante] [nombre del hablante] [edad del hablante] [genero del hablante]

Por ejemplo, en el caso de que el autor de esta tesis fuera un hablante indígena, el directorio quedaría de la forma mostrada en la Figura 4.9.

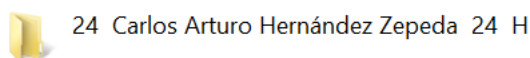


Figura 4.9. Ejemplo del nombre de una carpeta de un hablante, usando la correspondiente nomenclatura.

Una vez que los hablantes estuvieran listos, se empezaba grabar utilizando el software *Audacity*®¹¹ y el micrófono *SHURE PG42-USB* (ver Apéndice B). Antes de que la persona empezara a contestar el cuestionario, se realizaba una grabación del ambiente del lugar de unos *3 a 5 segundos* con el fin de que se pudiera crear un filtro que ayudase a disminuir el ruido ambiental del lugar de grabación (ver Sección 4.5). Posteriormente, en el caso de que las personas accedieran a contestar todo el cuestionario, la grabación se dividía en *4 sesiones*. Si sólo quería responder ciertas preguntas, la grabación se dividía en sesiones según las preguntas que el hablante eligiera.

¹¹ *Audacity*® es un software libre editor de grabación y edición de sonido, de código abierto y multiplataforma. Disponible para Windows®, Mac®, GNU/Linux® y otros sistemas operativos. Distribuido bajo la Licencia Pública General de GNU (GPL).

Capítulo 4. Corpus Entendámonos


Cada sesión corresponde a una grabación que comprende un cierto número de cuestiones. La primer sesión comprendía de la pregunta 1 a la 8. La segunda sesión comprendía de la pregunta 9 a la 12. La tercer sesión comprendía la pregunta 13. La cuarta sesión comprendía la pregunta 14. Jamás se dio el caso en donde un hablante quisiera responder sólo ciertas preguntas de cada una de las sesiones definidas.

Se hizo de este modo para que los hablantes tuvieran lapsos para descansar y para que pudieran pensar lo que iban a responder, además de que podían preguntar alguna duda que tuviesen.

Cada vez que se terminaba una sesión, esta se guardaba con la siguiente nomenclatura:

[id del hablante] p[intervalo de preguntas respecto su sesión]

Por ejemplo, en el caso de acabar un cuestionario completo, la carpeta del hablante tendría 4 archivos sus respectivos nombres mostrados en la Figura 4.10.



24 p1-8
24 p9-12
24 p13
24 p14

Figura 4.10. Archivos resultantes de una grabación de un cuestionario completo.

Una vez que se terminaba la grabación, se procedía a buscar otra persona que quisiera colaborar. En el caso de encontrarla, se repetía la mecánica del protocolo de grabación.

4.4 Especificaciones Técnicas

Las muestras de *Entendámonos* fueron grabadas a una frecuencia de muestreo de *44100 Hz*, con un *Bit Depth* de *16 bits* y un formato de muestra *pcm*, debido a dos cosas:

- Se tenía la teoría de que el sistema LID implementado en este trabajo (ver Capítulo 6) tendría un mejor rendimiento con muestras de mayor frecuencia de muestreo (antes de finalizar el corpus, el sistema se probó con el corpus OGI-TS) ya que el corpus utilizado para probar el sistema tenía muestras de *8000 Hz*.
- Debido a las grabaciones de prueba que se realizaron antes de empezar el desarrollo de *Entendámonos*, se consideró que el utilizar el estándar sería beneficioso en el caso de que se necesitara realizar un *down-sampling*¹² a las muestras del corpus.

¹² Término en inglés que consiste en la acción de remuestrear una señal a una frecuencia de muestreo más baja.

Sección 4.5 Post-procesamiento

El uso de un *bit depth* de *16 bits* se debe a que es el estándar recomendado para realizar grabaciones de voz según [1].

Las muestras del corpus también poseen una frecuencia de muestreo de *44100 Hz*. En el caso de que el usuario quiera utilizar una frecuencia diferente, la distribución de *Entendámonos* contempla la adición de los archivos *raw* de cada muestra con el fin de satisfacer esta necesidad.

El tipo de información que *Entendámonos* maneja es únicamente habla (*speech*) debido a que aún se carece de una suficiente cantidad de transcripciones. Se espera que en un futuro el corpus no sólo maneje habla, sino también texto (transcripciones de habla y fonemas).

La fuente de datos de *Entendámonos* es habla de micrófono (*microphone speech*), particularmente grabaciones de campo (*field recordings*).

El corpus sólo maneja 2 formatos de archivo: *wav* y *raw*. Las muestras están distribuidas en el formato *wav* ya que es un formato de distribución que la mayoría de los corpus utilizan, esto es debido a que es un formato sin compresión de datos, por lo que no pierde calidad respecto a la señal original. Los archivos *raw*, hacen referencia a archivos con formato *raw*, el cual es un formato con la capacidad de contener cualquier tipo de señal, son especialmente útiles para realizar pruebas sobre la señal original y así poder exportarla a otro formato.

Finalmente, *Entendámonos* maneja un total de 225 minutos de habla libre: 55 minutos de náhuatl, 77 minutos de tének, 93 minutos de xi'iuy.

4.5 Post-procesamiento

Una vez que se decidió finalizar el proceso de recolección de muestras, las etapas de post-procesamiento realizadas fueron: filtrado, editado, asignación de formato y asignación de nombre.

Es importante mencionar que los archivos *raw* de las muestras no poseen estas etapas de post-procesamiento, si el usuario de *Entendámonos* necesita trabajar sobre la señal original de la muestra distribuida en su versión del corpus, tendrá que aplicar las etapas explicadas en esta sección sobre el o los respectivos archivos *raw*.

Asimismo, todas las etapas de post-procesamiento fueron realizadas utilizando el software libre *Audacity*®.

4.5.1 Filtrado

Una de las desventajas de realizar grabaciones de campo, es la cantidad de ruido ambiental que el área o lugar de la grabación puede tener o generar, y desafortunadamente, es algo que uno no puede controlar.

Como forma de tratar ese problema, la solución más práctica que se encontró fue el realizar una grabación del ambiente previa a la grabación de cada persona con el fin de poder crear un filtro capaz de disminuir a gran medida el ruido ambiental. Esto consistía en realizar una grabación de *3 a 5 segundos* del sonido del ambiente del lugar donde se iba a realizar dicha grabación.

Concretamente, el realizar filtrado del ruido ambiental consistía en:

1. Seleccionar el segmento de la señal considerada como ruido ambiental (Figura 4.11).
2. Crear un *perfil de ruido* con el segmento seleccionado en 1 (Figura 4.12).
3. Aplicar este *perfil de ruido* como filtro a toda la señal (Figura 4.13).

Los parámetros recomendados para la reducción del ruido fueron:

- Reducción de ruido: *40 dB*
- Sensibilidad: *-2.00 dB*
- Suavizado de frecuencia: *100 Hz*
- Tiempo de ataque o decaimiento: *0.05 seg*

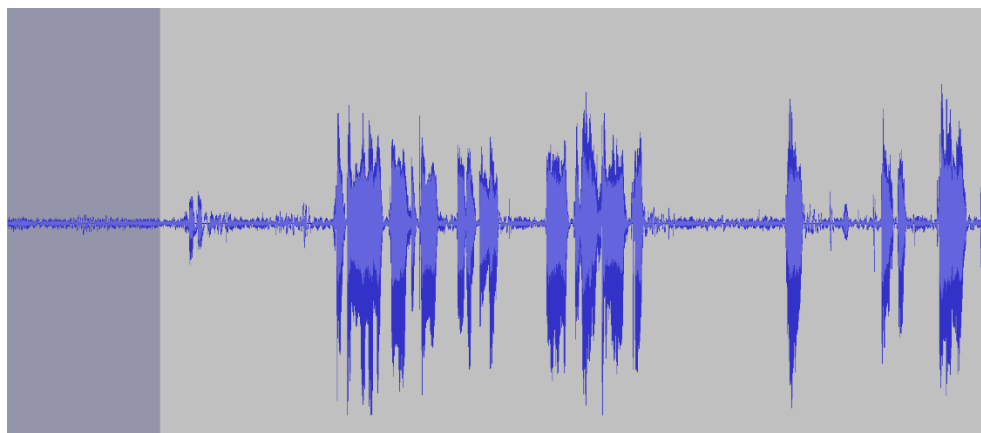


Figura 4.11. Señal de 20s de habla. La parte seleccionada al principio con gris oscuro es el segmento (de 3s en este caso) que se considera como ruido ambiental.

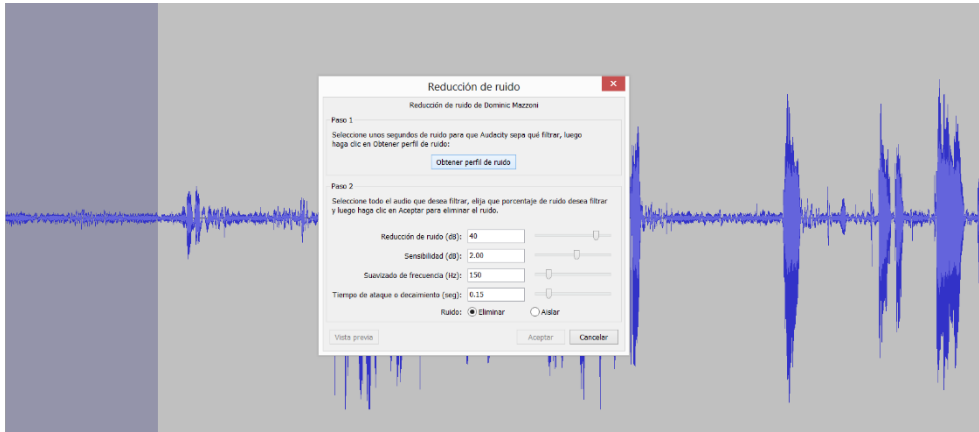


Figura 4.12. Para crear un *perfil de ruido* se le da clic en el botón ‘Obtener perfil de ruido’.

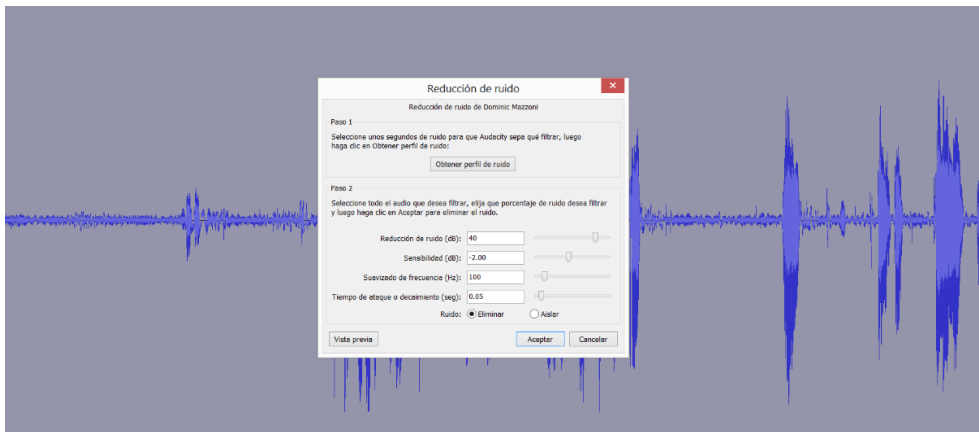


Figura 4.13. Esta figura se puede apreciar como el color gris oscuro se expande sobre toda la señal a diferencia de un reducido segmento como en la Figura 4.11, lo que significa que toda la señal ha sido seleccionada. Dando clic en el botón ‘Aceptar’ se aplica el filtro de ruido creado sobre toda la selección.

El filtrado puede resolver a cierto nivel el problema del ruido ambiental, pero no ayuda con todo el tipo de ruido ambiental. Por ejemplo, ruidos tales como el sonido de los animales como las aves son difíciles de remover, al igual que el sonido que los carros generan al pasar cerca del lugar de la grabación. Todos los segmentos con ruido, ya sea del tipo que no se podía filtrar o que se consideraba que afectaba la calidad de la grabación, fueron removidos de la señal en la etapa de *editado*.

4.5.2 Editado

Debido a la naturaleza de las grabaciones (*field recordings*), existe una gran diversidad de ruido ambiental que no se puede filtrar de manera aceptable. Es por esto que aquellos segmentos con esa clase de ruido fueron silenciados o cortados de la señal.

Técnicamente, silenciar un segmento de la señal consiste en reducir a 0 todos los valores de dicho segmento. Este proceso se puede apreciar en la Figura 4.14.

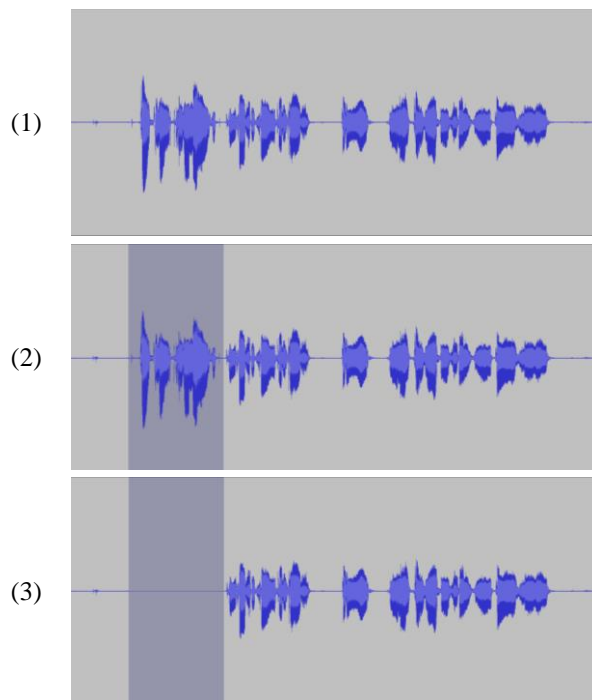


Figura 4.14. Señal de 11s. (1) Señal Original. (2) Selección del segmento cuyo ruido no pudo ser filtrado. (3) Resultado del segmento después de aplicarle el proceso de silencio.

El otro proceso utilizado fue el cortar ciertos segmentos de la señal, aunque esto fue hecho en sólo aquellos casos en donde había segmentos largos sin habla.

4.5.3 Asignación de formato

La asignación del formato fue algo muy simple de decidir ya que el formato estándar que la mayoría de los corpora (por ejemplo, los mostrados en el *LDC*) utilizan es el *wav* para los archivos de las muestras de audio y *raw* para los archivos con las señales de cada una de las muestras de audio.

4.5.4 Asignación de nombre

La idea sobre como la asignación de nombres a las muestras finales del corpus se tomó de [2]. La estructura del nombre de los archivos es la siguiente:

[idioma]-[ID de hablante]-[tipo de habla].wav

Un ejemplo de nombre de uno de los archivos del corpus podría ser:

t-01-lib.wav

En donde los prefijos que pudiese adquirir el primer campo son:

- *t* para el idioma *tének*.
- *n* para el idioma *náhuatl*.
- *x* para el idioma *xi'iuy*.

Los prefijos que pudiese adquirir el *ID del hablante* son números solamente, hasta el momento estos tienen un rango de *[01,50]*, lo que quiere decir que el prefijo sólo puede tener un id entre el *01* y el *50*. Un detalle importante es que este *ID* depende del prefijo del primer campo, lo que significa, por ejemplo, que la persona del archivo '*t-32-lib.wav*' no es la misma que la del archivo '*n-32-lib.wav*'.

Los prefijos que pudiese adquirir el tercer campo, encargado de especificar el tipo de habla, son:

- *esp* para archivos de tipo de habla específica.
- *des* para archivos de tipo de habla descriptiva.
- *lib* para archivos de tipo de habla libre.
- *pd* para archivos de tipo de habla específica que corresponde al dictado de las partes del cuerpo.

4.6 Anotaciones

Debido a la falta de tiempo y de personal, las anotaciones de *Entendámonos* no han sido terminadas al momento en el que se redacta este documento. Asimismo, la estructura de las anotaciones aún no está totalmente definida debido a que 2 de los 3 idiomas indígenas a tratar aún carecen de ciertas características lingüísticas, por ejemplo, el idioma *tének* aún no tiene una fonología totalmente definida y el idioma *xi'iuy* aún sigue trabajando con el establecer un abecedario bien estructurado.

Sin embargo, esto no quiere decir que no se hayan realizado ya algunas anotaciones de prueba. A continuación se mostrarán algunas anotaciones realizadas para los tres idiomas con el fin de mostrar algunas de las ideas que se tomarán como base para definir qué características se utilizarán en las anotaciones. Por el momento, todas las anotaciones que el proyecto dispone han sido hechas utilizando el software *Praat*.

En la Figura 4.15 se puede apreciar 4 niveles de caracterización los cuales son: fonema, palabra, sílabas, número. De los 3, el idioma *náhuatl* es el único que posee una estructura lingüística casi completa, siendo un idioma con un abecedario y estructura gramatical bien definidos. El único inconveniente del idioma es la fonología, la cual aún presenta algunos

detalles con ciertos fonemas, pero, es un inconveniente mínimo ya que este no impide que el equipo de trabajo no pueda realizar las anotaciones.

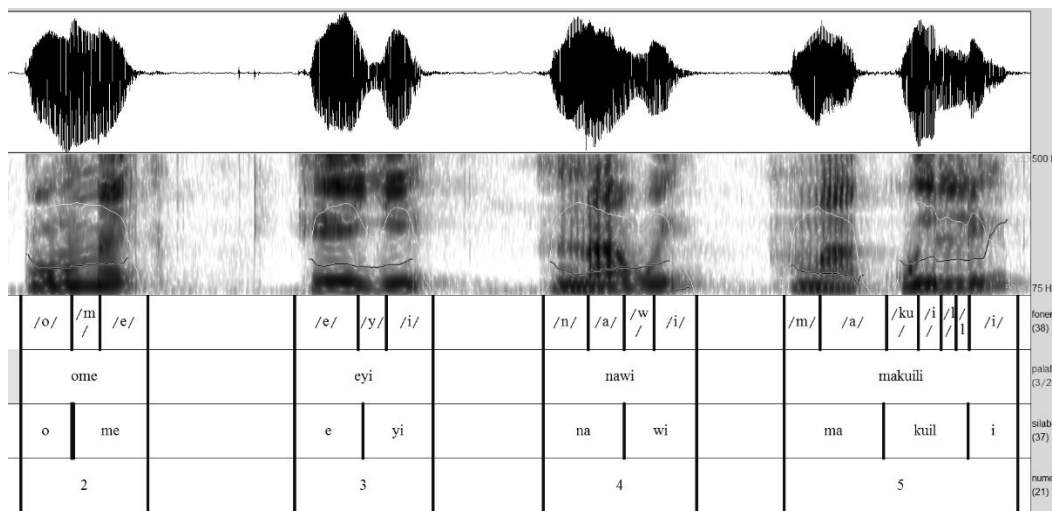


Figura 4.15. Parte de la anotación de una muestra (correspondiente a la pregunta 12) del idioma náhuatl.

Asimismo, la Figura 4.16 muestra una anotación hecha a las partes del cuerpo en el idioma náhuatl etiquetando sólo la palabra en el idioma náhuatl y en el idioma español. La diferencia en usar sólo dos características de etiquetado respecto a la estructura de la Figura 4.15 es que se tiene contemplado realizar un tipo de etiquetado que ayude a los lingüistas a formar un diccionario digital.

Entonces, surgirá la pregunta ¿Por qué no realizar un sólo archivo de anotaciones que contenga todos los niveles de etiquetado contemplados? El hacer un solo archivo que contenga todos los niveles de etiquetado contemplados resulta muy incómodo debido a que las etiquetas se empiezan a empalmar unas con otras, volviendo las anotaciones un trabajo un poco pesado para los lingüistas al momento de hacer el etiquetado.

La Figura 4.17 y 4.18 muestran anotaciones hechas para muestras del idioma tének utilizando distintos etiquetados. El tének es un idioma que todavía no cuenta con sus unidades fonológicas bien definidas, sin embargo ya se tiene un poco de trabajo de investigación acerca de ellas. Además, los lingüistas del IELIIP aún están batallando con la escritura de varias palabras ya que diversas personas de los pueblos tének del estado San Luis Potosí aún tienen problemas para poder escribir en su idioma. La ventaja es que el idioma tének ya cuenta con un abecedario estructurado (propuesto por el IELIIP), por lo que es posible realizar un etiquetado de sílabas acercándose a su lengua hermana náhuatl.

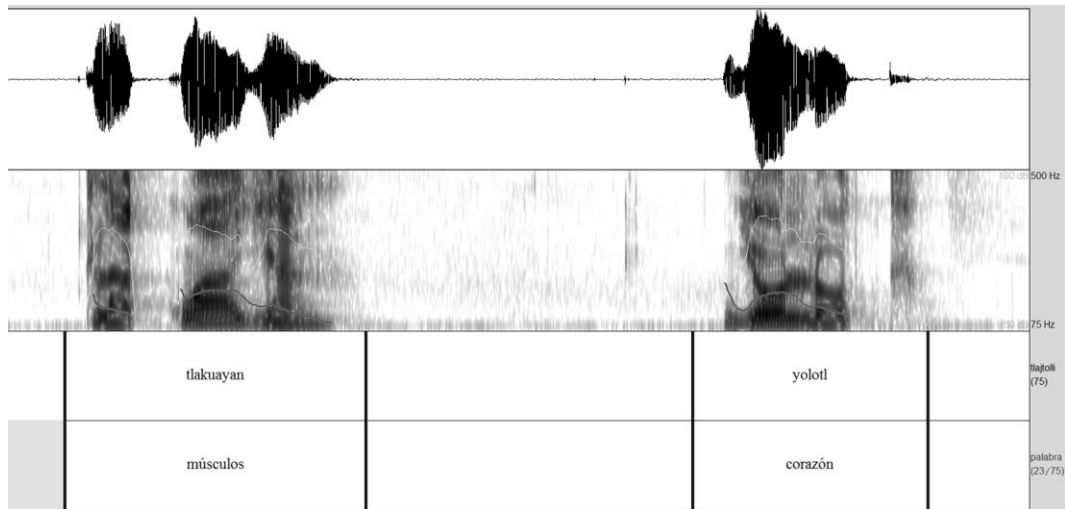


Figura 4.16. Parte de la anotación de una muestra (correspondiente a la pregunta 14) del idioma náhuatl.

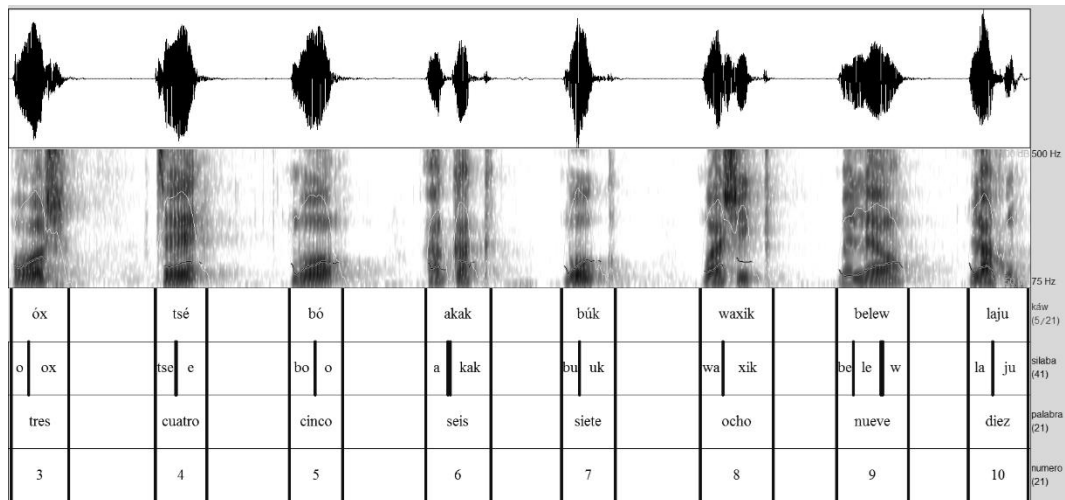


Figura 4.17. Parte de la anotación de una muestra (correspondiente a la pregunta 12) del idioma tének.

