UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

"INTERFAZ DE VOZ PARA PERSONAS CON DISARTRIA"

T E S I S

PARA OBTENER EL TÍTULO DE:

INGENIERO EN COMPUTACIÓN

PRESENTA:

GLADYS BONILLA ENRÍQUEZ

DIRECTOR:

DR. SANTIAGO OMAR CABALLERO MORALES

HUAJUAPAN DE LEÓN, OAXACA, MAYO DE 2012

Dedico esta tesis a mis padres, Socorro y Enrique, por el apoyo que me han brindado de siempre y sus continuas motivaciones.

Agradecimientos

Agradezco al Dr. Santiago Omar Caballero Morales por su dirección y enseñanza, por compartir su conocimiento, experiencia, y corregirme cuando era necesario.

Por supuesto, también mi agradecimiento especial al Maestro José Antonio Moreno Espinosa, al Dr. Raúl Cruz Barbosa, y al Dr. Felipe de Jesús Trujillo Romero, por ser objetivos y puntuales en sus comentarios y observaciones para la culminación de este proyecto.

A los profesores y personal de la Universidad Tecnológica de la Mixteca que durante mi estancia siempre me apoyaron.

Agradezco a mis padres y hermanos, Enrique y Baruc, porque siempre han estado conmigo, y por haberme inspirado el sentido de superación. A mi esposo, te doy gracias por tu comprensión y palabras de ánimo cuando más lo necesitaba, te amo.

Resumen

La voz o el habla es una de las formas básicas de intercambiar información entre los seres humanos. Daños neuronales ocasionados por un infarto, una embolia o trauma cerebral, pueden afectar la voz del individuo, alterando su articulación, resonancia y respiración. A este cuadro patológico de la voz se le conoce como disartria.

Investigaciones realizadas en el extranjero han demostrado los beneficios de la tecnología computacional para la comunicación y asistencia de personas con disartria, especialmente de Sistemas de Reconocimiento Automático del Habla (SRAH). Sin embargo, no hay desarrollo de dicha tecnología que aborde el tema de la disartria en México, y particularmente, de personas cuya lengua materna sea el español mexicano.

El desarrollo de un sistema robusto de RAH para voz disártrica implica solucionar los siguientes problemas: (1) tasas variables de precisión en el reconocimiento de voz (25% - 95%) para usuarios con niveles alto-medio de disartria; (2) conforme el tamaño del vocabulario del sistema aumenta ($>$ 30 palabras), el nivel de precisión de reconocimiento disminuye; (3) desempeño poco significativo de técnicas de adaptación de usuario; (4) los síntomas asociados a la disartria dificultan la recopilación de muestras de voz (corpus) para un entrenamiento supervisado robusto del sistema.

La presente tesis describe el diseño y desarrollo de: (1) un SRAH para el español mexicano, y (2) una interfaz gráfica para la administración y configuración del SRAH. Esto para realizar las funciones de apoyo a la comunicación, terapia, y diagnóstico de mexicanos con el trastorno de disartria. La metodología de desarrollo abordó los problemas mencionados anteriormente, y como solución se propusieron los siguientes puntos: (1) la planeación del corpus de entrenamiento del SRAH puede repercutir en un entrenamiento robusto, incluso con recursos acústicos limitados (p.e., de un solo hablante); (2) adaptación de usuario, aplicada de manera dinámica sobre el SRAH, puede incrementar el nivel de precisión del sistema; (3) la manipulación en tiempo de ejecución de elementos estructurales del SRAH (no visibles en sistemas comerciales), puede mejorar su desempeño para usuarios con diferentes niveles de disartria, especialmente para vocabularios $>$ 100 palabras; (4) aunque se han probado metodologías en donde se crea un SRAH para cada usuario (haciéndolo dependiente del mismo), hay evidencias de mejor desempeño cuando el sistema es independiente del usuario.

Para la implementación de estos puntos, se desarrollaron los siguientes tres módulos

principales para la interfaz gráfica:

1. Creación y Adaptación del Reconocedor de Voz. Se programaron las siguientes funciones: captura de número de componentes gaussianos para los modelos acústicos del SRAH y entrenamiento supervisado del mismo; captura de datos de usuario nuevo, grabación y parametrización de muestras de voz para la adaptación "estática" del usuario, creación y actualización de registros personales, y ejecución de adaptación de usuario.