La Figura 4.18 muestra una parte de la anotación realizada a una de las muestras de habla libre. Viendo la figura se puede apreciar que la anotación aún no está terminada, y es que las anotaciones hechas para las muestras de habla libre son las que poseen el mayor grado de dificultad. Esto se debe a varias razones, entre las que destacan:

- A veces la persona habla muy rápido e incluso los lingüistas no entienden lo que dijo.
- Debido a que habla muy rápido, las señales de cada palabra están demasiado juntas, lo que dificulta el etiquetado de cada una de ellas.
- En los idiomas tének y xi'iuy, se tiene que ser muy cuidadoso al momento de traducir una palabra, ya que estos idiomas usan la correlación de sus palabras para generar otras. Por ejemplo, en la Figura 4.18, la frase *pat'al abatnom* hace referencia a una computadora.

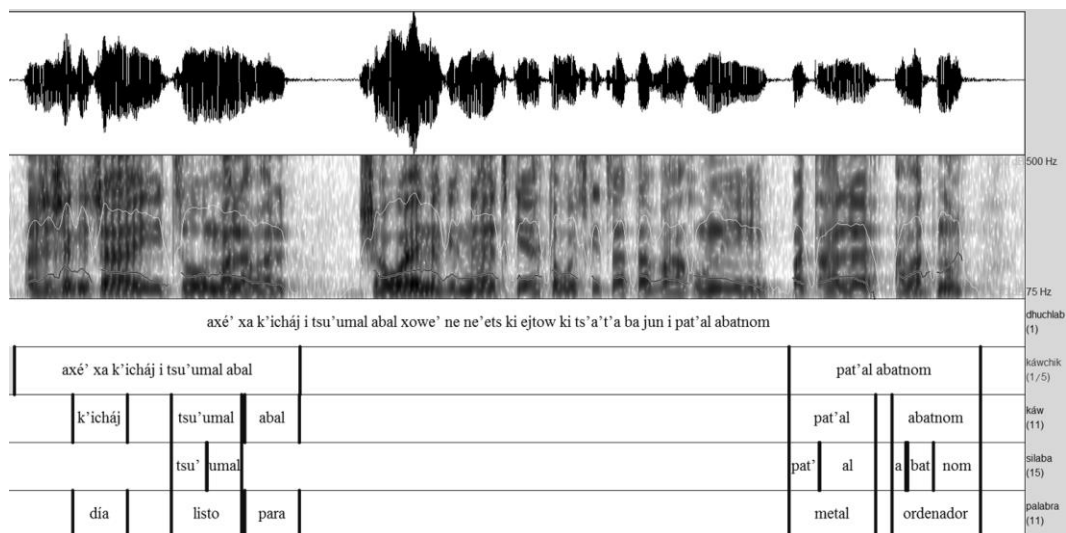


Figura 4.18. Parte de la anotación de una muestra (correspondiente a la pregunta 13) del idioma tének.

El idioma xi'iuy es el idioma que más carece de fundamentos lingüísticos. A pesar de poseer un abecedario propuesto por el IELIIP, la mayoría de los hablantes de este idioma aún tienen discrepancias en algunas sus características e incluso aún se está realizando trabajo de investigación para poder escribir varias palabras. No cuenta con trabajo de investigación respecto a su fonología en ningún nivel y, asimismo es el idioma que presenta una mayor pérdida de hablantes.

La Figura 4.19 muestra la anotación hecha a una muestra del idioma xi'iuy. En ella se utilizaron los mismos niveles de etiquetado que los de la Figura 4.17. El etiquetado de la figura anterior corresponde a la variante sur del idioma xi'iuy. Este idioma presenta otro nivel de dificultad, ya que la escritura de la variante sur y la variante norte es diferente.

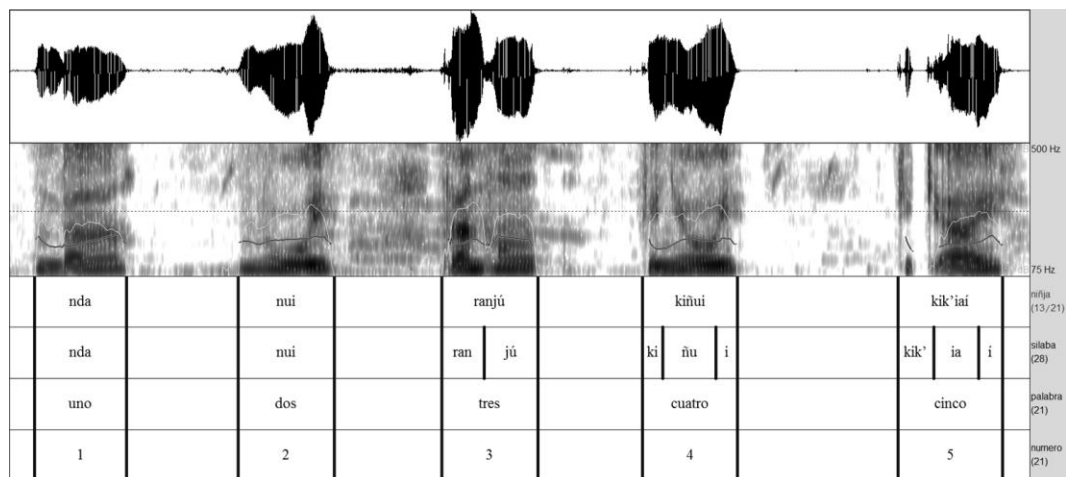


Figura 4.19. Parte de la anotación de una muestra (correspondiente a la pregunta 12) del idioma xi'iuy.

4.7 Metadatos

Los metadatos recolectados para cada sesión de grabación siempre se redujeron al mínimo debido a diferentes razones más que nada sociales, por ejemplo: El hablante no quería colaborar, por lo que la cantidad de tiempo que nos brindaba para la grabación era lo mínimo, entre otras. En la Figura 4.20 se puede apreciar el archivo *xml* que contiene los metadatos de todas las grabaciones recolectadas.

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <Entendámonos>
3   <subhost1>
4     <nl>
5       <nameX' . . . . . ' />
6       <age>38</age>
7       <genre>bc</genre>
8       <sesID>c</sesID>
9       <dateR>120927</dateR>
10      <timeR>1641</timeR>
11      <eCond>El lugar carece de ruido.</eCond>
12      <comns>La primer persona de Entendámonos. La fiesta de la radiodifusora XENUTF aun no comienza. El cuarto de grabacion que nos prestaron carece de ruido. La persona a veces dice algunas palabras en español.</comns>
13    </nl>
14  </subhost1>
15  <subhost1>
16    <nl>
17      <nameX' . . . . . ' />
18      <age>22</age>
19      <genre>MC</genre>
20      <sesID>3</sesID>
21      <dateR>120927</dateR>
22      <timeR>1718</timeR>
23      <eCond>La fiesta de la radio aun no comienza. El lugar está tranquilo.</eCond>
24      <comns>El ruido que se escucha es en general murmullos, afortunadamente nada que afecte en un mayor grado a la grabacion.</comns>
25    </nl>
26  </subhost1>
27  <subhost1>
28    <nl>
29      <nameX' . . . . . ' />
30      <age>49</age>
31      <genre>MC</genre>
32      <sesID>79</sesID>
33      <dateR>130128</dateR>
34      <timeR>1249</timeR>
35      <eCond>La grabacion se realizo en una escuela, afuerita del salon de clases. Hay ambiental.</eCond>
36      <comns>El ruido ambiental no es ni muy periodico ni muy fuerte. La grabacion no se ve afectada.</comns>
37    </nl>
38  </subhost1>
39 </Entendámonos>

```

Figura 4.20. Metadatos de *Entendámonos*.

Hasta el momento, los metadatos utilizados para el corpus son:

- *name*: Nombre completo del hablante.
- *age*: Edad del hablante.
- *genre*: Genero del hablante.
- *sesID*: Id de la sesión de grabación.
- *dateR*: Fecha de la grabación.
- *timeR*: Hora de la grabación.
- *eCond*: Condiciones ambientales del lugar donde se realizó la grabación.
- *comns*: Comentarios acerca de la grabación.

Los 8 metadatos anteriores están dentro del elemento que representa al *ID del hablante*, que a su vez, todos éstos están dentro del elemento idioma. Por lo tanto, si uno desea hacer referencia hacia algún dato en específico, la ruta a seguir estaría dada por los elementos:

```
<idioma>
  <ID del hablante>
    <nombre>
    <age>
    <genre>
    ⋮
    <comns>
```

Los elementos idioma son *nahuatl*, *tenek* y *xiiiuy*, nombrados así para los respectivos idiomas incluidos en el corpus.

El *ID del hablante* es el mismo tipo de dato que se utiliza en la Sección 4.5.4 para la asignación del nombre de las muestras. El único detalle diferente recae en la nomenclatura, la cual tiene la siguiente estructura:

$$[\text{letra del idioma}][\text{ID del hablante}],$$

en donde la letra del idioma corresponde a los caracteres *n*, *t* o *x* y el *id del hablante* corresponde a los valores en un rango de [1,50], por ejemplo: *x9* para el hablante 9 del idioma *xi'iuy*, *n1* para el hablante 1 del idioma náhuatl, *t43* para el hablante 43 del idioma *tenek*, entre otros.

Los elementos *name*, *eCond* y *comns* son datos del tipo *string*, representados por cadenas de caracteres. Los elementos *age* y *sesID* son datos representados por números enteros. El elemento *genre* es un dato representado por los caracteres *h* o *H* en el caso de los hombres, y *m* o *M* en el caso de las mujeres.

El elemento *dateR* es un dato *string*, conformado por una cadena de 6 caracteres (2 para cada bloque), el cual tiene la siguiente estructura:

$$[\text{año}][\text{mes}][\text{día}],$$

por lo que, por ejemplo, la cadena *130124* representaría la fecha *13/01/24*, la cual corresponde al *24 de Enero del 2013*.

El elemento *timeR* es un dato *string*, conformado por una cadena de 4 caracteres (2 para cada bloque), el cual tiene la siguiente estructura:

$$[\text{hora}][\text{minuto}],$$

por lo que, por ejemplo, la cadena *1642* representaría las *16:42 hrs*, o bien, las *4:42 pm*.

4.8 Estructura del Corpus

Entendámonos de distribuirá por medio de un archivo *iso* para facilitar la forma de compartirlo, sin embargo los términos y condiciones para que una persona externa pueda adquirir una copia del corpus queda en manos de la *División de Estudios de Posgrado e Investigación del Instituto Tecnológico de Ciudad Madero*. Hasta el momento, la estructura del corpus consta de los siguientes directorios y archivos:

- *leeme.txt* es un archivo con información sobre: las especificaciones generales y técnicas del corpus, la estructura del corpus y detalles sobre lo que contiene cada directorio del *iso*.
- *voces/* es el directorio que contiene todas las grabaciones en formato *wav* de una forma estructurada.
 - ↳ “*idioma*”/ es uno de los 3 directorios encargados de almacenar subdirectorios con los respectivos archivos de audio de su idioma, los cuales son: *nahuatl/*, *tenek/* y *xiiuy/*.
 - ↳ *especifica/* es el directorio que almacena los archivos en respuesta a información específica.
 - ↳ *descriptiva/* es el directorio que almacena los archivos en respuesta a información descriptiva.
 - ↳ *libre/* es el directorio que almacena los archivos de habla libre.
 - ↳ *partes_cuerpo/* es el directorio que almacena los archivos del dictado de las partes del cuerpo.
- *raws/* es el directorio que almacena las señales de cada archivo del directorio *voces/*. Su estructura es similar al del directorio *voces/* y el formato de archivo utilizado es *raw*, el cual es exportado utilizando el software *Audacity®*.
- *stuff/* es el directorio encargado de almacenar todo tipo de documentación y paquetería de software. En el caso de la documentación, esta se encontrará en formato *pdf*, y entre algunos de los documentos que habrá están: protocolo de grabación, procesos de validación, guía del micrófono.
- *metadatos/* será un directorio paralelo al directorio *voces/*, con la diferencia de que en vez de encontrar archivos de audio *wav*, se encontrará el archivo *xml* cuyo contenido serán los metadatos de su respectivo archivo *wav* del directorio *voces/*.

5

Evaluación

La forma de comprobar las cualidades de un corpus consiste en utilizarlo para realizar tareas específicas con el fin de evaluar sus resultados. La tarea específica en este caso es la *Identificación Automática del Lenguaje Hablado*, para la cual fue desarrollado el corpus *Entendámonos*.

El sistema LID desarrollado está basado en la metodología propuesta en [8], la cual hace un análisis acústico de las muestras utilizando la transformada wavelet para extraer las características de las bajas frecuencias. El uso de esta metodología se justifica con el nivel de desarrollo en el que se encuentra *Entendámonos* debido a que aún carece de una cantidad suficiente de anotaciones. El desarrollo del sistema LID propuesto comenzó antes del proceso de recolección de muestras de *Entendámonos* debido a que se contempló la falta de tiempo que el proyecto presentaría al final. El sistema LID primero fue probado con el corpus OGITS antes de usarse con *Entendámonos*, estos resultados se pueden encontrar en [24].

5.1 Metodología

Tomando como base la estructura general propuesta en la Figura 2.4, la estructura del sistema LID propuesto está definida en la Figura 5.1.

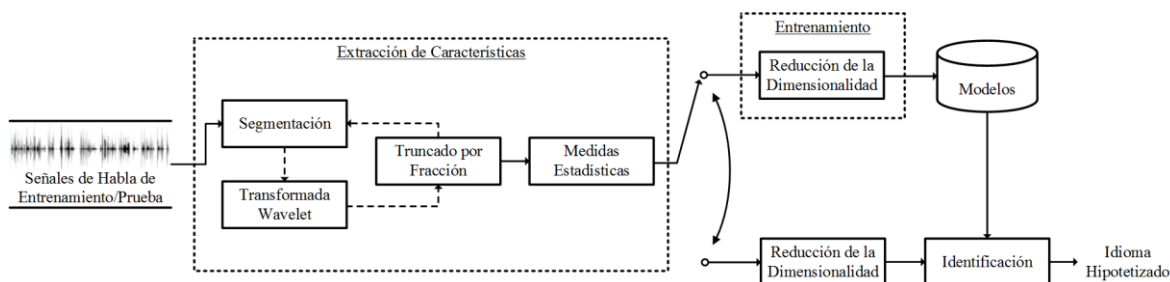


Figura 5.1. Estructura del sistema LID propuesto.

5.1.1 Señales de Haba de Entrenamiento/Prueba

El bloque representa las muestras, las cuales son señales de audio contenidas en archivos *wav*. Todas las muestras pasan por el bloque de *extracción de características* y de ahí pueden pasar a la fase de *entrenamiento* o la fase de *prueba*. Las muestras dirigidas a la fase de entrenamiento son aquellas que formaran los *modelos*, los cuales son los que se utilizan para *identificar* el idioma de cualquier otra señal. Las muestras dirigidas a la fase de prueba son aquellas cuyo idioma hablado se quiere identificar.

5.1.2 Segmentación

Esta parte consiste en dividir cada señal en segmentos de $1s$ (Figura 5.2). Cada señal es representada como un arreglo de tamaño equivalente a la duración (medida en segundos) de su respectivo archivo *wav* por su frecuencia de muestreo τ , por ejemplo, una señal de $10s$ de duración con una frecuencia de muestreo de 8000 Hz tiene un arreglo de 80000 valores de tamaño. De tal modo, la segmentación toma el valor en las posiciones del arreglo correspondientes al segundo que está siendo segmentado.

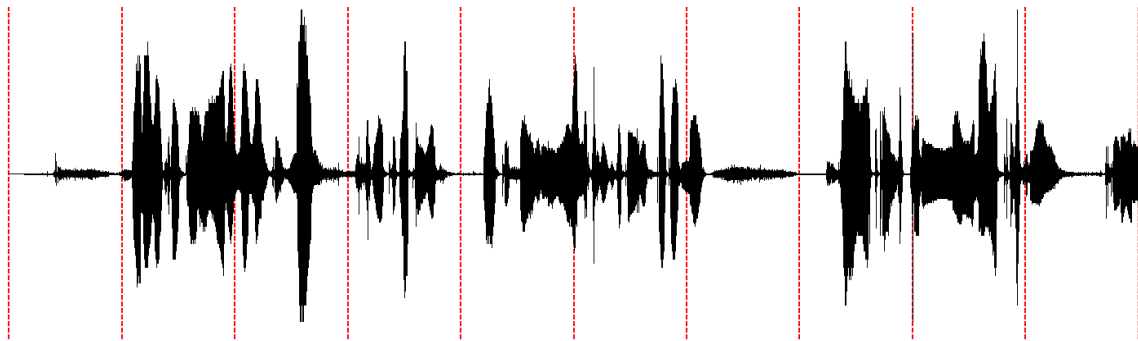


Figura 5.2. Señal de $10s$, segmentada en bloques de $1s$. Si la frecuencia de muestreo es de $\tau\text{ Hz}$, esto significa que cada bloque (segundo) tiene τ valores.

5.1.3 Transformada Wavelet

Esta parte aplica la *transformada wavelet Db4* con η niveles de *descomposición* a los arreglos de un $1s$ obtenidos en la segmentación, lo que nos retorna una señal dividida en dos partes iguales, la primera parte la conforman los coeficientes cA (*coeficientes de aproximación*) y la segunda parte los coeficientes cD (*coeficientes de detalle*), esto significa que el tamaño de $cA=cD$. Este proceso se puede apreciar en la Figura 5.3.

Cuando $\eta=1$, la función aplica la transformada sobre cada arreglo que se obtiene a través de la segmentación. Cuando $\eta>1$, la función aplica la transformada sobre los coeficientes cA resultantes de cada nivel de descomposición. Este proceso de descomposición se explica en la Figura 5.4.

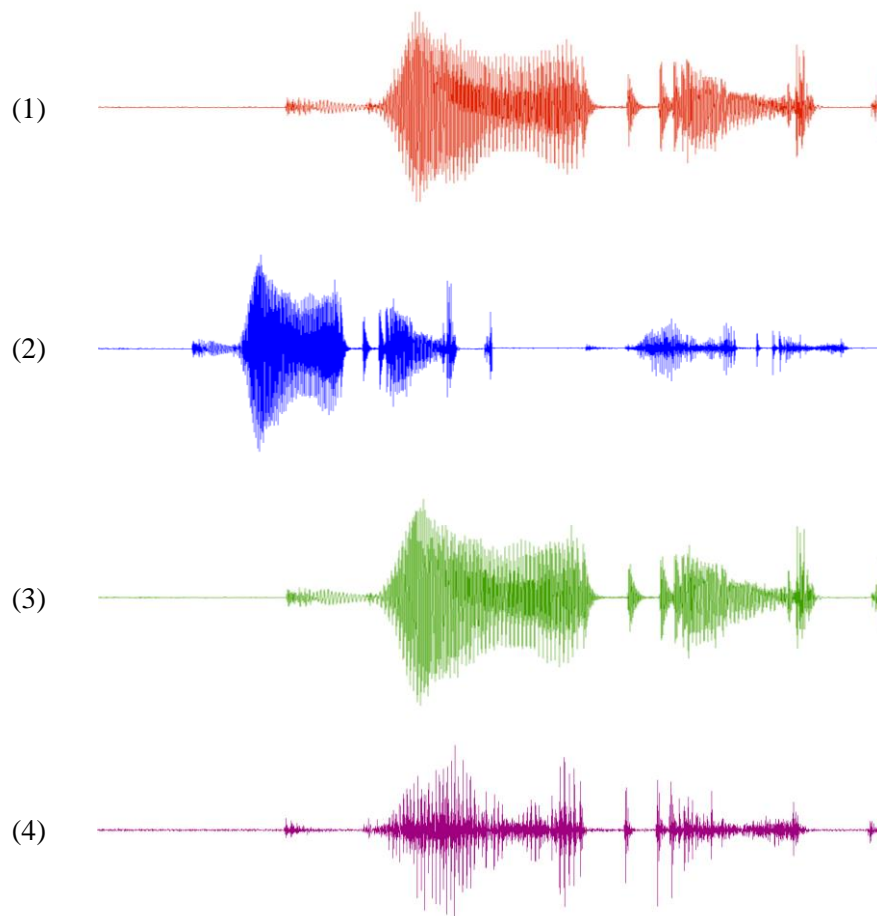


Figura 5.3. (1) Señal correspondiente al octavo bloque de la Figura 5.2. (2) Señal resultante del proceso de la transformada wavelet con $\eta=1$, conformada por los coeficientes cA y cD . (3) Señal que representada por los coeficientes cA . (4) Señal representada por los coeficientes cD .

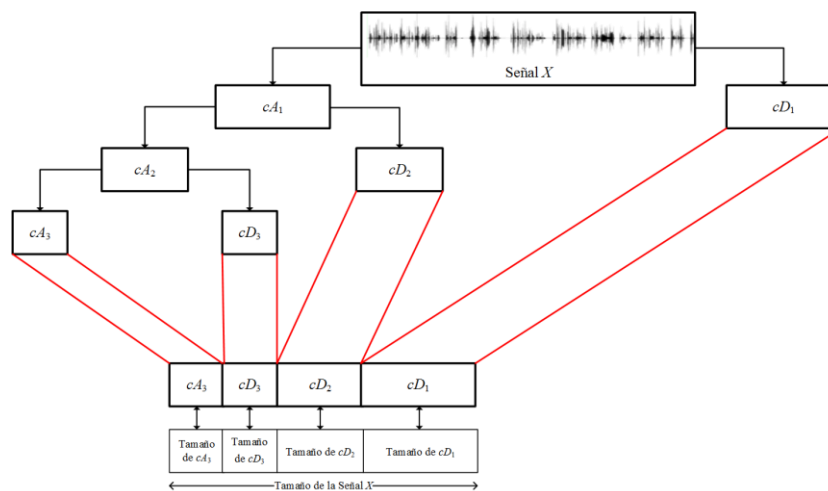


Figura 5.4. Proceso de descomposición de la transformada wavelet con $\eta=3$.

5.1.4 Truncado por Fracción

Esta parte efectúa un truncado sobre los arreglos conformados por los coeficientes cA , el cual conserva una fracción establecida del conjunto original de los valores inalterados de cada arreglo, mientras que los demás se establecen a 0. De los coeficientes cA resultantes, se eligen los que tienen mayor magnitud (Figura 5.5). Según los resultados obtenidos de diversas pruebas (en las cuales se varia el porcentaje de truncado), se ha visto que la mejor opción ha sido usar truncados entre el 1% y 5%, permitiéndonos usar sólo el porcentaje de los coeficientes más representativos de las bajas frecuencias.

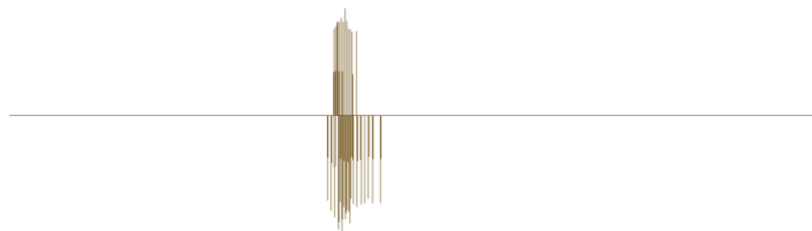


Figura 5.5. Truncado por fracción del 1% aplicado a la tercera señal de la Figura 5.3 (señal verde).

5.1.5 Medidas Estadísticas

Incluso utilizando el truncado, la problemática del manejo de la gran cantidad de coeficientes persiste, lo cual nos lleva a enfrentarnos con la dimensionalidad de los datos [8]. Esta parte se encarga de calcular las medidas estadísticas a una matriz A de tamaño $\Gamma \times \Delta$, la cual contiene los valores de los arreglos resultantes del proceso de truncado (Figura 5.6), donde Γ es igual al análisis en segundos que se va a realizar sobre las muestras y Δ es igual al tamaño los arreglos cA . El uso de medidas estadísticas sobre los coeficientes cA truncados es una manera de reducir su cantidad y a la vez describirlos, caracterizando la señal y estableciendo así nuestra manera de reducir la dimensión, capturando el contenido original de los datos. Las medidas utilizadas en este trabajo son: *media*, *desviación estándar*, *máximo*, *mínimo*, *sesgo* y *curtosis*.

5.1.6 Entrenamiento y Modelos

Una vez que las medidas estadísticas han sido calculadas, se construyen las instancias dependiendo de Γ . Cada muestra analizada genera cierto número de instancias y cada modelo contiene únicamente información de 2 idiomas (razón por la cual en el resumen se menciona como identificación mediante torneos, ya que cuando una instancia quiera categorizarse respecto al modelo, sólo habrá un ganador). En un principio, cada modelo contiene cierto número de instancias que después se irán removiendo dependiendo de qué tan significativa

sea la instancia para el modelo. Cada instancia posee $\mu \times \Delta$ atributos, donde μ representa el número de medidas estadísticas usadas y Δ el tamaño del arreglo cA .

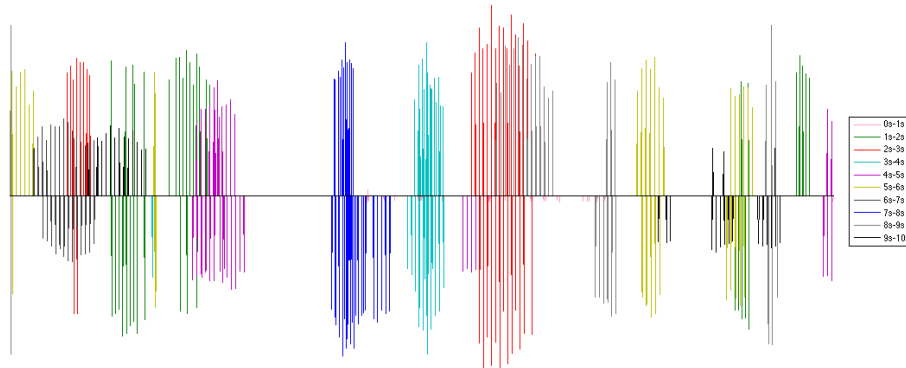


Figura 5.6. Grafica de una matriz A . La grafica de color azul, correspondiente al segmento $7s-8s$ es aquella que fue calculada en la Figura 5.5, asimismo, esta matriz contiene todos los arreglos truncados de cada bloque de la Figura 5.2.

A pesar de la reducción mediante el truncado y medidas estadísticas, las instancias a clasificar son demasiado grandes (p. ej. $24,006$ atributos por instancia para $\Gamma=10s$). El proceso de *reducción de la dimensionalidad* se encarga de aplicar la técnica de *ganancia de información* para seleccionar sólo los atributos más significativos de cada modelo. Esta técnica funciona como filtro, cuyo umbral elegido para filtrar fue de 0 , lo que indica que aquellos que tengan ganancia de información 0 o debajo de 0 serán eliminados.

Una vez realizado lo anterior, se procede a entrenar el modelo, para hacer esto utilizamos la validación cruzada, el cual es un método que divide un conjunto de datos en subconjuntos iguales, donde cada subconjunto es clasificado con el resto, lo que permite clasificar un conjunto completo sin tener que dividirlo o tener que analizar sólo una parte. La configuración para la validación cruzada fue de 10 subconjuntos (*folds*) utilizando el clasificador *Naive Bayes* debido al rendimiento presentado en [10].

Una vez concluido el entrenamiento se prosigue al bloque de *Modelos*, que consiste en escribir los modelos en la memoria para que estos puedan ser cargados cada vez que se necesite realizar el proceso de *identificación*, sin la necesidad de tener que realizar de nuevo el proceso de entrenamiento.

5.1.7 Identificación

El bloque de identificación se encarga de identificar el idioma hablado de alguna muestra de prueba. Previo al proceso de identificación se realiza un proceso de ganancia de información a sus instancias similar al realizado en la fase de entrenamiento. Cada instancia de extraída de la muestra de prueba es comparada con cada modelo, devolviendo un porcentaje de *likelihood*, el idioma que obtenga una *likelihood* mayor será el *idioma etiquetado* de su

respectivo modelo. Para este caso particular, *Entendámonos* sólo cuenta con 3 idiomas, por lo que sólo puede haber 3 modelos posibles: *náhuatl-tének*, *náhuatl-xi'iuy* y *tének-xi'iuy*. La *likelihood* que sea etiquetada 3 veces será el *idioma hipotetizado* del sistema LID.

5.2 Pruebas y Resultados

Una vez que se terminó la etapa de *post-procesamiento* (Sección 4.5) se realizaron 5 pruebas para evaluación del corpus. Lamentablemente, el proceso de recolección de muestras del corpus fue muy tardado, e independientemente de las limitaciones que dicho proceso presentó, se considera que aún falta mucho trabajo por elaborar respecto a la LID aplicada a *entendámonos*. Estas pruebas consistieron en análisis de *3s* y *10s* con niveles de descomposición (η) wavelet *1* y *2*, y sus respectivos casos de identificación.

Las pruebas se desarrollaron sobre un conjunto de *40* muestras por idioma (*120* muestras en total), la duración de cada muestra es de mínimo *50* segundos, y la frecuencia de muestreo de cada muestra es de *44100 Hz*, se omitió realizar un *downsample* al conjunto de muestras debido a que se quería trabajar con la mayor cantidad de información posible.

Además es importante mencionar algunas características de las pruebas hechas. Lo primero es la cantidad de información que las muestras contienen, en los reportes de actividades del proyecto se presenta una prueba piloto la cual expone el incremento abrupto de coeficientes wavelet (*22051* en contraste a los *4001* que se obtenían de las muestras del corpus *OGITS*) con $\eta=1$, debido a la frecuencia de muestreo, lo que genera *132306* coeficientes por instancia si se utilizan las *6* medidas estadísticas propuestas. Con $\eta=2$, la cantidad de coeficientes se reduce a *11027* y a la vez se mejoran los resultados en algunos casos.

Además, la cantidad de instancias que se obtienen por idioma ha incrementado, obteniendo *640* instancias para la prueba de *3s* y *200* instancias para la de *10s*, dando *1280* instancias para los modelos de la primera prueba y *400* para los modelos de la segunda. En adición, lo anterior ha causado un incremento considerable en el tamaño de los modelos llegando a tener modelos que pesan *1.8Gb*, necesitando que *Java* requiera alrededor de *4Gb* de memoria asignada para poder realizar los procesos de clasificación.

En las siguientes tablas se puede ver los resultados de los porcentajes arrojados por el proceso de validación cruzada de la fase de entrenamiento de las pruebas.

Tabla 5.1. Prueba de 3 segundos con $\eta = 1$.

	Tének	Xi'iuy
Náhuatl	79.6	75.9
Tének		77.7

Tabla 5.2. Prueba de 3 segundos con $\eta = 2$.

	Tének	Xi'iuy
Náhuatl	75.5	75.7
Tének		77.8

Tabla 5.3. Prueba de 10 segundos con $\eta = 1$.

	Tének	Xi'iuy
Náhuatl	96.2	80.2
Tének		90.0

Tabla 5.4. Prueba de 10 segundos, con $\eta = 2$.

	Tének	Xi'iuy
Náhuatl	95.5	83.2
Tének		95.0

Por ejemplo, para la Tabla 5.2, sus resultados nos dicen que el 75.5% de las instancias del modelo náhuatl-tének fueron clasificadas correctamente, o bien, su idioma hipotetizado fue correcto. Hay que recordar la cantidad de información que se está analizando, por ejemplo, ese 75.5% de la Tabla 5.2 representa a 966 instancias clasificadas correctamente de un total de 1280, asimismo, el 95.0% de la Tabla 5.4 representa a 380 instancias clasificadas correctamente de un total de 400.

Ahora bien, las instancias que no fueron clasificadas correctamente fueron removidas de los modelos y posteriormente estos pasan a ser escritos en la memoria. Una vez hecho esto, se procede a trabajar en la fase de prueba, la cual consiste en identificar el idioma de las muestras restantes de cada idioma.



Occāsus

Esta parte se encarga de recopilar las secciones finales en toda tesis: las conclusiones, las aportaciones de la investigación y el trabajo futuro. El ocaso que muestra este proyecto no es más que el comienzo de una investigación que podría solucionar muchos de los problemas de las lenguas indígenas mexicanas, particularmente las del estado de *San Luis Potosí*.

6.1 Conclusiones

Realizar un corpus de habla de lenguas indígenas mexicanas no es una tarea sencilla, sino más bien un tanto ardua. Los beneficios que un corpus de habla ofrece a los idiomas indígenas no son únicamente de conservación del idioma, sino que entre estos, se encuentra la introducción de los idiomas indígenas al mundo de las tecnologías del habla e información en general. Como ejemplos están [17], [18] y el presente trabajo; las lenguas indígenas mexicanas son una caja de pandora esperando ser abierta.

Un detalle importante en el desarrollo de un corpus de esta índole, se encuentra en el proceso de recolección de muestras. Por ejemplo, el hacer un único cuestionario para las lenguas indígenas es una característica absurdamente limitada, un cuestionario elaborado específicamente para cada idioma y variante del mismo sería lo más correcto para realizar las grabaciones, además de que se podría extraer información con un contenido lingüístico más rico.

Otro detalle importante es la manera de acercarse a la mayoría de la gente de este tipo de habla para que colaboren donando su voz. Desafortunadamente, para muchas de estas personas la satisfacción de colaborar en un proyecto que ayude a estudiar y preservar su idioma, es opacada por la necesidad económica; muchas personas requieren un incentivo meramente monetario para acceder a donar su voz. Afirmar que ésto aplica para la mayoría de las personas sería un tanto trillado, sin embargo el considerar una estrategia económica o de propaganda previa a la fase de recolección de muestras sería una gran ayuda.

El estudio que se hizo al momento de la selección del micrófono vocal (*Shure PG42-USB*) se considera un éxito. Realizar las grabaciones con este tipo de micrófonos tiene la ventaja de obtener grabaciones cuya calidad es extraordinaria a pesar de que el hablante no mantenga una distancia constante respecto al diafragma del micrófono al momento de realizar la grabación. El considerar realizar un corpus utilizando grabaciones telefónicas es una opción que ayudaría a solucionar ese detalle, sin embargo se debe considerar que en muchas de las comunidades y pueblos indígenas de *San Luis Potosí* no tienen servicio telefónico o red telefónica.

Respecto a los idiomas náhuatl, tének y xi'iu, se tiene la concepción de que son lenguas que carecen de una estructura lingüística bien definida, lo que es una mentira. Lo que es una realidad es la falta de conocimiento verbal y escrito en una cantidad muy amplia de personas que hablan (aunque sea a un nivel muy limitado) uno de estos idiomas. Por ejemplo, muchas personas hablantes tének no saben escribir lo que dicen, o bien, personas que estudian el idioma no saben cómo pronunciar muchas palabras. Esto confirma que el desarrollo de *Entendámonos* es necesario puesto que puede ser utilizado con fines de enseñanza, específicamente se podría empezar a utilizar las muestras con las anotaciones de las partes del cuerpo.

Pasando a la implementación del sistema LID, el análisis de las bajas frecuencias como forma de identificar el idioma habla es un área poco investigada, sin embargo los resultados presentados en esta tesis son alentadores ya que un 72% de identificación no es tan malo considerando el hecho que se omite el uso características más estructuradas como los fonemas y *phonotactics* explicadas en la Sección 2.1.2.

6.2 Aportaciones de la Investigación

La principal aportación de esta tesis es el desarrollo de *Entendámonos*, el cual es la primera colección de muestras de habla diseñada con el propósito de formar parte de un corpus de habla.

Además, el proceso de elaboración de un corpus también contempla, a cierto nivel, investigación sobre los idiomas que lo conformaron, en este caso, la gramática y las historias. El contenido hablado de las muestras posee alto valor gramatical y una vez que este se vea reflejado en las anotaciones podría considerarse un gran aporte a la lingüística de los 3 idiomas de *Entendámonos*. En el caso de las historias, principalmente aquellas que las personas decidieron relatar como habla libre, podrían pasar a formar parte de algún documento de un carácter más narrativo, por ejemplo [25] o [26].

Al ser un proyecto pionero en su campo, la estructura ejercida para el corpus puede utilizarse como base para trabajos futuros que se pretendan desarrollar. Asimismo, la investigación

presenta una *identificación automática del lenguaje hablado* para 3 lenguas indígenas mexicanas.

Esta investigación, además de concretar el análisis de las bajas frecuencias de la señal de voz, también desarrolló otro sistema LID que realizaba un análisis espectral sobre la señal, segmentándola en partes muy cortas y utilizando como característica discriminadora los *coeficientes cepstrales en la escala de Mel* y *máquinas de soporte vectorial* como técnica de aprendizaje automático. Sin embargo este sistema, así como sus resultados, se omitieron en esta tesis debido a la falta de tiempo para refinar su estructura.

Este proyecto de investigación concursó en el *MIT Techonology Review Innovadores Menores de 35 México 2012*, el cual es un concurso cuyo objetivo es premiar la innovación (el desarrollo de nuevas tecnologías o la aplicación creativa de las ya existentes para resolver los problemas actuales), el ingenio y los avances sobre asuntos que preocupan a nivel mundial.

Por el momento se han publicado tres artículos respecto al sistema LID implementado con el corpus OGITS.

6.3 Trabajo Futuro

Debido a que sólo se contó con un año para el desarrollo de *Entendámonos*, el corpus como tal requiere todavía de mucho trabajo. A continuación se enlistan varias ideas que se consideraron en la investigación.

6.3.1 Anotaciones

Como el autor de este proyecto, considero que el primer paso para mejorar el corpus es comenzar por buscar personas o instituciones que gusten trabajar con las anotaciones restantes del corpus, que desafortunadamente, son muchas. El culminar las anotaciones de la mayoría de las muestras, le abriría a el corpus las puertas a varias otra ramas de las tecnologías del habla. Por ejemplo, una vez que las anotaciones sean concluidas, *speech recognition* es una rama que podría elaborar mucho trabajo con las partes del cuerpo, los números y las palabras más utilizadas del corpus.

6.3.2 Número de Hablantes

Los corpora que pueden encontrarse en páginas como la del LDC, nos muestran colecciones de corpora cuyo contenido hablado es inmenso, y en su mayoría se debe a la cantidad de hablantes que donaron su voz.

Uno de los problemas con los que se podría empezar a trabajar, sería el tratar de emparejar la cantidad de hombres y mujeres que tiene el corpus, ya que por ejemplo, las muestras del idioma xi'uiy están conformadas por mujeres en su gran mayoría. Esta desproporción hace que los resultados de ciertas tecnologías del habla no sean tan finos.

Para que *Entendámonos* se torne un corpus de habla fundamental en el uso de las tecnologías de habla, el estado del arte nos muestra que debe contar con alrededor de 150 hablantes por idioma.

6.3.3 Metadatos

La información personal que se le pedía a cada hablante fue mínima, debido a que en varios casos los hablantes se comportaron difíciles al momento de colaborar. Sin embargo, sería prudente considerar pedir más información personal de cada hablante, esto con el fin de elaborar un trabajo lingüístico más fino. La sección *Metadata* de [1] muestra ejemplos de otros tipos de información adicional que se podría solicitar al hablante previo al momento de realizar la grabación.

6.3.4 Tecnologías del Habla

La Figura 2.1 nos muestra diferentes tareas del procesamiento digital de la voz, una de las más comunes es el *reconocimiento del habla* y es una de las tareas que podría brindar un acercamiento popular al campo de la investigación debido a las tendencias actuales de las tecnologías de habla como los son los famosos reconocedores *Siri* de la compañía *Apple Inc.* y *Google Now* de *Google*. Sin embargo, como se menciona en la sección 6.3.1, lo primero es que avanzar con el trabajo restante de las anotaciones de *Entendámonos*, ya que sin ellas, sus aplicaciones serían bastante limitadas.

A

Marco Teórico

Este apéndice se encarga de recabar todos los conceptos teóricos que se utilizaron o que se consideran necesarios para entender todos los tópicos tratados en los Capítulos del presente trabajo.

A.1 Desarrollo de un Corpus de Habla

Esta sección describe el proceso completo que conlleva el desarrollo o producción de un corpus de habla de una manera un tanto cronológica. La Figura A.1 muestra los principales pasos del proceso de desarrollo.

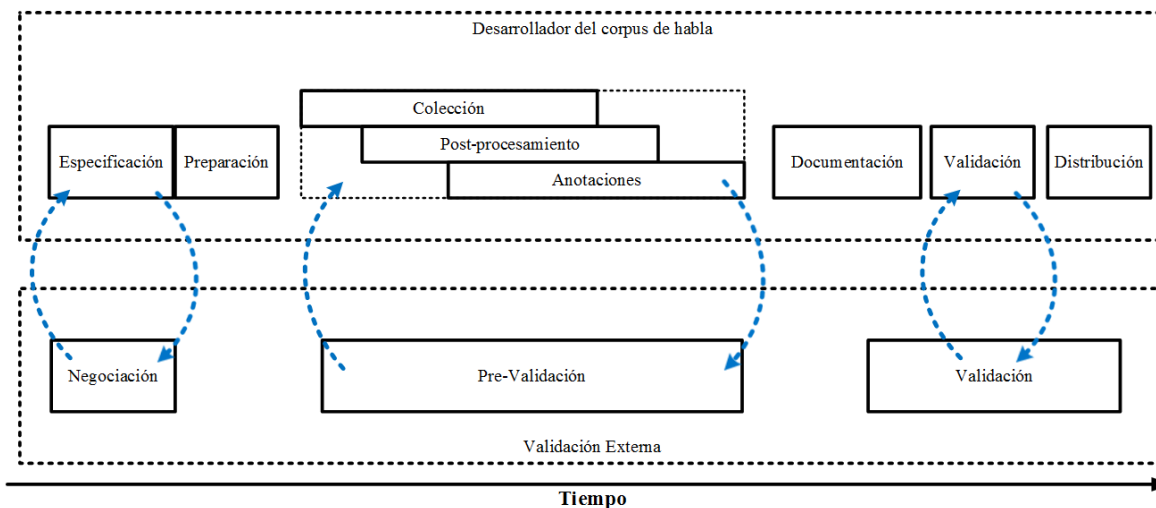


Figura A.1. Proceso de desarrollo típico de un corpus de habla.

Como se puede ver en la figura anterior, algunos pasos tienen un orden estricto porque ellos dependen de los resultados o información producida en el paso anterior, mientras otros pueden ser llevados a cabo en paralelo. Por ejemplo, no tiene caso realizar la etapa de post-

procesamiento si no se ha terminado de recolectar las muestras, aunque, esto no quiere decir que algunos proyectos no realicen varios de los pasos de forma paralela para ahorrar tiempo.

Otro aspecto que vale la pena mencionar de la Figura A.1, es el concepto ideal de las validaciones externas de al menos dos validaciones realizadas por una alguna institución. Aunque, es importante aclarar que no es tampoco un requerimiento fundamental ya que muchos proyectos no cuentan con los fondos o medios para que una institución u organismo externo pueda colaborar con el desarrollo del corpus realizando validaciones.

A.1.1 Especificaciones del Corpus

Al igual que con todos los proyectos que requieren grandes esfuerzos en términos de carga de laboral, el dinero es absolutamente esencial para iniciar la producción o desarrollo de corpus de habla con una especificación detallada de todas sus características deseadas, sus procedimientos, su control del proceso y su validación final. Si se está actuando como contratista, ésta es probablemente la fase del proyecto, donde tendrá el mayor contacto con el cliente. Es muy importante fijar todas las especificaciones por escrito (en su mayoría en forma de un anexo técnico al contrato) y que su cliente firme el presente anexo y todas las modificaciones posteriores.

Los altos costos de producción corpus del habla pueden ser explotados de manera tan óptima mediante la especificación de muchas características diferentes en una colección de habla como sea posible. Por ejemplo, los costos totales de un corpus basado en habla telefónica con el objetivo de reconocer los dígitos y los números no aumentan dramáticamente con algunas grabaciones adicionales sin guion o incluso con grabaciones de habla espontáneo dentro de las mismas sesiones de grabación. Sin embargo, la reutilización del corpus será mucho mayor que con un corpus que sólo contiene leer dígitos y los números.

Las siguientes secciones proporcionan una visión general acerca de los requisitos básicos de cualquier especificación discurso corpus. Puede haber puntos adicionales a cubrir en las especificaciones ya que esto es en función de la naturaleza del corpus de habla a desarrollar.

En este capítulo, los siguientes términos se usan con frecuencia:

- Señal de audio: información binaria digitalizada de audio, por ejemplo, archivos *wav*, *aiff*, *raw*, *flac*, etc.
- Texto: información textual libre de formato, a menudo referido como información *ASCII*.
- Texto de marcado: los datos de texto que contienen símbolos de marcador de un vocabulario cerrado con una sintaxis determinada, por ejemplo, *xml* o *html* de texto.
- Texto con formato: texto con instrucciones de formato tipográfico, a menudo en formato binario, por ejemplo, *Microsoft Word* o *PDF*.

El texto de marcado es el tipo más flexible de texto, ya que puede ser leído por ambas máquinas y los seres humanos, debido a que hace cumplir las limitaciones mínimas de consistencia, y debido a que es independiente de la plataforma y el software.

A.1.1.1 Perfil de los Hablantes

Un corpus de habla consiste en grabaciones de personas hablando. Por lo tanto, lo primero a especificar son las características y las distribuciones de estos hablantes. Es de gran importancia que las características documentadas del hablante sean tan elaboradas como sea posible. A pesar de que estos detalles no parezcan interesantes o relevantes al momento en donde el hablante está siendo grabado, su importancia surge inevitablemente después. Por otra parte, un corpus de habla bien documentado también puede ser usado para fines de investigación, por ejemplo, investigación sociológica. Algunas características importantes (en orden de importancia) son:

- Distribución de sexo. En la mayoría de los casos es 50:50.
- Distribución de edad. Por ejemplo:
 - Por encima de los 16 y menores de 50 años.
 - Distribución equitativa en los siguientes conjuntos: 12-22, 23-30, 31-40, 41-55.
 - Menores de 12 años.
- La lengua materna. Aunque la mayoría de las corpora implican hablantes nativos de una lengua determinada, es conveniente mencionarla en las especificaciones. También se recomienda especificar el porcentaje máximo de hablantes no nativos, por ejemplo:
 - Corpus idioma: francés
 - Porcentaje máximo de hablante no nativos: 5%
- Distribución dialectal. Se puede dar el caso de que un corpus deba cubrir una cierta distribución de un número de dialectos clasificados de un idioma. En general, es muy difícil controlar la afiliación dialectal de los hablantes. La mayoría de los hablantes tienen una preconcepción muy rígida de que dialecto están hablando (la mayoría no sabe lo que el término dialecto significa). Asimismo, incluso los expertos a menudo no están de acuerdo con ciertos rasgos dialectales y por lo tanto es muy difícil validar características *como el 10% de los hablantes del corpus están hablando Québécois*. Algunas recomendaciones prácticas son:
 - En vez de utilizar la etiqueta clase dialectal, utilizar lugares como lugar de escuela primaria y secundaria. En la mayoría de los casos los hablantes mantienen el dialecto adquirido durante el periodo en el que cursaron los grados escolares mencionados. Otra etiqueta recomendable sería utilizar los lugares, regiones, distritos, ciudades, etcétera de donde son los hablantes, aunque para estas etiquetas, también es recomendable especificar la procedencia de cada hablante.
 - Realizar una clasificación dialectal en la etapa de post-procesamiento. Esto requiere un experto en dialectos y algo de tiempo (más costos).

- Realizar una recolección usando el los medios locales, por ejemplo, los noticieros, periódicos o radios locales.
- Educación/Capacidad/Profesión. Algunos corpus de habla requieren ciertos factores sociales como ciertas capacidades (experto en informática, técnico en computación), un nivel mínimo o una distribución de diferentes niveles de educación (escuela primaria, escuela secundaria, preparatoria, universidad) o incluso los hablantes de una determinada profesión (radiólogo, locutor de noticias). Asegúrese de sólo especificar ciertas características si no tiene un proceso de recolección bien definido.

Otros posibles factores pueden ser: *patologías, acento de fuera o extranjero, ritmo de habla*, etc.

A.1.1.2 Número de Hablantes

El número de hablantes es una de las características más importantes de un corpus de habla. Las corpora de habla se pueden dividir en tres categorías:

- Corpora de habla con 1 a 5 hablantes: se utilizan a menudo en el desarrollo de sistemas de síntesis de voz o para investigación básica, por ejemplo, cuando se deben hacer mediciones invasivas.
- Corpora de habla con alrededor de 5 a 50 hablantes: se utilizan a menudo en la investigación experimental factorial. En general, el número de hablantes y el número de repeticiones de los fenómenos del habla que se investigan deben ser lo suficientemente grande para un procesamiento estadístico significativo si se han previsto diseños experimentales factoriales.
- Corpora de habla con más de 50 oradores son las necesarias para entrenar y probar adecuadamente el reconocimiento de voz, reconocimiento del hablante o identificación del idioma.

Algo no menos importante es que debe tener en cuenta que un pequeño número de hablantes no significa, necesariamente, un pequeño corpus.

A.1.1.3 Contenido Hablado

El contenido hablado de un corpus de habla es otra característica importante que determina el posible uso del recurso. Por supuesto, esta característica no es totalmente ortogonal a otras especificaciones, por ejemplo, el tipo de habla. Básicamente, hay cuatro enfoques principales que definen el contenido hablado de un corpus: vocabulario, dominio, tarea o distribución fonológica. Estos pueden ser aplicados de una manera combinada en algunos casos.

A.1.1.3.1 Vocabulario

Probablemente la forma más sencilla de especificar el contenido hablado es con el vocabulario, ya que, más o menos, deriva automáticamente la intención del uso del corpus.

Por ejemplo, si se utiliza el corpus para entrenar a un reconocedor de voz en 11 dígitos alemanes y tres palabras de comando, entonces, la definición de contenido más probablemente requerirá una distribución igual para todos los 14 elementos del vocabulario y sus repeticiones por hablante, por ejemplo, en este caso serían 14 palabras de 500 hablantes de 10 repeticiones daría un total de 70,000 elementos.

A.1.1.3.2 Dominio

Otro método para controlar el contenido hablado de un corpus habla es por dominio. El dominio en este contexto significa el tema, campo de los temas o la situación en la que una comunicación verbal toma lugar. El dominio podría ser, por ejemplo:

- Tiempo.
- Restaurantes de la Ciudad de México.
- Los discursos en la cámara de senadores.
- Cuentos de hadas.
- El programa de televisión de anoche.

Aunque el vocabulario exacto no puede ser determinado por este método, es un buen método para lograr un vocabulario más bien cerrado, sin restringir demasiado a los hablantes.

A.1.1.3.3 Tarea

Al instruir al hablante para resolver una determinada tarea (ya sea junto con uno o más compañeros de diálogo humano(s) o con una máquina virtual en un experimento de WOZ), el contenido hablado de un corpus de habla puede ser reducido a unos pocos cientos de palabras sin el problema de que los hablantes se sientan limitados por la situación. Una vez más, la instrucción de los hablantes define más o menos el tamaño del vocabulario resultante.

Algunos ejemplos de tareas en un corpus de habla pueden ser:

- Programar una reunión de negocios.
- Planificación de viajes.
- Comprar equipos.
- Programe la alarma de su teléfono celular.

A.1.1.3.4 Distribución Fonológica

En algunos casos (muy a menudo en el contexto científico o en combinación con síntesis de voz) el contenido hablado de un corpus de habla tiene que ser específico no en términos de vocabulario, sino en términos de unidades fonológicas, como los fonemas, sílabas, morfemas.

Por ejemplo, un sistema de reconocimiento de habla requerirá un número mínimo de repeticiones de cada posible fonema en diversos contextos por cada hablante. O bien, un corpus para síntesis de voz concatenativa requerirá cada combinación de los difonos pronunciados desde el mismo hablante con un mínimo de 20 contextos izquierdos y derechos diferentes.

A.1.1.4 Estilos de Habla

Los estilos de habla son otra característica clave que define los posibles usos del corpus de habla. Por ejemplo un corpus que contiene habla espontánea o habla no-guiada no será útil para una tarea de dictado.

Desafortunadamente muchos cuerpos de voz contienen sólo un estilo de habla, por lo que están restringidos en su reutilización para diferentes aplicaciones. Esto es una pena teniendo en cuenta el hecho de que la recolección y grabación de los hablantes es la parte más costosa en el desarrollo de un corpus de habla. Por lo tanto se recomienda que se definan, por lo menos, dos estilos diferentes de habla en el desarrollo de un corpus. La siguiente lista ofrece una visión general de los principales estilos de habla, listados en relación al aumento de la complejidad.

- Habla leída (*read speech*).
- Habla en respuesta a información específica (*answering speech*).
- Habla de control (*command/control speech*).
- Habla descriptiva (*descriptive speech*).
- Habla no-apuntada (*non-prompted speech*).
- Habla espontánea (*spontaneous/free speech*).
- Neutral vs. Emocional.

Hay que tener en cuenta que el/los estilos de habla elegidos van a interferir con otras especificaciones, como el tipo de grabación, perfiles de los hablantes, etc.

A.1.1.4.1 Habla Leída

La mayoría de los corpus de habla contienen habla leída, ya sea por razones prácticas, ya que la obtención de habla no leída (*non-read speech*) es más difícil, o simplemente porque prevista aplicación o investigación requiere habla leída. El habla leída puede ser grabada

usando los documentos llamados guion(es) o mostrando texto en un dispositivo de salida gráfica.

El habla de dictado (*dictation speech*) es un caso especial del habla leída: Se pide a los hablantes leer un texto como en una tarea de dictado. Este texto debe poseer instrucciones exactas sobre cómo debe ser leído el dictado y así evitar casos especiales como acrónimos y números, ya que todo el dictado debe ser dicho de forma coherente.

A.1.1.4.2 Habla en Respuesta a Información Específica

El habla en respuesta cubre a todas las grabaciones que están apuntadas mediante una pregunta. Estas preguntas pueden ser diseñadas de manera que sólo puedan ser respondidas mediante la selección de un determinado conjunto de opciones de vocabulario cerrado. Por ejemplo:

- Un sistema bancario solicita el número de tarjeta de crédito de un cliente.
- *Si/No* a preguntas como *¿Eres mujer?*
- Un conjunto determinados de números para preguntas como *¿Qué edad tienes?*

O bien, estas preguntas pueden ser diseñadas para ser respondidas como texto libre, por ejemplo, *¿Qué comiste en el desayuno?* Se debe tener en cuenta que la calidad de expresión difiere considerablemente de un texto leído y de habla libre de ser utilizado en un diálogo.

A.1.1.4.3 Habla de Control

El habla de control y/o mando es utilizada por los hablantes en un escenario en el que se les pide para controlar un dispositivo con un conjunto de comandos de voz conocidos, en la mayoría de los casos ocurre dentro de un experimento Wizard-of-Oz.

A.1.1.4.4 Habla Descriptiva

El habla descriptiva puede ser provocada mostrando una imagen, un gráfico o una película al hablante y pedir una descripción de los elementos que se muestran, también puede ser provocado pidiendo instrucciones que describan un lugar u objeto muy común (de preferencia), por ejemplo, Describe lo mejor que puedas la ciudad donde vives. El habla descriptiva es más espontánea que el habla leída, en respuesta o de control, pero puede fácilmente restringida a un vocabulario de un tema determinado.

A.1.1.4.5 Habla No-Apuntada

El habla no-apuntada cubre todos los estilos de habla que no utilizan ningún guion o texto escrito que será dicho o reproducido palabra por palabra, pero no es totalmente espontánea, es decir, sin ninguna restricción. Por ejemplo, el diálogo entre un piloto y una torre de aeropuerto no se basa en ningún texto escrito, pero tiene que seguir ciertas reglas (sólo una persona habla en un momento dado) que restringen el habla de ambas personas.

A.1.1.4.6 Habla Espontánea

El habla espontánea real sólo se puede grabar en un diálogo cara a cara o en un entorno Wizard-of-Oz muy elaborado. El hablante no tiene restricciones en su habla salvo el tema o tarea que fue proporcionada por el entrevistador o supervisor. Este estilo de habla es uno de los más utilizados en el Procesamiento Digital del a Voz debido a que, idóneamente es el estilo de habla que una persona utilizaría al momento de interactuar con una computadora o dispositivo.

Por ejemplo, cuando una persona dicta unas palabras a un sistema de reconocimiento de habla, su objetivo es comunicarle las palabras en sí, preocupándose más por enunciar las palabras con cuidado e interesándose más en cómo se están produciendo las palabras, sin embargo, cuando la intención/objetivo del hablante es resolver un problema de forma interactiva con dicho sistema, el objetivo no es (o no debería) ser el habla en sí misma, sino el problema que se quiere resolver. Por lo tanto, el habla es producida menos cuidadosamente y de forma más espontánea, más casual, por así decirlo, existiendo más variabilidad en todo el proceso, borrando o reduciendo ciertos segmentos, sílabas, y formas gramaticales *no estándar* son pronunciadas con mayor frecuencia (interjecciones, dudas, repeticiones, falsos comienzos, etc.).

Principalmente, las variaciones típicas del habla espontánea podrían resumirse en los siguientes casos:

- Variaciones fonológicas, es decir, formas alternativas y efectos al pronunciar una palabra. Existen variaciones más o menos regladas en cada lengua, respondiendo a condicionantes geográficos o culturales, que pueden tenerse en cuenta al generar los modelos acústicos de un sistema relacionado al procesamiento digital de la voz. Sin embargo, algunos otros efectos son debidos a la propia espontaneidad del habla generada, a la producción de habla condicionada por la intención de resolver un problema interactuando con un sistema sin preocuparnos por el hecho de cómo se le está hablando al mismo, es decir, más preocupados de transmitirle el contenido o la información sobre el problema que cómo se la estamos comunicando.
- Variaciones Gramaticales, es decir, la producción de formas gramaticales que no forman parte de estándar de un vocabulario. Al hablar de forma espontánea, teniendo que pensar en cómo transmitimos la información sobre el problema al mismo tiempo

que lo hacemos, genera formas que no pertenecen a la frase en sí, efectos propios del hecho de pensar al tiempo que hablamos. Inclusión de interjecciones, la producción de falsos comienzos, expresiones hechas, pausas, suspiros, respiración, etcétera, acompañan a la frase en sí. Incluso en dominios restringidos la interpretación correcta de una frase es algo complicado. Por tanto, parece razonable y útil permitir algo de flexibilidad en las restricciones impuestas por la gramática, con la contrapartida de necesitar sistemas mejores y más eficientes para compensar la relajación sufrida por el módulo gramatical.

- Anáfora, con referentes omitidos en la frase que se encuentran en frases anteriores (historia). A veces, para transmitir la información sobre el problema a resolver el usuario no envía toda la información necesaria empleando una sola frase sino varias. Un módulo de diálogo puede encargarse de interactuar con el usuario para conseguir toda la información y completarla antes de procesar o ejecutar la interpretación de la misma.

A.1.1.4.7 Neutral vs. Emocional

Para algunos corpus de habla, puede ser necesario que el habla contenga o no partes emocionales. Provocar verdadera habla emocional es muy difícil (en general, sólo es viable con un entorno Wizard-of-Oz) y (en algunos casos) legalmente problemática.

A.1.1.5. Tipos de Grabación

Antes de que especifiquemos las características técnicas de las grabaciones es importante definir el tipo de grabación adecuado para el desarrollo del corpus de habla. Básicamente, el tipo de grabación define las características acústicas del corpus resultante y por lo tanto también su capacidad de uso de los datos para ciertas aplicaciones o investigaciones. Uno puede distinguir entre grabaciones abiertas vs. grabaciones en secreto. La mayoría de las personas que saben que están siendo grabadas cambian su comportamiento de habla. Por otro lado, las grabaciones en secreto imponen un problema ético, además de correr el riesgo de pasar mucho tiempo y esfuerzo a cambio de nada, si después los hablantes no dan su consentimiento sobre el uso de las grabaciones. Por lo tanto, las grabaciones en secreto se deben usar sólo cuando no hay otra alternativa. Un buen método para provocar discurso muy natural y espontáneo es hacer que los hablantes realicen una tarea que requiera un poco de actividad cognitiva, ya que por lo general la persona olvida de que está siendo grabada y tiene la ventaja de que uno pueda elegir el equipo de máxima calidad disponible y no aquel que se utilizaría para realizar una grabación en secreto.

Además, el tipo de grabación tiene un impacto en el reclutamiento de los hablantes, por ejemplo, es mucho menos costoso reclutar hablantes para grabaciones telefónicas que para un estudio de grabación ya que uno debe considerar los costos de traslado de los hablantes al estudio.

El tipo de grabación especifica las siguientes características generales:

- Entorno acústico. No especificarlo en términos técnicos como reverberación, índice señal-ruido, etcétera, sino más bien como una descripción de la ubicación en sí: estudio de cancelación de eco, estudio, oficina tranquila, oficina con impresora/teléfono, oficina con n-empleados, sala (si es posible, especificar los muebles y si tiene las ventanas abiertas/cerradas), cuarto con televisión o computadora encendida, sala de estar con los niños jugando, coche parado, coche en marcha, coche en marcha sobre la ciudad, cabina telefónica en la calle, cabina de teléfono en centros comerciales, etc.
- Guion. Define cómo el hablante actúa en ambiente de grabación. En la mayoría de los casos, la única cosa que se especifica aquí es que si el hablante sigue las instrucciones mientras que no cambia de posición. En algunos casos, el guion define las acciones del hablante en paralelo a la grabación: el hablante conduce un carro, el hablante se mueve en la sala, el hablante señala ciertos objetos mientras habla de ellos, el hablante usa un teléfono, etc.

El guion también puede definir el orden de la grabación y por lo tanto tiene un impacto en las características de la voz en sí. Considere, por ejemplo, un guion que presenta frases cortas en grupos de 6 cada una. El hablante leerá estos grupos a partir del documento o de una pantalla y lo más probable es que su agrupación influirá en su prosodia significativamente, por ejemplo mediante la reducción de su tono de voz en la última frase de cada grupo. Para evitar este efecto uno puede superponer las frases de los grupos de manera que la última frase de cada grupo también se encuentra al principio o dentro de otro grupo o utilizar frases de relleno.

Por último, se recomienda que el guion contenga una fase de entrenamiento antes de que la o las grabaciones se inicien, y posiblemente también algunas pausas que se efectuaran durante la grabación. El hablante se acostumbra a dicha situación de ser grabado en la fase de entrenamiento y los efectos de adaptación al sentirse grabado no están representados en el corpus. Las pausas en el guion permiten al hablante relajarse y tal vez tomar agua para mantener su voz sin cambios.

- Ruido definido. Algunas corpora requieren un ruido de fondo definido (por su tipo y/o nivel). Esto sólo se puede lograr en un estudio de grabación.
- Tipo, número, posición y distancia de los micrófonos. Siempre que sea posible, es recomendable el uso de más de un micrófono en el desarrollo de un corpus de habla. El utilizar sólo los micrófonos de alta calidad (y de alto costo) podría aumentar la calidad de las grabaciones, pero no necesariamente su usabilidad. Por lo tanto, es aconsejable utilizar también al menos un micrófono de bajo costo, ya que podría ser utilizado en un producto.

Los tipos de grabación más comunes son:

- Grabaciones telefónicas.
- Grabaciones de sitio.
- Grabaciones de campo.

- Grabaciones en un entorno Wizard-of-Oz.

Es importante señalar que los tipos de grabación mencionados anteriormente pueden ser mezclados, lo que le daría al corpus una amplia reusabilidad.

A.1.1.5.1 Grabaciones telefónicas

Grabar habla mediante redes telefónicas es barato y fácil de realizar. Hoy en día la mayoría de las tarjetas PC *RDSI* permiten configurar un servidor de habla automático que redirige la llamada del hablante a través de una sesión de grabación, y la forma de reclutar personas se reduce a ofrecer algún incentivo al grupo de hablantes. Sin embargo, las grabaciones telefónicas tienen algunas desventajas también: la calidad de la grabación está restringida a una frecuencia de muestreo de 8 kHz con un *bit depth* de 12 bits (comprimido a 8 bits), además hay perturbaciones técnicas que no se pueden evitar, es difícil controlar/validar el ambiente acústico.

A.1.1.5.2 Grabaciones de sitio

Las grabaciones de sitio cubren todas las grabaciones de voz que se realizan en uno o un número limitado de sitios de grabación (ya sea estudios profesionales, el cuarto de una casa, etc.). Estas grabaciones la ventaja de que la calidad de las señales grabadas se puede controlar sin restricciones. Por ejemplo, se puede usar una amplia gama de micrófonos, ya sean de baja o alta calidad, en una sesión de grabación.

Básicamente, este tipo de grabación de sitio puede tener todas las características posibles de configuración enumeradas antes. Aparte de eso, se debe distinguir entre:

- Grabaciones supervisadas, donde el entrevistador o supervisor está presente y puede controlar la sesión de grabación e interferir en caso de errores (y poder repetir grabaciones individuales).
- Grabaciones no supervisadas, donde el hablante sigue un procedimiento automatizado y los errores no se puede corregir al momento de la sesión. Este subtipo de grabación es más rentable porque el mismo supervisor podría realizar hasta tres grabaciones en diferentes salas a la vez, y los errores serán señalados en la fase de anotación después de la recolección de muestras. Sin embargo, si las características definidas de la voz en las grabaciones son 100% esenciales, entonces es recomendable seguir el método anterior.

Las grabaciones de sitio requieren más trabajo (mano de obra) que las grabaciones telefónicas (para supervisión y recolección de muestras). Además, en la mayoría de los casos sólo hay unas pocas habitaciones de grabación disponibles (mientras que en la grabación telefónica se pueden manejar muchas llamadas en paralelo) y por lo tanto se le debe atribuir más tiempo a la fase de recolección de muestras. Las grabaciones de sitio en diferentes lugares requieren

una planificación y capacitación de los distintos equipos que pueda haber para evitar diferencias específicas del lugar en la señal de habla de las grabaciones. Se debe asegurar que el mismo hardware sea usado en todas las ubicaciones y el monitoreo de las grabaciones se debe realizar en un lugar central.

A.1.1.5.3 Grabaciones de campo

Las grabaciones de campo cubren todas las grabaciones realizadas en el *mundo real*. La gran ventaja de estas grabaciones es que todas las características del medio ambiente, y en la mayoría de los casos incluso los perfiles de los hablantes, coinciden exactamente con las necesidades de una determinada aplicación. Sin embargo, los costos son mucho más altos y la reutilización es baja debido a que, usualmente, los corpus con este tipo de grabación están diseñados para una tarea muy específica.

Muy a menudo, las grabaciones de campo son de tiempo crítico en el sentido de que la ubicación y/o los hablantes no están disponibles todo el tiempo. Vale la pena realizar un ensayo unas pocas semanas antes de la fecha de grabación para asegurarse que los dispositivos de grabación y sus procedimientos no tengan problemas.

A.1.1.5.4 Grabaciones en un entorno Wizard-of-Oz

El término *Wizard-of-Oz (WOZ)* se utiliza para el tipo de grabación donde el comportamiento de un sistema o aplicación (por ejemplo, un sistema de lenguaje hablado por computadora) es simulado de manera que el hablante cree que él o ella está interactuando con el sistema real. De hecho, el comportamiento del sistema es controlado por una o más personas llamadas *wizards*.

La gran ventaja de este método es que el comportamiento de la voz del hablante es muy cercana a la de los futuros usuarios de la aplicación prevista. Además, los diferentes aspectos de diseño de la aplicación pueden ser probados de antemano y los datos pueden ser analizados para modelar las reacciones de los usuarios en la aplicación con mucho más éxito.

Dependiendo del esfuerzo que se pone en este tipo de grabación, el entorno acústico se puede adaptar muy de cerca al de una situación real. Por lo tanto, los datos recogidos en la técnica WOZ suele ser lo mejor que se puede conseguir para una aplicación compleja.

Por otro lado, las grabaciones WOZ requieren mano de obra mucho mayor a la de los demás tipos de grabación mencionados anteriormente, y también puede requerir un costo mucho mayor debido a que el ajuste del sistema WOZ debe ser tan convincente que los usuarios ingenuos no sospechan que están siendo engañados, la grabación en sí requiere como mínimo dos personas (un supervisor y un asistente), y finalmente, debido a que el hablante está siendo dirigido por el sistema WOZ, requiere que las personas tengan la capacitación necesaria para manejar el sistema WOZ y realizar las grabaciones.

Las grabaciones WOZ se pueden diseñar en muchos entornos diferentes en función de sus necesidades, por lo que es difícil dar instrucciones detalladas sobre cómo especificarlas. Además, poseen muchos problemas que no se pueden prever, porque las grabaciones WOZ son grabaciones definitivamente no estandarizadas. Lo mejor que se puede aconsejar es que uno quede tan impreciso como sea posible acerca de la técnica WOZ. Se debe tratar de concentrar en la intención general -por ejemplo, que los usuarios no tienen que estar conscientes de que es una simulación, que la configuración del entorno WOZ coincide con la situación real en la mejor manera posible, y así sucesivamente- y no dar algún dato o hecho terminante. Por otra parte, se debe tratar de obtener la mayor información de las especificaciones de la aplicación prevista como sea posible; esta es la base que se debe trabajar más, si no se sabe exactamente cómo se supone que trabaja dicha aplicación, estás perdido(a).

A.1.1.6 Anotaciones

El término *anotaciones* en un corpus de habla, hace referencia a toda la información simbólica que está relacionada con la señal de voz, por ejemplo, transcripciones ortográficas, transcripciones fonéticas, todos los tipos de segmentaciones, etc.

Dado que en la mayoría de los casos algún tipo de anotación es una parte integral del corpus de habla, se debe definir el contenido de la anotación deseada en las especificaciones. También podría ser una buena idea para definir los procedimientos para lograr una anotación, así como el control de calidad de la anotación o la institución, organización o persona encargada de realizarlas.

A.1.1.7 Especificaciones Técnicas

Las especificaciones técnicas definen las propiedades formales de los datos del corpus de habla. Básicamente, todas las señales y anotaciones deben ser especificadas en esta parte. Esta sección ofrece una visión general acerca de las posibles categorías y valores en un corpus de habla estándar. Se debe tener cuenta que esta lista podrá ampliarse en otras categorías y valores si se utilizan dispositivos de grabación especiales distintos a los de las señales acústicas, por ejemplo: señales de vídeo, señales de farinogramas, señales bioeléctricas como electroencefalogramas o electrogastrogramas, etc.

A.1.1.7.1 Frecuencia de Muestreo

El teorema de Shannon o Ley de Nyquist requiere que la frecuencia de muestreo sea mayor que el doble de la frecuencia máxima en la señal digitalizada. Dado que el habla está más o menos situada por debajo de los 8 kHz, la mayoría de los corpus de habla tienen una frecuencia de muestreo de mínimo 16 kHz. Una excepción son las grabaciones telefónicas

donde el ancho de banda se reduce técnicamente 300 Hz - 3300 Hz y por lo general utiliza una frecuencia de muestreo de 8 kHz. Dado que el estándar del CD de audio fue introducido con 44.1 kHz, los divisores 22.05 kHz y 11.025 kHz también son utilizados debido a que algunos dispositivos de audio no procesan otras frecuencias de muestros distintas a las mencionadas. Todo esto es también la razón por la que se recomienda evitar, siempre que sea posible, las frecuencias de muestreo de muestreo exóticas.

Las señales de los farinogramas generalmente se muestrean a la misma frecuencia que las de las señales de voz. Debido a que el ancho de banda de sus datos de movimiento del habla es bajo y se puede ser muestreado a unos 200 Hz.

Algunas recomendaciones de frecuencias de muestreo son:

- Grabaciones Telefónicas: 8 kHz
- Grabaciones de sitio: 16 kHz o 22.05 kHz mínimo
- Grabaciones de campo: 16 kHz mínimo

A.1.1.7.2 Tipo y Formato de Muestra

El tipo de muestra define el formato de cada uno de los valores de la frecuencias de muestreo, y a la vez define el número de bits (*bit depth*) requeridos para representar cada valor de la frecuencia de muestro en un medio de almacenamiento. Por supuesto, ambos son dependientes. Los estándares típicos son:

- Grabaciones telefónicas: *ULAW* (EUA) o *ALAW* (resto del mundo), 8 bits.
- *PCM* (lineal), 16 bits ya sea con signo (valores con un rango $[-32768,32767]$ o sin signo (valores con un rango $[0,65535]$).
- *GSM*.

A.1.1.7.3 Número de Canales, Intercalado

El almacenar varios canales en un sólo archivo de audio se llama intercalado o multiplexado. Sin embargo, intercalar archivos de señales puede resultar, en algunos casos, muy difícil de procesar. Por ende, es recomendable guardar la señal de cada canal en su respectivo archivo de la grabación, claro, si es que se requiere utilizar archivos de audio con más de 1 canal (mono).

A.1.1.7.4 Formatos de Archivo

El formato de los archivos define el estándar específico por el cual los archivos serán almacenados. Dado que un corpus de habla siempre contiene señales, y puede contener

anotaciones, metadatos y, en muchos casos, un diccionario, se describirán cada uno de ellos separadamente.

A.1.1.7.4.1 Formatos de Señales

Existe un buen número de formatos de archivo de señales estandarizados. Esta sección se centra en los formatos más comunes del Procesamiento Digital de la Voz.

En la mayoría de los casos, el formato de una señal consiste en: una cabecera (*header*) la cual contiene información sobre la señal, por ejemplo, la frecuencia de muestreo, el tipo y formato de la muestra, el número de canales, etc.), y un cuerpo (*body*) que contiene las muestras digitalizadas de la señal.

Los formatos más comunes son:

- **RAW.** El formato más sencillo, no tiene *header*, sólo el *body* de la señal digitalizada. La desventaja es que uno tiene que obtener las especificaciones de la señal de otro lugar. Algunas corpora agregan una extensión que define dichas especificaciones de las señales. Por ejemplo: *.raw*, tipo *pcm a 16 bits*.
- **WAV.** El formato de archivo de audio estándar de Microsoft. Fue creado por Microsoft e IBM y es actualmente uno de los más utilizados en las corpora de habla ya que no tiene compresión de datos. Su extensión de archivo es *.wav* y puede ser creado a partir de diferentes formatos de muestra como: PCM, ADPCM, GSM, ALAW, ULAW, CCUIT, entre otros.
- **NIST SPHERE.** El formato NIST SPHERE fue definido por el *Speech Group* en el *National Institute for Standards and Technology, USA*. Se compone de un *header* legible en texto sin formato (*7 bits US ASCII*), seguido de los datos de la señal en forma binaria. Debido al formato simple y extensible, se utiliza ampliamente en la comunidad científica que trabaja el habla y en muchas corpora de habla. La mayoría de los programas científicos que trabajan el habla pueden reconocer este formato automáticamente; otros programas comerciales no. La gran ventaja de este formato es que la información de la cabecera está en texto, por lo que es muy fácil de extraer e insertar los valores en algún otro lado o en la señal misma. La gran desventaja, la modificación del *header* requiere modificar el archivo completo.
- Sus extensiones comunes de archivo son: *.nis* o *.nist*.

A.1.1.7.4.2 Formatos de Anotaciones

Las anotaciones son datos simbólicos asociados con las señales registradas del corpus del habla. El formato en que se almacenan estos datos tiene que ser incluido en las especificaciones del corpus. Dado que no existe un estándar ampliamente aceptado y puesto que muy a menudo las corpora de habla contienen una nueva forma de anotación que nunca

se ha aplicado antes, muchas corpora incluyen sus formatos particulares que se definen sólo para esa ocasión. Algunos de estos formatos han sido comúnmente aceptados y se han reutilizado en otras colecciones.

En esta sección se indican algunos formatos de archivo estándar para la anotación de datos y sus respectivas propiedades. Estas propiedades son descritas por los siguientes criterios:

- Independencia de la plataforma.
- Autodescriptivo.
- Texto plano vs Texto de marcado.
- Disponibilidad de la herramienta(s).

Los formatos más comunes son:

- SAM (utilizado en la popular *SpeechDat telephone speech corpora*).
- EAF (Eudico Annotation Format).
- BPF (BAS Partitur Format).
- AGS (Annotation Graphs).
- TextGrid (formato de anotación utilizado por la popular herramienta *Praat*).

A.1.1.7.4.3 Formatos de Metadatos

Para los metadatos, no existe un formato que sea aceptado comúnmente. En general, se recomienda un formato de texto de marcado basado en XML y Unicode.

A.1.1.7.4.4 Formatos de Lexicon

Dependiendo de la complejidad de su corpus de habla, puede agregar un *lexicon* que cubra a todo el corpus o a partes de él. Una vez más no hay un estándar comúnmente aceptado sobre el formato de archivo para léxicos o diccionarios.

En la mayoría de los casos, las corpora de habla vienen con una simple lista de tres columnas para cada forma de cada palabra en su representación ortográfica, el número de veces que se dice la palabra en el corpus y su pronunciación más probable. Esto parece muy sencillo, pero es un verdadero dolor para aquellos idiomas que no tienen una ortografía estandarizada, que su ortografía no se puede establecer sin ambigüedad respecto a su habla o que su ortografía no se basa necesariamente en palabras. Estos son algunos consejos para el *lexicon*:

- Ortografía: siempre que sea posible, utilice Unicode.
- Pronunciación: siempre que sea posible, utilice SAMPA o X-SAMPA.
- Especifique claramente lo que se entienda por la pronunciación *más probable* o *canónica* y cómo va a producirlas.

- Especifique si puede haber más de una posible pronunciación de la misma forma de la palabra en el léxico.
- Utilice una simple lista de texto sin formato o XML como formato de archivo (todo el mundo es feliz utilizando los formatos *txt* o *xml* porque pueden ser fácilmente importados a cualquier sistema y/o base de datos).

A.1.1.8 Estructura de un Corpus de Habla

La estructura del corpus define la estructura interna de la versión final del corpus, la nomenclatura del nombre de sus archivos y los medios de distribución. Si se tiene planeada una recolección de muestras de larga duración, es recomendable pensar en la forma en que distribuirá las versiones de su corpus.

Definir la estructura de corpus es de suma importancia si se está esa trabajando con un consorcio de clientes o socios.

A.1.1.8.1 Estructura

Como se mencionó antes, es una buena idea mantener por separado las señales y las anotaciones. La razón de esto es que muy a menudo los usuarios sólo necesitarán acceder las anotaciones del corpus, ya que es más probable que las anotaciones estén sujetas a cambios o actualizaciones que las señales de habla. El tener separadas las señales y las anotaciones también permite realizar mantenimientos o actualizaciones de manera más sencilla.

Las corpora pequeñas normalmente siguen una estructura común, muy parecida a la siguiente:

- *DATA*: contiene todos los archivos de señales de habla.
- *ANNOT*: contiene todos los archivos de anotaciones.
- *META*: contiene todos los archivos de metadatos.
- *DOC*: contiene toda la documentación del corpus.
- *LEX*: contiene el lexicon (si es que existe la posibilidad de hacerlo)
- *TOOLS*: contiene el software para acceder a las señales, anotaciones y lexicon. También contiene las configuraciones de los archivos *raw*.

Es importante recordar que cada punto de la lista anterior representa un directorio, los cuales normalmente están situados en la raíz del medio de distribución (CD, USB, etcétera).

Es recomendable que los directorios *DATA* y *ANNOT* tengan una cantidad mínima de subdirectorios, también es recomendable que estos directorios tengan un orden de lo más natural o entendible al usuario. Dependiendo de las especificaciones de su corpus de habla, el nombre de los subdirectorios dentro de los 2 mencionados anteriormente pueden ser:

- Hombres/Mujeres.
- Sesiones de grabación.
- Hablantes.
- Entornos acústicos.
- Idiomas.
- Tipos de dialecto.
- Estilos de habla.
- Tipos de grabación.
- Tipos de anotación (en el caso de las anotaciones).

A.1.1.8.2 Nombres de Archivo

La nomenclatura de los nombres de archivo define el nombre permitido para los archivos y directorios dentro de su corpus de habla. Un enfoque muy común es utilizar una nomenclatura basada en el contenido de los archivos, otra alternativa es utilizar nombres de archivos generados por una secuencia.

La nomenclatura de los nombres basada en el contenido es construida en base a las características del corpus y permite acceder a fragmentos específicos del corpus simplemente filtrando los nombres de los archivos. Por supuesto, la nomenclatura para el nombre de cada archivo tiene que ser entendible y fácil de extraer. Un problema con los nombres de archivos basados en contenido es la dependencia del medio o plataforma de distribución, por ejemplo, 8.3 para CDs ISO 9960.

Finalmente, usted puede ver la Sección 4.5.4 para darse una idea de cómo definir la nomenclatura de su corpus. Asimismo, otras características comunes a considerar para la nomenclatura son:

- Escenario o entorno: *a* (sala de cada), *b* (estudio), *c* (salón de clases), etc.
- Descripción del medio de grabación: *c* (celular), *m* (micrófono), *o* (micrófono del ordenador), etc.
- Número de canales: de *1* a *n*.
- Momento del día: *m* (mañana), *t* (tarde), *n* (noche), etc.

A.1.1.8.3 Medios de Distribución

Se deben especificar sobre cuales medios de comunicación se distribuirá el corpus del habla a los usuarios, clientes, etc. Básicamente, se tiene la posibilidad de elegir entre *CD-ROM* (650 MB), *DVD-R/RW* o *DVD + R / RW* (4.7 GB), *DVD-RAM* (5.2 GB), unidades removibles como *USBs* (2 GB o más) o *discos duros* (1 TB o más), o bien, subir el corpus a un servidor *FTP* y permitir su descarga por *internet*.

También puede especificar el número de copias que se producirán y quién va a cubrir los costos de los medios de almacenamiento y distribución. También hay que tener en cuenta que podrían ser necesarios dispositivos especiales para producir los medios de distribución y por lo tanto esto debe tenerse en cuenta en el plan de financiación.

A.1.1.9 Metadatos

El término *metadatos* (*metadata* en inglés), en el caso de grabaciones de habla, no se refiere sólo a la información de habla grabada, sino a la información acerca de esa información grabada. El énfasis aquí reside sobre el término en inglés *data* (información) debido a que los metadatos no incluyen documentación del corpus de habla. Los metadatos consisten en información categorizada y leíble por computadora que pueda ser usada para clasificar la información del habla contenida en el corpus.

En consecuencia, los metadatos consisten en códigos (contrario a *texto libre*), a excepción de los comentarios. Cuando se especifica la meta data de un corpus de habla, es importante no sólo especificar el tipo sino también el conjunto de valores posibles.

Constantemente se producen nuevas corpora de habla y muchas de ellas se vuelven accesibles para los científicos y desarrolladores, por lo que su diversidad está creciendo rápidamente. Como consecuencia, se hace más y más difícil para el usuario decidir qué corpus es óptimo para su trabajo. Por lo general, no es posible acceder a la señales de habla de un corpus sólo para corroborar si el corpus es más calificado para el trabajo. Pero en algunos casos si es posible acceder a los metadatos, ya que son una descripción formal del corpus subyacente y por sí solos son de poco valor comercial. Por lo tanto, los metadatos juegan un papel importante en la fase de planificación y para la adquisición de corpus orales.

Por desgracia, en el pasado muchos corpus de habla se produjeron bajo un paradigma diferente: en la mayoría de los casos el objetivo era producir datos que se utilizaran en una aplicación específica, de la forma más rápida y económica posible, no se aplicó ningún énfasis sobre los metadatos; estos eran considerados como una minúscula parte de la documentación (en el mejor de los casos), y en consecuencia los metadatos a menudo no eran analizables, carecían de una estructura y estaban incompletos. Mediante este paradigma, un corpus se vuelve prácticamente inútil en términos de reutilización, simplemente porque después de un rato, ya no habrá persona o documento que conozca las especificaciones exactas o circunstancias bajo las que el corpus fue desarrollado.

En general, el termino *metadatos* se refiere a muchos tipos de datos o información acerca de categorías más generales de *recursos lingüísticos*, desde las cuales, un corpus de habla no es más que una subcategoría. Sin embargo, en el contexto de las corpora de habla, los metadatos pueden estar restringidos a 3 principales tipos: *protocolos de grabación*, *perfiles de los hablantes* y *comentarios*.

A.1.1.9.1 Protocolos de Grabación

Cada grabación tiene que tener un protocolo de grabación en el que se registra toda la información importante acerca de la grabación real. En el caso de ser legible por algún software en particular, el protocolo debe tener una forma estandarizada (analizable) en *XML*. Si el corpus de habla contiene grabaciones bajo exactamente las mismas condiciones, sólo se necesita un protocolo de grabación para todo el corpus.

Si el esfuerzo para proporcionar metadatos se reducirá a un mínimo, cinco requisitos mínimos son indispensables para obtener un protocolo de grabación útil. Estos requerimientos mínimos son:

- ID de sesión.
- ID del hablante.
- Fecha de grabación.
- Condiciones ambientales.
- Condiciones técnicas de grabación.

El *ID de sesión* identifica un registro o grabación en especial en el corpus. A menudo se compone de letras que le dan amplia información categórica sobre la grabación (por ejemplo, idioma, sexo del hablante, tipo de grabación, dominio, etcétera) y un número. El *ID de sesión* se utiliza a menudo en los nombres de archivo dentro del corpus de datos para indicar que pertenecen a la misma sesión de grabación.

El *ID del hablante* es un código de identificación reemplazar el nombre real del hablante para asegurarse de su anonimato. La asignación de los *ID del hablante* a los nombres reales no deben serán publicados con el corpus. El *ID del hablante* también se utilizará en el perfil de hablante, así como en cabeceras de los archivos de audio y comentarios.

La *fecha de grabación* se realiza automáticamente cada vez que se realiza una grabación. Cada grabación tendrá su respectiva información acerca de cuándo y a qué hora se grabó.

Las *condiciones ambientales* son parámetros que describen las condiciones del lugar donde se grabó dicha entrevista, algunas de ellas son la acústica de la sala o lugar, las fuentes de ruido y la presencia de interferencias de otros hablantes. Por lo general, sólo se especifica algún ruido persistente o habla entredicha por otra persona.

Las *condiciones técnicas de grabación* es información acerca del equipo para realizar las entrevistas del corpus, por ejemplo el micrófono, equipo de grabación, volumen, especificaciones técnicas de la grabación (*bit depth*, frecuencia de muestreo, número de canales), colocación y distancia del micrófono.

Otros parámetros recomendables podrían ser:

- Nombre o ID del supervisor/entrevistador.
- Detalles acerca del dominio de la grabación.
- Detalles acerca de las instrucciones que se le proporcionaron al hablante.
- Duración de la sesión de grabación.
- En el caso proporcionar un guion al hablante: Tipo de guion (papel, cara a cara, pantalla, voz).
- Si/No en caso de que haya habla emocional.
- Si/No en caso de haber un supervisor/entrevistador presente.
- Si/No en caso de haber un intérprete presente.
- En el caso de aplicar un entorno WOZ: Detalles acerca de la configuración WOZ.
- Estilo de habla.
- Comentarios libres.

A.1.1.9.2 Perfiles de los Hablantes

No confundir con el *Perfil de los Hablantes* de la *Sección A.1.1.1*. Esta sección está relacionada únicamente a su empleo en los metadatos.

Los rasgos característicos de cada hablante se deben de recoger en un conjunto de perfiles de los hablantes del corpus. Esto puede ser un archivo para cada hablante, una tabla que resuma las características de cada hablante en una línea o columna, o puede estar de manera indexada a un archivo único de metadatos con todas las características por cada hablante. Para que estos perfiles sean compatibles con la mayoría de los sistemas, lo recomendable es utilizar un formato estructurado de texto estandarizado, por ejemplo, el *xml*. Si es una persona que está produciendo corpora de habla de manera regular, puede considerar la inclusión de estos perfiles en un sistema de base de datos.

Los requerimientos mínimos de los perfiles de los hablantes son: su *ID o nombre completo*, *Sexo* y *Fecha de Nacimiento*.

En algunos casos, existen proyectos de desarrollo de corpora de habla que se enfoquen a ciertas comunidades en donde puede haber varios casos donde los hablantes no cuenten con su *fecha de nacimiento* por alguna razón. En estos casos, en vez de solicitar la *fecha de nacimiento*, resulta mejor solicitar únicamente la *edad* del hablante.

Otros parámetros que podrían resultar útiles podrían ser:

- Lengua madre del hablante.
- Segunda lengua del hablante.
- Lengua madre de los padres del hablante.
- Patologías.
- Prótesis dental.
- *Piercings* (sólo en el caso de que haya tenido alguno en el cuello, labios o alguna parte interna de la boca, especificar la cantidad de ellos y la parte del cuerpo).

- Lugar de escuela primaria.
- Dialecto.
- Nivel de educación.
- Capacidad ante cierta tarea.
- Profesión.
- Peso.
- Altura.
- Si/No en el caso de que sea fumador.
- Problemas auditivos.
- Comentarios libres.

A.1.1.9.3 Comentarios

La definición de comentarios en este contexto se refiere a toda la información extra que no encaja en las categorías del Protocolo de Grabación o en los Perfiles de los Hablantes. Esto quiere decir que los comentarios muchas veces contienen información acerca de los eventos, características y observaciones no previstas por el diseñador del corpus. Como tales, en la mayoría de los casos muy valiosos, por lo que debe tener un procedimiento para capturar los comentarios de los hablantes, experimentos, etiquetas, etc. Los comentarios no tienen un formato estructurado como otros metadatos, por lo tanto es discutible si pertenecen a los metadatos en absoluto. Sin embargo, por razones prácticas, están incorporadas a esta sección, ya que es muy fácil de insertar un texto de entrada de campo libre en un archivo, ya sea del Protocolo de Grabación o de los Perfiles de los Hablantes. Del mismo modo es posible añadir estos campos de comentario en las etiquetas y archivos de transcripción.

Los comentarios deben ser mantenidos en su versión original con texto original, algunos resúmenes también son posibles, pero debe ser reconocible si el comentario es una versión resumida o la versión original. Más allá de eso debe ser evidente si los comentarios se han recogido de forma sistemática (por ejemplo, en forma de un cuestionario) o por coincidencia (por ejemplo, un sujeto expresa algo acerca de la grabación sin que se le haya preguntado explícitamente).

Los comentarios deben ser mantenidos en la distribución del corpus de habla de manera que sean accesibles para los usuarios. Es una buena idea para mantenerlos en una forma (por ejemplo, archivos de texto sin formato) que pueden ser buscados por palabras clave.

Los más comunes son los comentarios sobre el comportamiento de los hablantes, por ejemplo:

- ¿Ha mostrado alguna emoción?
- ¿Hizo algún gesto después de una pregunta?
- ¿Posee alguna clase de comportamiento nervioso o parecido?

Otros comentarios pueden provenir del investigador, la persona encargada de realizar el etiquetado de las grabaciones, en la etapa de Post-procesamiento, o incluso un grupo de validación externa.

Por último, todos los comentarios recogidos durante la producción corpus pueden formar una buena fuente de información para la documentación del corpus.

A.1.1.11 Procesos de Validación

El término validación se ha mencionado ya unas cuantas veces en este trabajo. Esta sección brinda algunos consejos básicos acerca de la validación tan lejos como sea relevante para un corpus de habla. Los puntos principales son:

- Validaciones internas vs. externas.
- ¿Cuándo realizar una validación?
- ¿Qué se debe validar?

Además, la especificación del corpus (o el contrato/acuerdo con el cliente, si es el caso) debe tener un plan o calendario estricto, el cual marca las principales etapas de desarrollo del corpus, así como los planes o procesos de validación. Si se tiene planeado realizar una pre-validación y/o una validación final (altamente recomendado), se debería también incluir los detalles y fechas para los procedimientos de validación.

También es sabio proponer procesos de *validación interna* durante el transcurso de la etapa de recolección de muestras, y de anotaciones y transcripciones. Especialmente, las anotaciones y transcripciones deberían ser sometidas a una etapa de corrección final una vez que éstas hayan concluido, preferentemente realizada por una sola persona experta, para asegurar una calidad buena y consistente de las anotaciones.

A.1.1.11.1 Validaciones Internas Vs. Externas

Una *validación interna* (*internal validation*) hace referencia básicamente al control de calidad durante o después del desarrollo de un corpus de habla, y es realizada por los miembros de su grupo de trabajo o institución. Definitivamente, es mucho más económica que una *validación externa* (*external validation*) que requiere de más dinero, tiempo y esfuerzo.

Sin embargo, es recomendable hacer uso de una *validación externa* siempre que sea posible, ya que las validaciones internas tienden a no ser muy efectivas. La razón de esto es similar a, por ejemplo, el hecho bien conocido de un programador que no puede encontrar el error de su programa/código debido a que él se encuentra muy metido en el proceso interno del mismo, sin embargo, un observador externo frecuentemente brinda una solución simple y un tanto obvia (es por eso que los programadores tienen a hablar demasiado acerca sus

problemas de programación). Lo mismo sucede con los errores en el desarrollo de un corpus de habla, por lo tanto es muy importante realizar validaciones externas tan a menudo como sea posible.

También se puede dar el caso en donde no haya fondos disponibles para llevar a cabo validaciones externas, o peor aún, que no haya nadie que pueda actuar como una institución de validación externa. Si el corpus del habla se produce para un cliente o empresa, la solución obvia es hacer que el cliente o empresa actúe como una institución de validación externa. Sin embargo, si lo hace, asegúrese de que en el contrato existan pautas precisas que especifiquen lo que tiene que ser validado, cuándo y cómo se debe de hacer.

A.1.1.11.2 ¿Cuándo Validar?

¿Cuándo es el mejor momento para validar los resultados del desarrollo de un corpus de habla? Esto depende del tamaño, escala de tiempo y tecnología o tarea revista para la cual está diseñado el corpus de habla. Las corpora pequeñas, tomando en cuenta que sigan una línea de tiempo de trabajo óptima en todo sentido, toman alrededor de 3 a 4 meses en ser finalizadas y requiere solamente de una prevalidación (*pre-validation*) y una validación final (*final validation*). Las corpora más grandes requieren de procesos de validación periódicos, y gozan de una *validación de lanzamiento o estreno (release validation)* por cada nueva distribución o versión liberada del corpus.

Una *prevalidación* se debe hacer después de que se hayan realizado un cierto número de grabaciones de los hablantes, y todos sus datos ya han sido anotados, transcritos y post-procesados. Normalmente se efectúan durante el proceso o fase de recolección de muestras.

La *validación de lanzamiento o estreno* toma lugar en momentos definidos del desarrollo del corpus para mejorar la consistencia interna de cada versión o distribución que se va a lanzar o estrenar del mismo. Es recomendable esperar los resultados de la validación antes de liberar la versión o distribución prevista del corpus, así como también para empezar a realizar las fases de recolección de muestras o anotaciones para la siguiente versión/distribución. Si se está trabajando con un cliente, recuerde especificar en el contrato las fechas y tiempos para cada validación de este tipo. Asimismo, el contrato debe especificar cómo manejar los errores encontrados, por ejemplo, si se encuentra un error sistemático que ha existido a través de varias versiones, ¿Debería ser corregido o no? ¿Habrán suficientes fondos disponibles para tales correcciones o mejoras?

La *validación final* se lleva a cabo después de que el desarrollo de un corpus de habla se declare completo, finalizado o terminado. Es recomendable mantener una cierta cantidad de fondos de financiación que le permitan realizar otra validación después de la validación final. Después de todo, una validación final tampoco está libre de errores.

A.1.1.11.3 ¿Qué Validar?

Básicamente, cada muestra, anotación, metadato, lo que sea que está incluido en las especificaciones del corpus de habla puede ser objeto de validaciones. Qué será validado y qué será considerado como un error son tópicos que deben estar plasmados en las especificaciones del corpus (o en el contrato, en el caso de estar trabajando con un cliente). Por lo general, las siguientes partes del corpus de habla se validan de la siguiente forma:

- Documentación: coherencia, integridad, estructura.
- Metadatos: integridad, compatibilidad, contenido (muestras).
- Señal de datos: integridad, calidad técnica, calidad acústica, contenidos.
- Anotaciones: integridad, compatibilidad, contenidos (muestras).
- Diccionario: integridad y calidad.

A.1.1.11 Documentación

Aunque no es muy común, los procedimientos de documentación también pueden ser especificados de antemano. Esta es probablemente una buena idea para grandes proyectos relacionados a varios socios que trabajan en un objetivo común.

La documentación de un corpus de habla resume toda la información importante relativa a la producción y el uso previsto del corpus. No contiene anotaciones, metadatos o cualquier tipo de datos simbólicos directamente relacionados con las señales de voz. La documentación consta de un texto descriptivo (preferentemente en inglés), figuras y opcionalmente imágenes.

Sin embargo, la distinción entre la documentación y los metadatos en la definición anterior es a menudo difusa. En muchos casos los datos de las corpora de habla (que son esencialmente metadatos), pueden encontrarse en la documentación del corpus y la razón simple: En la mayoría de los casos, estos metadatos son constantes durante todo el corpus del habla y por lo tanto no están incluidos en todos los perfiles de los hablantes o en el protocolo de grabación. Por otra parte, algunos autores definen los metadatos de una forma mucho más amplia de lo que normalmente se hace en la práctica: por ejemplo también incluyen parámetros que describen al corpus como el número de hablantes, el número de artículos registrados, especificaciones técnicas, etc.

Este capítulo se limita a dar una visión general de lo que se considera como partes esenciales de cualquier documentación de un corpus del habla. Esto no es una extensa lista porque no se pueden prever todas las necesidades de un corpus de habla en particular o de futuras producciones de corpus de habla.

En resumen, la documentación de un corpus consiste de las siguientes partes:

- Introducción.
- Derechos de Autor, Descargos de Responsabilidad y Reglas de Uso.
- El número de versión y fecha.
- Lista de los archivos de la documentación.
- Descripción de corpus.
 - Números (hablantes, grabaciones, etc).
 - Estructura.
 - Contenido.
 - Terminología (nombres de archivos).
 - Especificaciones técnicas del formato de las señales.
 - Otras: diccionarios, traducciones, etc.
- Reclutamiento.
 - Perfil de los Hablantes.
 - Técnicas para recolectar muestras (la forma que se utilizó para reunir gente).
 - Aspectos Legales.
- Grabaciones.
 - Tipos de Grabación.
 - Guion.
 - Técnica.
 - Fechas de los archivos.
- Post-procesamiento.
- Anotaciones.
 - Contenido.
 - Procedimiento.
 - Formatos de archivo.
- Metadatos.
 - Contenido.
 - Formatos de archivo.
- Guiones visuales, archivos de guiones, etc.
- Informes de validación.
- Publicaciones, reportes internos.
- Comentarios.
- Historia Corpus.
- Errores conocidos.

A.1.2 Recolección

Después de haber establecido todas las especificaciones necesarias del corpus de habla, se necesita algún tiempo para prepararse para la *fase de recolección*, a este lapso de tiempo de le llama *fase de preparación*. No hay que subestimar el tiempo requerido para esta *fase preparación*, es posible que se necesite más tiempo para preparar esta fase que aquel requerido en la *fase de recolección*.

Ahora bien, esta sección se centra en la *fase de recolección*, la cual describe las actividades principales de todo desarrollo o producción de un corpus de habla, concretamente, la grabación de las señales de voz. La mayor parte de *la asistencia técnica y eso* se encuentra en la Sección A.1.1. Esta parte resume algunos consejos un tanto prácticos que podrían ser útiles durante la fase de recolección. Es importante aclarar que el orden de las secciones en este capítulo no están destinadas a que se tomen de manera cronológica, lo recomendable es que primero se lean todas y después uno decida cuáles y en qué orden se deben hacer.

A.1.2.1 Instrucciones para el Hablante

Todos los hablantes que participen en las grabaciones del corpus necesitan algún tipo de instrucción antes de que comience la grabación. Las instrucciones pueden variar de una instrucción muy reducida basada en experimentos inspirados psicológicamente a instrucciones muy detalladas y estrictas para una recolección supervisada por red telefónica. Si es posible, evite hacer uso de instrucciones verbales dadas por el supervisor o entrevistador antes de comenzar grabación. Siempre use una instrucción por escrito, o utilice las instrucciones pregrabadas.

Haga las instrucciones tan simples y tan inequívocas como sea posible. No sobrecargue a los hablantes con información básica sino de una breve reseña de lo que está y va a pasar. Describa el contenido, el estilo del habla, el tipo de grabación y la duración estimada de la grabación.

Muchos corpus piden al hablante de decir expresiones específicas. Esto se puede hacer acústicamente (resultando en estilos de habla imitadas) o en el papel o un monitor (que resulta en discurso leído). Excepto con fines especiales para una determinada prosodia, volumen o el estilo de habla, se recomienda utilizar siempre guiones escritos en papel o mostrados mediante alguna pantalla.

A.1.2.2 Técnicas de Grabación

Las siguientes dos secciones brindan información que podría ser útil al momento de realizar las grabaciones. Debido al modo de trabajo que recibí Entendámonos, sólo se tomarán en cuenta las grabaciones en sitio y de campo. Si requiere la información adicional referente a las grabaciones telefónicas y en entornos WOZ, puede encontrarla en [1].

A.1.2.2.1 Grabaciones en Sitio

Si realiza grabaciones en un estudio profesional, lo más probable es que éste cuente con el personal capacitado para realizar las grabaciones. Si es así, puede saltarse esta sección por completo.

Sin embargo, si tiene planeado establecer su propio sistema de grabación, los siguientes consejos tal vez puedan ser útiles. Una buena cosa para evitar errores previos a las sesiones de grabación son las listas de verificación. Proporcionarle una lista de verificación al supervisor o entrevistador de la grabación resulta algo útil, más cuando no se tuvo el tiempo o fondos para capacitar completamente al personal.

Antes de elegir un lugar para grabar, asegúrese de que el entorno acústico sea el adecuado. No altere los muebles y decoraciones del lugar durante la fase de recolección a no ser que eso forme parte de algún experimento o técnica para las muestras del corpus. Recuerde que el lugar no debe poseer eco, o bien, debe tener la menor cantidad posible. Si es posible, documente el lugar de grabación con algunas fotos en diferentes ángulos.

Otro aspecto importante a considerar es el micrófono. Lamentablemente recomendar algún micrófono en especial para realizar las grabaciones es algo que no tiene mucho sentido, micrófonos hay muchos y si uno va a una tienda especializada, lo más probable es que el personal de dicha tienda pueda ayudarles a conseguir el micrófono indicado para realizar el tipo de grabación que necesite. Por ejemplo, debido a que Entendámonos consiste en grabaciones de campo, se obtuvieron varias opiniones con las que se decidió adquirir el micrófono *SHURE PG42-USB* (ver Apéndice B), el cual está documentado en el Apéndice B. Pasando a consejos un poco más prácticos, intente mantener siempre la misma distancia y posición del micrófono respecto al hablante. Si el micrófono cuenta con una interfaz de volumen o ganancia de información/sonido, intente mantenerlos siempre al mismo nivel o con la misma configuración. Se recomienda no utilizar diferentes tipos de micrófono para la fase de recolección, utilice siempre el mismo micrófono para realizar todas sus grabaciones, en el caso de que se deba cambiar el modelo del micrófono, especifíquelo en la documentación.

Por último asegúrese de contar con el software indicado para realizar las grabaciones, este software debe tener las capacidades suficientes acorde a las especificaciones técnicas de su corpus, además de que debe ser compatible con todos sus equipos de trabajo y con su micrófono(s).

A.1.2.2.2 Grabaciones de Campo

Para el caso de las grabaciones de campo, los consejos expresados en la sección anterior también pueden serle útiles. Asimismo, a continuación se expresan algunos consejos adicionales:

- La mayoría de los dispositivos estarán alimentados por batería. Asegúrese de mantener siempre pilas de reserva para su equipo. Pruebe los dispositivos antes de cada grabación (puede hacer uso de la lista de verificación).
- Sólo en algunos casos, se encontrará con instalaciones que no cuenten con un sistema de corriente eléctrica óptimo. Líneas de alta tensión al aire libre muy a menudo tienen

mala conexión a tierra y puede causar zumbidos de 50 Hz o 60 Hz en sus señales. Si va a realizar una prueba preliminar, no lo haga en un laboratorio o parecido, si no en el lugar donde va a realizar las grabaciones. Es sutil mencionar que actualmente existen equipos de grabación que cuenten con medidas preventivas en su diseño para que usted no se tenga que preocupar por este punto.

- Este preparado para enfrentar entornos acústicos incontrolables. Siempre busque el mejor lugar para grabar. Siempre intente buscar lugares que carezcan, en su menor medida posible, ruido, corrientes de aire, eco, etc. Es recomendable que le pida al hablante o a todas las personas del lugar donde realizará la grabación que apaguen su teléfono celular, dispositivo de videojuegos, etc.
- También esté preparado para hacer frente a condiciones meteorológicas adversas o desagradables. Así como también a ubicaciones geográficas pesadas, o difíciles de llegar. Intente conseguir un vehículo capaz de poder llegar a todas los lugares y zonas contempladas.
- Siempre lleve consigo un dispositivo de almacenamiento USB (ya que se considera lo más práctico) que tenga la capacidad suficiente para realizar una copia de seguridad de sus archivos.
- Planee con tiempo el viaje a los lugares de grabación. Esto le permitirá realizar pruebas suficientes previas a las grabaciones. Las grabaciones de campo se caracterizan por ser de las más pesadas, así que también es recomendable que agregue lapsos de descanso a su agenda.

A.1.2.3 Cuestionarios y Formularios

Se deben preparar una serie de documentos para la fase de recolección. Como mínimo se debe proporcionar:

- Cuestionario, guion y formulario al hablante y supervisor/entrevistador.
- Formulario con el protocolo de grabación y/o metadatos.
- Cuente con una cantidad de copias de los formularios y cuestionarios que esté por encima de la suficiente.
- En caso de requerir una *declaración de transferencia de derechos de propiedad intelectual*, esta debe ser firmada por el hablante.
- En el caso de que cuente con los fondos para poder proporcionar incentivos al hablante, tenga listos los formularios y recibos necesarios.

A.1.2.4 Aspectos Legales

Debido a que los aspectos legales cambian dependiendo del país, el dar consejos detallados sobre esto resulta difícil. Sin embargo, esté preparado para enfrentar todo tipo de problemas legales en la fase de recolección. Consulte asesores jurídicos para obtener información sobre cómo formular documentos y declaraciones que necesitará en relación a los hablantes. Tenga

cuidado de que los documentos firmados se guarden en un lugar seguro. Si usted paga incentivos a los hablantes, confirme que recibieron el dinero y, en caso de ser necesario, solicite un documento donde ellos firman de recibido.

A.1.2.5 Plan de Reclutamiento

Si usted no es la persona encargada de hacer el reclutamiento de los hablantes, entonces usted es un suertudo. En caso contrario, si le es posible asignar esta tarea a una agencia externa, se lo recomiendo. Finalmente, si usted es la persona que realizará el reclutamiento de los hablantes, esté preparado para encontrarse con el problema de no obtener a los hablantes correctos, en el lugar indicado y en el tiempo indicado. En [1] se afirma que la técnica utilizada para reclutar personas depende del tipo de habla que el corpus necesita, lo cual, obviamente, tiene algo de verdad. Pero, a opinión propia, dicha afirmación siempre se verá opacada respecto a la cantidad de fondos que usted designe a los incentivos.

A.1.2.5.1 Incentivos

En el dado caso que se haya decidido retribuir al hablante con un incentivo, el dinero sigue siendo el incentivo más eficaz: *El dinero mueve al mundo*, dicen. Sin embargo, el usar el dinero como incentivo no es algo que sea fácil de distribuir por correo o servicios de entrega de paquetería, en el caso de que se estén recolectando las muestras por medio de grabaciones telefónicas. Una alternativa a lo anterior es utilizar incentivos que sean fáciles de enviar por correo o paquetería, por ejemplo, vales de descuento en alguna compra vía internet, *bitcoins*, etc.

El valor de los incentivos más o menos se puede estimar respecto al tiempo que el hablante tiene que gastar en la grabación. Algo muy útil es el realizar, previamente a la fase de recolección, una encuesta cuyo fin sea el averiguar la cantidad mínima de dinero que deba ser utilizada sobre cada incentivo. Debido a que cada país cuenta con su propia moneda nacional (excepto algunos casos), el recomendar una cantidad es algo inútil gracias a la variable estabilidad económica en muchos países.

A.1.3 Post-procesamiento

El *post-procesamiento* incluye todas las etapas de procesamiento posteriores a la fase de recolección, estas etapas van desde las señales de los archivos de las grabaciones a la versión o distribución final del corpus. Algunas de las siguientes tareas de procesamiento pueden que no sean del todo necesarias para la finalidad de su corpus; sin embargo, algunas de ellas sí lo son (marcadas con un asterisco):

- Transferencia de archivos.
- Asignación de nombre a los archivos*.

- Filtrado.
- Editado.
- Remuestreo.
- Conversión de formato*.
- Conversión de formato especial para anotaciones.
- Detección automática de errores*.

Por favor, tenga en cuenta que algunas de estas tareas pueden ser aplicadas antes o después de la fase de anotación.

Considere esta sección muy relevante para la producción de un corpus del habla, debido a que los costos, mano de obra y tiempo necesarios para el post-procesamiento son a menudo descuidados o, al menos, groseramente subestimados. Revise detenidamente esta sección antes de calcular los costos totales del desarrollo de su corpus y para que tome en cuenta las etapas de post-procesamiento necesarias para la producción del mismo.

A.1.3.1 Transferencia de Archivos

Es posible utilizar dispositivos de grabación que no sean computadoras y almacenen los datos de las señales de voz en medios que no se pueden leer directamente desde su ordenador o interfaz de post-procesamiento. Si no es el caso, siéntase libre de poder saltarse esta parte. En caso contrario, el primer paso del post-procesamiento es, por supuesto, la transferencia los datos de su dispositivo de grabación a un archivo de señal. Contemple procedimientos de verificación para asegurarse de que no haya pérdida de datos durante la transferencia.

A.1.3.2 Asignación de Nombre a los Archivos

Lo primero que debe hacer después de una sesión de grabación es la correcta asignación de los nombres de archivo a los archivos de señales. Aunque es posible utilizar una terminología interna durante el post-procesamiento, se recomienda no hacerlo para evitar confusiones innecesarias. Si está utilizando listas de verificación, agregue un campo en el que el supervisor o entrevistador puedan comprobar que el o los nombres de archivos asignados en la respectiva sesión fueron correctos.

Para darse una idea de cómo asignar los nombres de los archivos, puede utilizar como ejemplo la estructura que siguieron los nombres de archivos de Entendámonos en la Sección 4.5.4.

A.1.3.2 Editado

Dependiendo de la finalidad de la grabación que se está tratando, podría ser necesario remover ciertos segmentos de una grabación larga. Por ejemplo, puede que usted se encuentre

con una persona que no tenga la capacidad para leer el cuestionario, lo que lo llevará a realizar una entrevista oral, lo que hará que su voz quede atrapada en la grabación. Entonces, usted terminará con un archivo de señal cuyo contenido será la grabación completa, no sólo del hablante, sino incluyéndolo a usted realizando las preguntas o instrucciones de la entrevista.

Para cortar los segmentos de señal correspondientes y dividir en archivos individuales con nombres correctos usted puede utilizar algún tipo de procedimiento automático o hacerlo manualmente con un editor de sonido o una combinación de ambos. Si el corpus consiste en grabaciones de campo, se recomienda realizar un editado manual, debido a la cantidad de sonidos que se pueden presentar es muy amplia, por lo que configurar algún proceso automático sería casi imposible.

A.1.3.3 Filtrado

En algunos casos puede que tenga que filtrar la señal de grabación. Esto es necesario cuando se tiene pensado realizar un *downsample* sobre la señal, también se puede dar el caso cuando uno encuentra algún tipo de ruido o perturbación constante sobre la señal. Usted puede diseñar su propio filtro o utilizar herramientas que vengan con el software de edición de sonido que decidió utilizar. En el caso de la versión final o distribución de su corpus contemple archivos de audio de grabaciones filtradas y no filtradas, se recomienda el uso de un sufijo diferente para los archivos de señales filtradas para evitar la confusión y el doble filtrado de los mismos archivos.

A.1.3.4 Remuestreo

Muy a menudo se encontrará con la situación en la que el dispositivo de grabación no se graba con la frecuencia de muestreo deseada como se especifica en el corpus final. Un caso típico es el de grabación con una *grabadora DAT*, que por lo general sólo permite que la frecuencia de muestreo sea de 48 kHz o 44.1 kHz . Estas tasas de muestreo altas se requieren para las grabaciones de alta calidad, sin embargo no para el habla, donde una frecuencia de muestreo máxima de 22.05 kHz es suficiente. Esto no quiere decir que no se deba utilizar una frecuencia de muestreo alta para las grabaciones de su corpus, pero es un factor que se debe considerar ya que podría ahorrar mucho espacio respecto al tamaño de las grabaciones y el corpus en general.

Para ahorrar espacio en la distribución final del corpus, las señales tienen que pasar por un proceso de *downsample*. Antes de realizar el *downsample* hay que estar seguros de que la señal no contiene frecuencias que sean más altas que la mitad de la frecuencia de muestreo deseada (*teorema de muestreo de Nyquist-Sannon*).

Tenga en cuenta que re-muestreo en la mayoría de los casos provoca una degradación de la calidad de la señal. La calidad de la señal después de haber sido remuestreada depende del algoritmo utilizado y el formato de muestra.

A.1.3.5 Conversión de Formato

Lo más probable es que usted tendrá que convertir sus archivos de señales finales en un formato estándar como se indica en la especificación de corpus. Recuerde que cada formato tiene sus propias características, por lo que sería prudente volver a revisar la Sección A.1.1.7.4.1 para volver a meditar la decisión sobre el formato final de los archivos de audio.

A.1.3.6 Detección Automática de Errores

La *detección automática de errores* denota todos los procedimientos de verificación que se pueden llevar a cabo de forma automática. Siempre que sea posible, incluya comprobaciones después de cada etapa del post-procesamiento, sobre todo después de aquellas etapas que requieran trabajo manual. Algunos de las posibles verificaciones que se pueden llevar a cabo de forma automática son:

- Buscar archivos vacíos
- Verifique la duración correcta de los archivos de audio
- Compruebe terminología correcta
- Compruebe si el archivo contiene habla o sólo silencio/ruido.
- Compruebe el número requerido de archivos por sesión de grabación
- Compruebe el espacio en disco total de una sesión de grabación
- Compruebe relación S/N (señal/ruido).
- Compruebe el formato de los archivos.

Una manera fácil de poner en práctica tales controles automáticos es el uso de un *shell* de su respectivo sistema operativo. La mayoría de los controles son comandos sencillos y se pueden implementar fácilmente en un lenguaje de script como *awk*, *perl* o *bash*.

A.2 Teoría de Wavelets

Para una mejor comprensión de la transformada Wavelet, se presentarán primero algunos conceptos matemáticos necesarios. Se definirán los conceptos de *Espacios de Hilbert*, *Ortogonalidad* y *Bases Ortogonales*.

Por razones de claridad comenzaremos definiendo el espacio métrico sobre el que vamos a trabajar, el cual es el espacio de $L^2[-\infty, \infty]$ de Hilbert.

Finalmente, al igual que en la Sección A.1, se aclara que los conceptos de esta sección están basados en los trabajos: [8] al [11], [13], [24], [28] al [30] (9 en total).

A.2.1 Espacios de Hilbert

El espacio H de Hilbert es un espacio vectorial cuyos elementos pertenecen al plano complejo \mathbf{C} . Sea \mathbf{H} el conjunto de elementos del espacio H . Los vectores complejos de este conjunto pueden ser sumados con las reglas usuales de la aritmética de vectores (propiedad aditiva) y multiplicados por escalares (números complejos).

El espacio H está dotado de una métrica y de un producto interno. Consideraremos en particular el espacio H formado por funciones vectoriales f_n . Si f y g son funciones del conjunto \mathbf{H} de H , el producto interno para este conjunto de funciones es un escalar definido por

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f^*(x)g(x)dx, \quad (\text{A.1.1})$$

donde $f^*(x)$ es el complejo conjugado de $f(x)$. El producto escalar o interno de la función f con sí misma es un número real no negativo. En particular, si la función $f \in \mathbf{C}$, entonces satisface la condición:

$$\int_{-\infty}^{\infty} |f(t)|^2 dt < \infty, \quad (\text{A.1.2})$$

este espacio métrico recibe el nombre de *Espacio de Hilbert* $L^2[-\infty, \infty]$.

A.2.2 Ortogonalidad y Bases Ortonormales

Se dice que dos vectores x e y son ortogonales en un Espacio de Hilbert H si su producto interno es cero:

$$\langle x, y \rangle = 0$$

Se le llama *conjunto ortogonal* a aquel conjunto de vectores que en cualquier par de sus elementos es *ortogonal*. Además, este conjunto es *ortonormal* si la norma de los vectores es igual a uno:

$$\|x\| = \sqrt{\langle x, x \rangle} = 1$$

También se define la *base ortonormal* de \mathbf{H} como un conjunto *ortonormal maximal* en \mathbf{H} si cualquier vector en \mathbf{H} puede ser representado como el límite de las combinaciones lineales de los elementos de una base ortonormal.

A.2.3 Bases de la Función de Escala

Las *funciones de escala* juegan el papel de funciones promedio. La correlación entre la función de escala y una función continua arbitraria produce la aproximación promediada de la última.

La función de escala básica $\phi(t)$, dilatada por un factor de escala 2^i , es desplazada con un factor de escala discreto de translación k ,

$$\varphi_{i,k}(t) = 2^{-\frac{i}{2}} \phi(2^{-i}t - k) \quad (\text{A.2.1})$$

Las funciones de escala básica $\phi(t)$ que se emplean satisfacen la condición de ortogonalidad, tal que las traslaciones discretas $\{\phi(t - k)\}$ con $k \in \mathbb{Z}$, forman un conjunto ortonormal. La proyección de una función $f(t) \in L^2(\mathbb{R})$ en la base ortonormal $\{\phi(t - k)\}$ es una correlación entre la función $f(t)$ original y la función de escala $\phi(t)$ muestreada a intervalos enteros.

Como resultado de la proyección de $f(t)$ en la base de la función de escala, se obtiene una aproximación menos detallada de $f(t)$. Todas las aproximaciones de $f(t)$ forman un subespacio $V_0 \in L^2(\mathbb{R})$. El espacio vectorial V_0 puede ser interpretado como el conjunto de todas las posibles aproximaciones de la función en $L^2(\mathbb{R})$ generado por el conjunto ortonormal $\{\phi(t - k)\}$.

Las funciones de escalas para todas las escalas $s = 2^i$ con $i \in \mathbb{Z}$, generadas a partir de la misma $\phi(t)$, son todas de forma similar. Debido a que la función de escala básica $\phi(t)$ genera la base ortonormal $\{\phi(t - k)\}$ de V_0 , con un paso de translación entero, la función de escala dilatada $\phi(t/2)$ generará la base ortonormal $\{\phi(2^{-1}t - k)\}$ de V_1 con un paso de translación igual a 2, y $\phi(t/4)$ generará la base ortonormal $\{\phi(2^{-2}t - k)\}$ de V_2 con un paso de translación igual a 4, y así sucesivamente. Existe entonces un conjunto de bases ortogonales de las funciones de escala. Cada base de la función de escala es ortonormal en el espacio de la misma escala:

$$\langle \phi_{i,k}, \phi_{i,n} \rangle = \delta_{k,n} \quad (\text{A.2.2})$$

para todo k y $n \in \mathbb{Z}$.

Las proyecciones $L^2(\mathbb{R})$ sobre el conjunto de bases Ortonormales de la función de escala, forman un conjunto de subespacios V_i . Cada subespacio V_i es el conjunto de todas las posibles aproximaciones de la función en $L^2(\mathbb{R})$ generado por la base ortonormal de la función de escala $\phi(2^{-i}t - k)$. El subespacio V_i es abarcado por la base ortonormal de la función de

escala en el nivel de resolución i . Por lo tanto, la función de escala $\phi(t)$ genera los subespacios del análisis multiresolución.

Las aproximaciones de una función $f(t)$ en diferentes resoluciones deben ser similares, ya que son todas generadas por la misma función de escala con escalas diferentes. Los espacios de aproximación V_i pueden ser, entonces, deducidos unos de otros por simple dilatación:

$$f(t) \in V_i \Leftrightarrow f(2t) \in V_{i-1} \quad (\text{A.2.3})$$

Toda la información útil para calcular la función de aproximación en el nivel de menor resolución i , está contenida en la función de aproximación en el nivel de mayor resolución $(i - 1)$. Entonces, V_i es un subespacio de V_{i-1} .

A.2.4 Análisis Multiresolución

El análisis multiresolución es una técnica que permite analizar señales en múltiples bandas de frecuencia. Consiste en una secuencia de subespacios cerrados V_i en $L^2(\mathbb{R})$:

$$\dots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \dots \subset L^2(\mathbb{R}) \quad (\text{A.2.4})$$

Cuando la resolución se incrementa con i tendiendo a $-\infty$, la función aproximada debería converger a la función original. Esto es:

$$\overline{\bigcup_i V_i} = L^2(\mathbb{R}) \quad (\text{A.2.5})$$

Por el contrario, cuando la resolución se decrecienta a cero con i tendiendo a ∞ , las aproximaciones contienen cada vez menos información y convergen a cero:

$$\overline{\bigcap_i V_i} = \{0\} \quad (\text{A.2.6})$$

con $i \in \mathbb{Z}$.

A.2.5 Bases Wavelet

Debido a que la proyección de una función sobre la base de la función de escala ortonormal es una aproximación menos detallada de la función en un nivel de resolución particular, se pierde algo de información en el proceso, esto significa que la función escala ϕ no es completa en cualquier nivel. Por lo tanto, se usan las proyecciones sobre otras funciones, denominadas *wavelets Ortonormales* (o simplemente *wavelets*), para obtener la información complementaria de los detalles de la función.

Como se verá más adelante, las wavelets son generadas a partir de la *wavelet madre* $\psi(t)$ por las traslaciones y dilataciones discretas

$$\psi_{i,k}(t) = 2^{-i/2}\psi(2^{-i}t - k) \quad (\text{A.2.7})$$

Cuando la transformada de Fourier $\psi(w)$ de la wavelet madre satisface la condición de ortogonalidad, las traslaciones discretas de la wavelet madre $\{\psi(2^{-i}t - k)\}$ forman una base ortonormal para cada escala 2^i . Más aún, en el mismo nivel de resolución, el conjunto de traslaciones wavelet es ortogonal al conjunto de traslaciones de la función de escala en el espacio de la misma resolución.

$$\langle \phi_{i,k}, \psi_{i,n} \rangle = 2^{-i} \int \phi_i(t - k)\psi_i(t - n) dt = 0 \quad (\text{A.2.8})$$

para todo k y $n \in \mathbb{Z}$.

La proyección de $f(t)$ sobre las bases wavelet ortonormales es una correlación entre $f(t)$ y $\psi(t)$ muestreada a intervalos discretos. Las proyecciones de las funciones en $L^2(\mathbb{R})$ sobre la base wavelet ortonormal $\{\psi(2^{-i}t - k)\}$, forman un subespacio W_i . El subespacio W_i es abarcado por $\{\psi(2^{-i}t - k)\}$.

Como la base wavelet $\{\psi(2^{-i}t - k)\}$ es ortogonal a la base de la función de escala $\{\phi(2^{-i}t - k)\}$, dentro de la misma escala, el subespacio W_i es el complemento ortogonal del subespacio V_i :

$$W_i \perp V_i \quad (\text{A.2.9})$$

Tanto V_i como W_i son subespacios de V_{i-1} : $V_i, W_i \in V_{i-1}$, y en razón de que W_i es el complemento ortogonal de V_i , el subespacio V_{i-1} es la suma directa de V_i y W_i :

$$V_{i-1} = V_i \oplus W_i \quad (\text{A.2.10})$$

A.2.5.1 Transformada Wavelet

De manera muy general, la *Transformada Wavelet* de una función $f(t)$ es la descomposición de $f(t)$ en un conjunto de funciones $\psi_{s,\tau}(t)$, que forman una base y son llamadas las “*Wavelets*”. La transformada Wavelet se define como:

$$W_i(s, \tau) = \int f(t)\psi_{s,\tau}^*(t) dt \quad (\text{A.2.11})$$

Las Wavelets son generadas a partir de la traslación y cambio de la escala de una misma función wavelet $\psi(t)$, llamada la “*Wavelet madre*”, y se define como:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right), \quad (\text{A.2.12})$$

donde s es el factor de escala, y τ es el factor de traslación.

Las wavelets $\psi_{s,\tau}(t)$ generadas de la misma función wavelet madre $\psi(t)$ tienen diferente escala s y ubicación τ , pero tienen todas la misma forma. Se utilizan siempre factores de escala $s > 1$. Así, cambiando el valor de s se cubren rangos diferentes de frecuencias. Valores grandes del parámetro s corresponden a frecuencias de menor rango, o una escala grande de $\psi_{s,\tau}(t)$. Valores pequeños de s corresponden a frecuencias de mayor rango o una escala muy pequeña de $\psi_{s,\tau}(t)$.

A.2.5.2 Wavelets Ortonormales y Discretas

Cuando la función $f(t)$ es continua y las wavelets son continuas con factor de escala y traslación discretas, la transformada Wavelet resulta en una serie de coeficientes wavelets, y es llamada la descomposición en *Series Wavelet*.

La función $f(t)$ puede ser reconstruida desde los coeficientes discretos la función $W_f(s, \tau)$, de la siguiente manera:

$$f(t) = A \sum_s \sum_\tau W_f(s, \tau) \psi_{s,\tau}(t), \quad (\text{A.2.13})$$

donde A es una constante que no depende de $f(t)$.

A estas funciones wavelets continuas con factores de escala y traslación discretos se les denomina *Wavelets discretas*. Los factores de escala y traslación de las wavelets discretas pueden ser expresados como:

$$s = s_0^i \text{ y } \tau = k\tau_0 s_0^i \quad (\text{A.2.14})$$

donde el exponente i y la constante k son enteros, y $s_0 > 1$ es un paso fijo de dilatación.

El factor de traslación τ depende del paso de dilatación s (Ecuación A.2.14). Entonces, a partir de la Ecuación A.2.12 y con la Ecuación A.2.14, las correspondientes wavelets discretas quedan expresadas como:

$$\psi_{i,k}(t) = s_0^{-\frac{i}{2}} \psi\left(s_0^{-i}(t - k\tau_0 s_0^i)\right) = s_0^{-\frac{i}{2}} \psi(s_0^{-i}t - k\tau_0) \quad (\text{A.2.15})$$

A través de la Ecuación A.2.11, la transformada Wavelet de una función continua es realizada a frecuencias y tiempos discretos que corresponden a muestreos con distintas traslaciones (tiempo) y distintas dilataciones (cambios de escala).

El paso de muestreo en tiempo es pequeño para el análisis utilizando wavelets de pequeña escala, mientras que es grande para el análisis con wavelets de gran escala. La posibilidad de variar el factor de escala s permite usar wavelets de escala muy pequeña para concentrar el análisis en singularidades de la señal. Cuando solo los detalles de la señal son de interés, unos pocos niveles de descomposición son necesarios. Por lo tanto el análisis wavelet provee una forma más eficiente de representar señales transitorias.

A modo de ejemplo, podemos hacer una analogía entre el análisis de Wavelet y el microscopio. Así, el factor de escala s_0^i corresponde al aumento o resolución del microscopio y el factor de traslación τ corresponde a la ubicación donde se hace la observación con el microscopio. Si queremos mirar detalles muy pequeños, el aumento y la resolución deben ser grandes, lo que se corresponde con un i grande y negativo. Esto da lugar a una función wavelet muy concentrada, y a pasos de traslación pequeños. Para un valor de i grande y positivo, la wavelet se extiende y los pasos de traslación son adaptados a esa amplitud.

Eligiendo adecuadamente $\psi(t)$ y los parámetros s_0, τ_0 es posible lograr que las funciones $\psi_{s,\tau}(t)$ constituyan una base ortonormal de $L^2(\mathbb{R})$.

De esta forma, si las funciones wavelets discretas forman una base ortonormal, una función $f(t)$ de soporte finito puede ser reconstruida como una suma de los coeficientes wavelets discretos $W_f(s, \tau)$ multiplicados por las funciones de la base, como sigue:

$$f(t) = \sum_s \sum_\tau W_f(s, \tau) \psi_{s,\tau}(t) \quad (\text{A.2.16})$$

Una descomposición wavelet ortonormal no posee información redundante y representa la señal en forma unívoca. Una base wavelet ortonormal es posible con wavelets con factores de traslación y dilatación discretos. Por lo tanto, para estas funciones wavelets discretas ortogonales, los productos internos son iguales a cero:

$$\int \psi_{i,k}^*(t) \psi_{m,n}(t) dt = \begin{cases} 1 & \text{si } i = m \text{ y } k = n \\ 0 & \text{en cualquier otro caso} \end{cases} \quad (\text{A.2.17})$$

En 1986, *Meyer* y *Mallat* demostraron que la descomposición y reconstrucción wavelet ortonormal podrían ser implementadas en el marco del análisis multiresolución de señales.

A.2.5.3 Relación dos-escala

Con sus traslaciones discretas, las funciones de escala y las de wavelets forman dos bases ortonormales en cada nivel de resolución. Las funciones de escala y las wavelets en múltiples niveles de resolución son la versión dilatada de la función de escala básica y de la wavelet madre, respectivamente.

Sea $\phi(t)$ la función de escala básica cuyas traslaciones generan el subespacio V_0 . Entonces $\phi(t)$ puede ser expresada como una combinación lineal de la suma ponderada del conjunto $\{\phi(2t - k)\}$ generado por $\phi(2t)$. Así las funciones de escala en dos niveles de resolución adyacentes satisfacen la relación dos-escala:

$$\phi(t) = \sum_k p(k)\phi(2t - k) \quad (\text{A.2.18})$$

que puede ser considerada como la proyección de la función $\phi(t) \in V_0$ en el subespacio de mayor resolución V_{-1} . Esta relación es la ecuación fundamenta en el análisis multiresolución. La secuencia $p(k)$ es el *coeficiente de interescala*, correspondiente a un filtro discreto paso-bajo.

Sea $\psi(t) \in V_0$ la wavelet madre, la cual puede ser desarrollada en la base ortonormal de la función de escala $\{\phi(2t - k)\}$ en V_{-1} como:

$$\psi(t) = \sum_k q(k)\phi(2t - k) \quad (\text{A.2.19})$$

Donde la secuencia $q(k)$ es el coeficiente de interescala, correspondiente a un filtro paso-alto. Esta relación dos-escala permite generar wavelets a partir de las funciones de escala.

En el lado izquierdo de las Ecuaciones A.2.18 y A.2.19, $\phi(t)$ y $\psi(t)$ son continuas. En el lado derecho de las relaciones, los coeficientes de interescala, $p(k)$ y $q(k)$ son discretos.

A.2.5.4 Descomposición Wavelet (Algoritmo piramidal)

Sea la función $f(t) \in V_0$ que puede ser representada como la combinación lineal de las funciones de escala trasladadas $\phi(t - k)$ en V_0

$$f(t) = \sum_k c_0(k)\phi(t - k) \quad (\text{A.2.20})$$

con los coeficientes

$$c_0(k) = \langle f, \phi_{0,k} \rangle = \int f(t)\phi(t - k) dt \quad (\text{A.2.21})$$

La función a ser analizada pertenece a V_0 , el cual corresponde al nivel de digitalización inicial al comenzar la descomposición. E el siguiente nivel de menor resolución $i = 1$, existen dos subespacios mutuamente ortogonales $\{\phi_{i,k}(t)\}$ y $\{\psi_{i,k}(t)\}$, respectivamente. Debido a que V_0 es una suma directa de V_1 y W_1 , existe una única forma de expresar una función $f(t) \in V_0$, como una combinación lineal de funciones v_1 y w_1 , donde $v_1 \in V_1$ y $w_1 \in W_1$. En

particular, la función $f(t) \in V_0$ puede descomponerse en sus componentes a lo largo de V_1 y W_1 :

$$f = (P_1 + Q_1)f \quad (\text{A.2.22})$$

donde las dos componentes son las proyecciones ortonormales de $f(t)$ sobre V_1 y W_1 :

$$(a) P_1 f = \sum_n c_1(n) \phi_{1,n} \quad (\text{A.2.23})$$

$$(b) Q_1 f = \sum_n d_1(n) \psi_{1,n}$$

Multiplicando ambos lados de la Ecuación A.2.22 por $\phi_{1,k}$ y calculando los productos internos, se obtiene:

$$\langle \phi_{1,k}, f \rangle = \langle \phi_{1,k}, P_1 f \rangle \quad (\text{A.2.24})$$

Haciendo lo mismo en la Ecuación A.2.23(a) pero multiplicando por $\phi_{1,n}$ y usando la Ecuación A.2.20, se obtiene:

$$c_1(k) = \langle \phi_{1,k}, f \rangle = \langle \phi_{1,k}, P_1 f \rangle = \sum_n \langle \phi_{1,k}, \phi_{0,n} \rangle c_0(n) \quad (\text{A.2.25})$$

donde el producto interno de los dos conjuntos de la función de escala $\{\phi_{1,k}\}$ y $\{\phi_{0,n}\}$ se puede calcular como

$$\begin{aligned} \langle \phi_{1,k}, \phi_{0,n} \rangle &= 2^{-1/2} \int \phi\left(\frac{t}{2} - k\right) \phi(t - n) dt \\ &= 2^{-1/2} \int \phi(t) \phi(2t - (n - 2k)) dt \end{aligned} \quad (\text{A.2.26})$$

Sustituyendo $\phi(t)$ por la relación dos-escala en la Ecuación A.2.26 y usando la ortonormalidad del conjunto $\{\phi(2t)\}$ se obtiene

$$c_1(k) = 2^{-1/2} \sum_n p(n - 2k) c_0(n) \quad (\text{A.2.27})$$

La secuencia $c_1(k)$ o tendencia contiene los coeficientes del desarrollo de la función continua $f(t)$ en la base de la función de escala continua $\{\phi_{1,k}\}$ en V_1 . La secuencia $c_1(k)$ representa la versión suavizada de los datos originales $c_0(n)$.

Simultáneamente, multiplicando ambos lados de las Ecuaciones A.2.22 y A.2.23(b) por la wavelet $\psi_{1,n}$ y calculando los productos internos, se obtiene:

$$d_1(k) = \langle \psi_{1,k}, Q_1 f \rangle = \langle \psi_{1,k}, f \rangle = \sum_n \langle \psi_{1,k}, \phi_{0,n} \rangle c_0(n) \quad (\text{A.2.28})$$

y siguiendo los pasos aplicados para la obtención de $c_1(k)$ se llega a que:

$$d_1(k) = 2^{-1/2} \sum_n q(n - 2k) c_0(n) \quad (\text{A.2.29})$$

De acuerdo con la Ecuación A.2.22, la proyección ortonormal $Q_1 f$ sobre W_1 es la *información de detalle* de $f(t)$. La secuencia $d_1(k)$ representa la diferencia entre la $f(t)$ original y la aproximación $P_1 f$, y se conoce como los *coeficientes wavelets discretos*.

La descomposición en aproximaciones suavizadas y detalles a menor resolución se puede continuar tanto como se desee. Generalizando,

$$\begin{aligned} P_{i-1} f &= P_i f + Q_i f = \sum_k c_i(k) \phi_{i,k} + \sum_k d_i(k) \psi_{i,k}, \\ c_i(k) &= 2^{-1/2} \sum_n p(n - 2k) c_{i-1}(n), \\ d_i(k) &= 2^{-1/2} \sum_n q(n - 2k) c_{i-1}(n). \end{aligned} \quad (\text{A.2.30})$$

Las secuencias $c_i(n)$ y $d_i(n)$ pueden ser calculadas a partir de $c_{i-1}(n)$ por filtrado iterativo.

De esta manera, iterando hasta un nivel de resolución M , donde M toma un valor determinado, se puede representar la función original $f(t)$ por una serie de funciones de detalle más una aproximación gruesa:

$$\begin{aligned} f(t) &= P_M f + Q_M f + Q_{M-1} f + \dots + Q_1 f, \\ f(t) &= \sum_{k \in \mathbb{Z}} 2^{-\frac{M}{2}} c_M(k) \phi(2^{-M} t - k) + \sum_{i=1}^M \sum_{k \in \mathbb{Z}} 2^{-\frac{i}{2}} d_i(k) \psi(2^{-i} t - k) \end{aligned} \quad (\text{A.2.31})$$

La Ecuación A.2.31 es la descomposición $f(t)$ en Series Wavelet. En esta descomposición wavelet las bases de la función de escala y las bases wavelet son todas continuas. Los coeficientes de aproximación $c_M(k)$ y los coeficientes de detalle $d_i(k)$ con $i = 1, 2, \dots, M$ y $k \in \mathbb{Z}$ son discretos.

Los coeficientes $c_1(n)$ y $d_1(n)$ se pueden calcular con un algoritmo discreto implementado por la aplicación recursiva de filtros discretos paso-alto y paso-bajo a las aplicaciones discretas $c_{i-1}(n)$. Este algoritmo es conocido como el *algoritmo piramidal de Mallat* o *algoritmo piramidal*. Los dos primeros pasos del algoritmo para calcularla descomposición

wavelet se muestran en la Figura A.2, asimismo en la Figura 5.4 se muestra una versión implementada del mismo.

A.2.6 Wavelet de Daubechies

Por lo que se vio anteriormente solo queda elegir el tipo de wavelet que se usará en el algoritmo piramidal, existen una gran variedad de wavelets. Una importante familia de wavelets es la *Daubechies* propuestas por *Ingrid Daubechies* en la década de los '80s.

La familia de wavelets Daubechies es conocida como la familia *dbn*, donde cada miembro posee un número $n = 1, 2, 3, \dots$ con $n \in \mathbb{N}$, y a la vez cada *dbn* cuenta con $2n$ coeficientes filtro. La wavelet *db1* es conocida como la wavelet de Haar, la cual posee 2 coeficientes filtro.

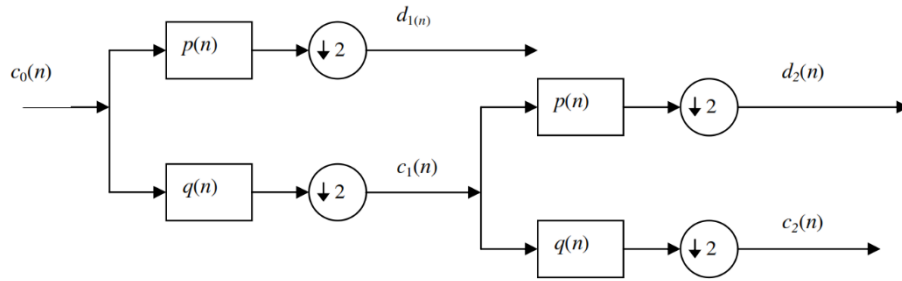


Figura A.2. Esquema de la descomposición en Series Wavelet.

La siguiente es la wavelet *db2*, y está formada por 4 coeficientes filtro. Los cuales son:

$$\begin{aligned}
 h_0 &= \frac{1+\sqrt{3}}{4} & h_1 &= \frac{3+\sqrt{3}}{4} & h_2 &= \frac{3-\sqrt{3}}{4} & h_3 &= \frac{1-\sqrt{3}}{4} \\
 g_0 &= \frac{1+\sqrt{3}}{4} & g_1 &= -\frac{3+\sqrt{3}}{4} & g_2 &= \frac{3-\sqrt{3}}{4} & g_3 &= -\frac{1-\sqrt{3}}{4}
 \end{aligned}
 \tag{A.2.32}$$

dando como resultado las siguientes funciones de aproximación y detalle

$$\phi(t) = \frac{1+\sqrt{3}}{4} \phi(2t) + \frac{3+\sqrt{3}}{4} \phi(2t-1) + \frac{3-\sqrt{3}}{4} \phi(2t-2) + \frac{1-\sqrt{3}}{4} \phi(2t-3)
 \tag{A.2.33}$$

$$\psi(t) = \frac{1+\sqrt{3}}{4} \phi(2t) - \frac{3+\sqrt{3}}{4} \phi(2t-1) + \frac{3-\sqrt{3}}{4} \phi(2t-2) - \frac{1-\sqrt{3}}{4} \phi(2t-3)
 \tag{A.2.34}$$

Las funciones de las Ecuaciones A.2.33 y A.2.34 se pueden apreciar en la Figura A.3.

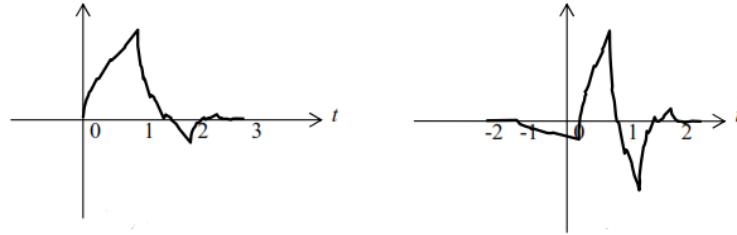


Figura A.3. Función de aproximación $\phi(t)$ (izquierda) y función de detalle $\psi(t)$.

A.3 Aprendizaje Automático

Esta sección aborda los tópicos de *aprendizaje automático* (*machine learning*) utilizados en el presente trabajo.

A.3.1 Aprendizaje Automático: Naive Bayes

El aprendizaje de máquina es parte de la Inteligencia Artificial, y se encarga de la cuestión de construir programas de computadora que sean capaces de mejorar con el uso de la experiencia [31].

Existen diversos métodos para construir estos sistemas de aprendizaje automático, algunos provenientes de métodos de minería de datos, y otros que son en sí mismos un área con fundamentos propios. Entre ellos podemos encontrar los métodos bayesianos, las redes neuronales, los árboles de decisión. Entre los métodos de tipo bayesiano destaca *Naive Bayes*, el cual es sencillo de implementar, se utiliza tanto para biclasificación como para multclasificación.

Aunque está diseñado para variables nominales, mediante una discretización es posible manejar variables numéricas. Su utilización en este trabajo obedece a las características antes mencionadas y a los resultados mostrados en [11].

Un clasificador Naive Bayes utiliza todas las variables y considera que las contribuciones de cada una de ellas son igualmente importantes en la decisión, además considera a cada variable independiente una de otra dada una clase, lo cual no corresponde en la realidad, no todas las variables son igualmente importantes e independientes una de otra [32]. La independencia de las variables permite la multiplicación de las probabilidades con el fin de calcular la verosimilitud (*likelihood*) de la clase.

Naive Bayes está basado en la *regla de Bayes*, la cual dice que si se tiene una hipótesis h y una evidencia E que es relevante para la hipótesis, entonces

$$P(h|E) = \frac{P(E|h)P(h)}{P(E)}, \quad (\text{A.3.1})$$

donde $P(h)$ es la probabilidad de h , $P(E)$ es la probabilidad de E , $P(E|h)$ es la probabilidad condicional de E dado h y $P(h|E)$ es la probabilidad condicional de h dado E .

Así, basados en la regla de Bayes, se puede calcular la hipótesis máxima a posteriori (*maximum a posteriori hypothesis*)

$$h_{MAP} = \arg \max_{h \in H} P(h|E) = \arg \max_{h \in H} \frac{P(E|h)P(h)}{P(E)} = \arg \max_{h \in H} P(E|h)P(h) \quad (\text{A.3.2})$$

donde H es el conjunto de todas las hipótesis y h_{MAP} la hipótesis máxima a posteriori. La probabilidad independiente E es omitida en h_{MAP} dado que representa una constante independiente de E .

En algunos casos se puede asumir que todas las hipótesis son igualmente probables a priori, por ejemplo $P(h_i) = P(h_j)$, así para toda $h_i, h_j \in H$, lo que quiere decir que se asume una *probabilidad a priori uniforme*, lo que facilita el cálculo de h_{MAP} dando como resultado la Ecuación A.3.3, la cual es llamada la *hipótesis de máxima verosimilitud* (*maximum likelihood hypothesis*).

$$h_{ML} = \arg \max_{h \in H} P(E|h) \quad (\text{A.3.3})$$

¿Cómo funciona el clasificador? Suponga que se tiene un conjunto de instancias de entrenamiento o prueba y un conjunto finito de clases C , la tarea del clasificador es predecir correctamente la clase de una nueva instancia con n atributos $\langle a_1, a_2, \dots, a_n \rangle$, lo que genera

$$\begin{aligned} c_{MAP} &= \arg \max_{c_j \in C} P(c_j | a_1, a_2, \dots, a_n) \\ c_{MAP} &= \arg \max_{c_j \in C} \frac{P(a_1, a_2, \dots, a_n | c_j) P(c_j)}{P(a_1, a_2, \dots, a_n)} \\ c_{MAP} &= \arg \max_{c_j \in C} P(a_1, a_2, \dots, a_n | c_j) P(c_j), \end{aligned} \quad (\text{A.3.3})$$

donde C es el conjunto de todas las clases.

Tomando la Ecuación A.3.3 y aplicándola a las 3 posibles combinaciones de las pruebas del Capítulo 5, las ecuaciones quedarían de la siguiente manera

$$c_{NB_1} = \arg \max_{c_j \in [Nahuatl, Tének]} P(c_j) \prod_i P(a_i | c_j) \quad (\text{A.3.4})$$

$$c_{NB_2} = \arg \max_{c_j \in [Nahuatl, Xiviyu]} P(c_j) \prod_i P(a_i | c_j) \quad (A.3.5)$$

$$c_{NB_3} = \arg \max_{c_j \in [Tének, Xiviyu]} P(c_j) \prod_i P(a_i | c_j) \quad (A.3.6)$$

A.3.2 Ganancia de Información

La técnica de selección de atributos llamada ganancia de información (*gain information*) es muy usada en aplicaciones de categorización de textos. Además, la ganancia de información y su modificación llamada tasa de información se basan en la *entropía*, la cual puede entenderse como el grado de desorden de la información o variables. Por tanto, si A es un atributo y C es la clase de ese atributo, las Ecuaciones A.3.7 y A.3.8 definen la entropía de la clase y la entropía de la clase dado el atributo A .

$$H(C) = - \sum_{c \in C} p(c) \log_2 p(c) \quad (A.3.7)$$

$$H(C|A) = - \sum_{a \in A} p(a) \sum_{c \in C} p(c|a) \log_2 p(c|a) \quad (A.3.8)$$

En la Ecuación A.3.9, la cantidad por la cual la entropía de la clase C decrece refleja la información adicional en la clase, la cual es suministrada por el atributo A y se le llama ganancia de información. A cada atributo A se le asigna un resultado basado en la ganancia de información entre sí misma y la clase.

$$IG_i = H(C) - H(C|A_i) = H(A_i) - H(A_i|C) = H(A_i) + H(C) - H(A_i, C) \quad (A.3.9)$$

Esta técnica usa variables nominales (no numéricas), por lo cual se requiere *discretización de las variables*. Para tomar la decisión de cuales variables deben ser seleccionadas, la ganancia de información usa un *criterio de selección (ranker)*, el cual enlista los atributos de acuerdo a su resultado IG y una función de restricción los filtra, cuyas variables restantes son las seleccionadas.

A.3.3 Validación Cruzada

Estimar la precisión de un modelo creado a partir de un algoritmo de aprendizaje supervisado es importante no solo para predecir su precisión respecto a futuras predicciones sino también para escoger un clasificador de un grupo dado.

La precisión de un clasificador es la probabilidad de clasificar correctamente instancias seleccionadas aleatoriamente. La validación cruzada es un método que sirve para calcular dicha precisión.

En la validación cruzada con k paquetes, el conjunto D es dividido aleatoriamente en k subconjuntos D_1, D_2, \dots, D_k mutuamente exclusivos de aproximadamente el mismo tamaño. El modelo es entrenado con el subconjunto D/D_t con $t \in k$ y evaluado $D_t k$ veces.

Lo anterior nos ayuda a entender una suposición que emerge cuando se usa validación cruzada: si el modelo es inestable para un conjunto de datos bajo un conjunto de perturbaciones introducidas por la validación cruzada, la precisión de predicción será poco estable. Si el clasificador es estable para cierto conjunto de datos, podemos esperar una estimación estable.

B

Equipo de Trabajo

Este apéndice brinda una descripción un tanto técnica de todo el equipo electrónico de trabajo, o bien *hardware*, utilizado para el desarrollo del presente trabajo. El equipo utilizado fue:

- Laptop MacBook Negra (modelo del 2008)
- Micrófono SHURE PG42-USB
- Laptop Asus N56V

B.1 Laptop: MacBook



Figura B.1. Imágenes del MacBook utilizado.

Dimensiones y peso:

- Alto: 2.75 cm
- Ancho: 32.5 cm
- Fondo: 22.7 cm
- Peso: 2,27 kg

Sonido:

- Altavoces estéreo incorporados.

Sección B.2 Micrófono: SHURE PG42-USB

- Micrófono omnidireccional incorporado.
- Entradas de audio de línea y óptico digital combinadas (miniconector).
- Salidas de audio de línea y óptico digital combinadas (miniconector).

Procesador y memoria:

- Procesador Core 2 Duo de Intel 2,4 GHz con 3 MB de caché de nivel 2 integrada y compartida que funciona a la velocidad del procesador
- Bus frontal a 800 MHz
- 2 GB (en dos módulos SODIMM de 1 GB) de SDRAM DDR2 a 667 MHz.

Almacenamiento y sistema operativo:

- Disco duro Serial ATA de 160 o 250 GB a 5.400 rpm
- Mac OS X 10.6 Snow Leopard.

Este ordenador portátil fue utilizado para realizar 107 de las 143 grabaciones que el corpus *Entendámonos* presenta en este trabajo. Concretamente, con este equipo se realizaron todas las grabaciones de náhuatl, 49 grabaciones de tének (la última se realizó con el ordenador *Asus*) y las primeras 12 grabaciones de xi'iuy.

El motivo por el cual se dejó de utilizar este portátil fue que tuvo una avería en su pantalla; concretamente el sistema electrónico del ordenador dejó de iluminar la pantalla (la pantalla si mostraba los gráficos y contenido pero sin iluminación).

B.2 Micrófono: SHURE PG42-USB

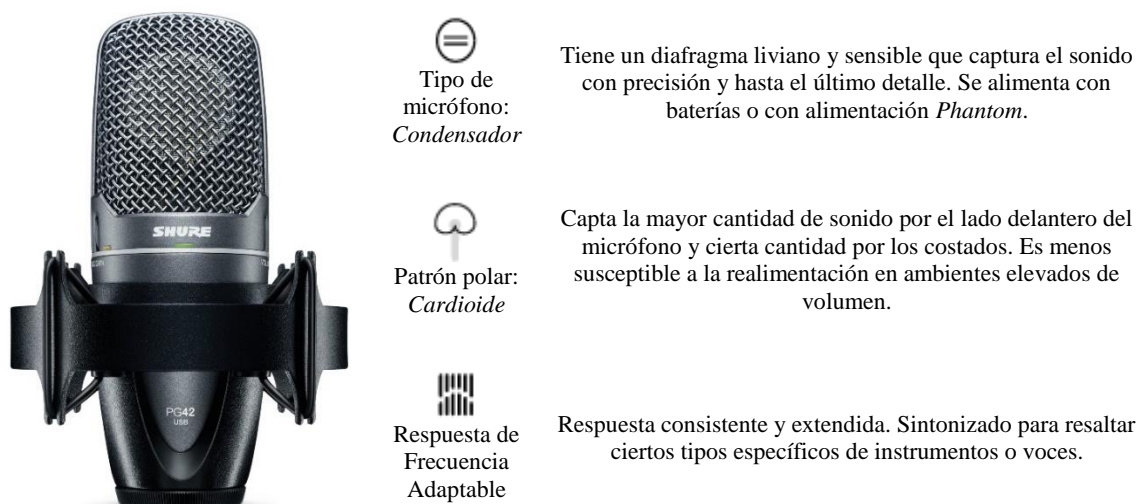


Figura B.2. Micrófono SHURE PG42-USB (con su *araña* enroscada) y algunas de sus características.

Apéndice B. Equipo de Trabajo

El micrófono está diseñado con alta sensibilidad para un rendimiento optimizado, se conecta fácilmente a al ordenador para grabación digital. Además posee una cápsula cardioide con diafragma grande ajustada para una respuesta de frecuencia adaptada a la voz.

El micrófono incluye un cable USB de *9.8 pies (3m)*, adaptador para pedestal de micrófono (araña) y estuche acolchonado.

Sus características son:

- Conectividad USB le permite grabar en digital a cualquier hora, en cualquier lugar y a donde quiera llevar su computadora.
- Monitoreo sin latencia para grabar en tiempo real sin la desorientación del retardo.
- Preamplificador integrado con control de ganancia.
- Control de mezcla para monitorear la señal del micrófono con el audio pregrabado.
- Conector para audífonos de *1/8"*.
- Compatible con *Windows 7, 8, Vista, XP, 2000* y *Mac OS X (10.1* o versiones posteriores).

Sus especificaciones son:

- Respuesta de Frecuencia: de *20* a *20,000 Hz*.
- Patrón polar: Cardioide.
- Nivel de salida: *-29 dBFS/Pa* (a mínima ganancia).

Este micrófono resultó ser una opción muy agradable para realizar las grabaciones, no presentó ningún tipo de problema y el maletín con el que viene para su transportación es cómodo y compacto. Tal vez, el único detalle incomodo del micrófono es su preamplificador de control de ganancia de información, ya que este se puede desajustar si uno no agarra el micrófono con cuidado. Pero, de nuevo, este es un detalle minúsculo ya que eso se soluciona fácilmente ajustando de nuevo el nivel de ganancia requerido.

B.3 Laptop: Asus N56V



Figura B.3. Imágenes de la laptop Asus N56VJ utilizada.

Sección B.3 Laptop: Asus N56V

Dimensiones y peso:

- 380 x 255 x 27.2 ~34.0 pulgadas.
- 2.7 kg (con la batería de 6 celdas).

Sonido:

- 2 altavoces y un micrófono integrados.
- Bang & Olufsen ICEpower®.
- SonicMaster Premium.
- Soporte de sub-woofer externo Asus N series.
- Soporte MaxxAudio.

Procesador y memoria:

- Procesador Intel® Core™ i7 3630QM a 2.40 GHz.
- 8GB DDR3 1600 MHz SDRAM.

Almacenamiento y sistema operativo:

- 1TB 5400 RPM.
- Windows 8.

Este ordenador portátil fue utilizado para realizar 36 de las 143 grabaciones que el corpus *Entendámonos* presenta en este trabajo. Concretamente, con este equipo se realizaron las últimas 35 grabaciones de xi'iuy y 1 grabación de tének (la última).

Asimismo este equipo fue el que se utilizó para realizar toda la etapa de post-procesamiento de *Entendámonos* y también se utilizó para realizar las pruebas finales, mostradas en el Capítulo 5.

Bibliografía

- [1] Florian Schiel et al, “*The Production of Speech Corpora*”, BAS Infrastructures for Technical Speech Processing, Institut für Phonetik und Sprachverarbeitung, 2004.
- [2] Yeshwant Kumar Muthusamy, “*A Review of Research in Automatic Language Identification*”, Technical Report No. CS/E 92-009, Center for Spoken Language Understanding, Oregon Graduate Institute, USA, 1992.
- [3] Jiří Navrátil, “*Multilingual Speech Processing Chapter 8: Automatic Language Identification*”, Academic Press, pp. 233-272, 2006.
- [4] Miguel Ángel Martínez, Juan Enrique García y Patricia Fernández, “*Indígenas en zonas metropolitanas*”, Consejo Nacional de la Población, 2003.
- [5] Ruth Rodríguez, “*Hay 8 mil indígenas presos que carecen de un intérprete*”, Sección: Nación, El Universal México, Domingo 10 de Febrero del 2013.
- [6] Yeshwant Kumar Muthusamy, “*A Segmental Approach to Automatic Language Identification*”, Oregon Graduate Institute of Science and Technology, Oregon, United States, 1993.
- [7] D. Cimarusti and R. B. Ives, “*Development of an automatic identification system of spoken languages: Phase I*”, Proceedings 1982 IEEE International Conference on Acoustics, Speech and Signal Processing, Paris, France, May 1982.
- [8] José Manuel Vargas Martínez, “*Un Método para la Identificación Automática de Lenguas Basado en la Transformada Wavelet*”, División de Estudios de Postgrado e Investigación, Maestría en Ciencias en Ciencias de la Computación, Instituto Tecnológico de Ciudad Madero, 2008.
- [9] César Medina Trejo, “*Un Método para la Multiclasificación de Idiomas Usando la Transformada Wavelet*”, División de Estudios de Postgrado e Investigación, Maestría en Ciencias en Ciencias de la Computación, Instituto Tecnológico de Ciudad Madero, 2011.
- [10] Carlos Arturo Hernández Zepeda, “*Identificación automática de lenguas, utilizando la familia de wavelets Dbn, medidas estadísticas y minería de datos*”, Departamento de Ingeniería en Sistemas Computacionales, Instituto Tecnológico de Ciudad Madero, Mayo del 2011.
- [11] Ana Lilia Reyes Herrera, “*Un Método para la Identificación Automática de Lenguaje Hablado Basado en Características Suprasegmentales*”, Instituto Nacional de Astrofísica Óptica y Electrónica, Puebla, México, 2007.
- [12] Joel Sherzer, “*Archive of Indigenous Languages of Latin America*”, Department of Anthropology, University of Texas at Austin, 2002.
- [13] Kurtis Gurley et al, “*Applications of Wavelet Transforms in Earthquake, Wind and Ocean Engineering*”, Department of Civil Engineering and Geological Sciences, University of Notre Dame, 1998.

- [14] Ekaterina Timoshenko, “*Rhythm Information for Automated Spoken Language Identification*”, Lehrstuhl für Mensch-Maschine-Kommunikation, Technische Universität München, 2012.
- [15] Instituto Nacional de Estadística y Geografía (INEGI) , “*Hablantes de lengua indígena en México*”, *Cuéntame... de México - Población de México*, 2010, <http://cuentame.inegi.org.mx/poblacion/lindigena.aspx?tema=P>
- [16] Instituto Nacional de Lenguas Indígenas, “*Catálogo de las Lenguas Indígenas Nacionales, Variantes Lingüísticas de México con sus autodenominaciones y referencias geoestadísticas*”, INALI, México, 2009, ISBN 978-607-7538-08-0.
- [17] Juan Carlos Flores Paulín, “*Técnicas para el Reconocimiento de Voz en Palabras Aisladas de la Lengua Náhuatl*”, Centro de Investigación en Computación, Instituto Politécnico Nacional, 2009.
- [18] Caballero Morales, “*On the Development of Speech Resources for the Mixtec Language*”, The Scientific World Journal, Vol, 13, 2013.
- [19] T. Gumede, “*Development of a telephone-based speech-driven information service for the South African Government*”, Lwazi Application Section Report, Technical Report, Meraka Institute, July 2008.
- [20] C. Kuun, “*Development of a telephone-based speech-driven information service for the South African Government*”, Lwazi Project Final Report, Contract Report, Meraka Institute, November 2009.
- [21] Jean-Luc Rouas, “*Automatic prosodic variations modelling for language and dialect discrimination*”, IEEE Transactions on Audio, Speech and Language Processing, V15, N6, p1904-1911, 2007.
- [22] Alan Davies et al, “*The Handbook of Applied Linguistics*”, Blackwell Publishig, Publisher: John Wiley & Sons, ISBNs: 0470756756 or 9780470756751, 2006.
- [23] Michael Kipp et al, “*Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*”, Springer, ISBN 3642047920 or 9783642047923, 2009.
- [24] Jesús Eduardo Carrillo Ibarra et al, “*Automatic Language Identification Using the Daubechies Wavelet Transform in a MATLAB and WEKA Environment*”, 11th Mexican International Conference on Artificial Intelligence, V Hybrid Intelligent Systems Workshop, 2012.
- [25] Jorge Monforte et al, “*Narraciones Mayas*”, Primera edición, Instituto Nacional de las Lenguas Indígenas, México, ISBN: 978-607-7538-14-1, 2010.
- [26] Jonathan D. Amith et al, “*Ok nemi totlahtōl*”, Primera edición, Instituto Nacional de las Lenguas Indígenas, México, ISBN: 978-607-7538-01-1, 2009.
- [27] Michael Kipp et al, “*Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*”, 6th International Conference for Language Resources and Evaluation, ISBN: 3642047920 9783642047923, Springer, 2009.
- [28] D. Lee Fugal, “*Conceptual wavelets in digital signal processing: An in-depth, practical approach for the non-mathematician*”, University of California, Space & Signals Technical Pub., 2009, ISBN: 0982199457, 9780982199459.
- [29] I. Daubechies, “*Orthonormal bases of compactly supported wavelets*”, Commun. on Pure and Appl. Math., 41:909–996, November 1988.

Bibliografía

- [30] Ionut Danaila et al, “*An Introduction to Scientific Computing*”, 2006.
- [31] Mitchell Tom M., Machine Learning, McGraw-Hill Science/Engineering/Math, March 1, 1997.
- [32] Witten Ian H et al, “*Data Mining: Practical machine learning tools and techniques*”, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.