2. Reconocedor de Voz. Se programó la creación, integración, y ejecución automática de los componentes del SRAH (p.e., Modelos Acústicos, Modelo de Lenguaje, Diccionario, Algoritmo de Búsqueda). De igual manera, se programaron las siguientes funciones: captura y parametrización de voz en tiempo de ejecución para su reconocimiento; ajuste y actualización del Modelo de Lenguaje para reducir su perplejidad (y mejorar la precisión del sistema); captura e integración de nuevo vocabulario; administración de nuevas muestras de voz para adaptación "dinámica" del usuario; síntesis de voz.

3. Patrones de Confusión. Se diseñó y programó un alineador para generar una matriz de confusión fonética que presente los errores de pronunciación del usuario. Esta parte se agregó como un apoyo al terapeuta.

Para algunas funciones de la interfaz (p.e., construcción del SRAH) se utilizó la biblioteca de HTK toolkit, siendo la técnica de modelado acústico los Modelos Ocultos de Markov (Hidden Markov Models, HMM's). Con esta interfaz se construyeron dos SRAH: (1) Dependiente de Usuario (DU, entrenado con las muestras de un usuario con disartria leve-moderada), y (2) Independiente de Usuario (IU, entrenado con muestras de un usuario con voz normal). En pruebas con un vocabulario de 275 palabras, el sistema DU (usado por el mismo usuario con disartria) tuvo un desempeño del 75% de precisión para 50 frases espóntaneas. Sin embargo, el sistema IU adaptado de manera estática y dinámica para dos usuarios con disartria tuvo un desempeño de 95%. Estos resultados son comparables a la precisión del reconocimiento humano y mejor que el de otros sistemas computacionales (comerciales y de investigación) bajo condiciones de prueba similares (tamaño de vocabulario, número de usuarios de prueba, y nivel de disartria).

Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

En nuestro país existe investigación hacia nuevas tecnologías, como lo es en Sistemas de Reconocimiento Automático del Habla (SRAH) y en Procesamiento de Lenguaje Natural (PLN). Uno de ellos es el proyecto DIME (Diálogos Inteligentes Multimodales en Español), desarrollado en el Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas (IIMAS) de la UNAM. Este proyecto desarrolló un robot (Golem) y su SRAH para permitir la interacción por medio de lenguage natural entre el robot y usuarios humanos [45].

En la Universidad Tecnológica de la Mixteca hay antecedentes de proyectos en SRAH. En [40], se llevó a cabo el entrenamiento dinámico de un reconocedor de voz con los corpora *DIMEx100 niños*, *DIMEx100 adultos*, y *Golem-Universum* del IIMAS para evaluar su desempeño con diferentes secciones de estos corpora. En tanto, en [14] se desarrolló un parser semántico para el módulo Golem-Universum.

Dentro de este campo de las Ciencias Computacionales, se busca desarrollar una aplicación directa para personas con discapacidades, en especial, del habla. De acuerdo al INEGI, al año 2010, aproximadamente el 5.1% de la población total mexicana tenía alguna discapacidad. De este porcentaje, el 58.3% correspondía a una discapacidad motora, y el 8.3% a una discapacidad para hablar o comunicarse[1]. Particularmente en el estado de Oaxaca, en donde hay alrededor de 3.8 millones de habitantes, aproximadamente 343,000 tienen alguna de estas discapacidades[2]. Algo importante es que una persona puede tener más de una discapacidad. Por ejemplo: los sordomudos tienen una

[1] http://cuentame.inegi.org.mx/poblacion/discapacidad.aspx?tema=P

[2] http://cuentame.inegi.org.mx/monografias/informacion/oax/poblacion/default.aspx

limitación auditiva y otra de lenguaje, o quienes sufren de parálisis cerebral presentan problemas motores y de lenguaje.

Dentro de las discapacidades motoras del habla y de comunicación se considera a la disartria, que se puede definir como el trastorno de la expresión verbal causado por una alteración en el control muscular de los mecanismos del habla, siendo este un problema del habla y no un problema del lenguaje. En el Centro SNDIF[3] de la H. Huajuapan de León, Oaxaca, se obtuvo asesoría de la Dra. María Luisa Gutierrez (Coordinadora) y de los terapeutas, Rocio Bazan Pacheco (Terapia del Lenguaje), y Diana Pérez Hernández (Terapia Ocupacional), para conocer acerca de las patologías de lenguaje y motoras en general de las personas con disartria. Este personal reconoció la utilidad práctica de contar con un sistema computacional que ayudara a los pacientes a comunicarse por medio de la voz. De igual manera, que a los terapeutas les ofreciera herramientas para diagnosticar de mejor manera la disartria de sus pacientes, y así, planear de mejor manera sus actividades de rehabilitación.

Al tener dicha información se realizó investigación en el campo de aplicaciones de RAH, encontrándose proyectos en otros países enfocados al desarrollo de sistemas para mejorar la comunicación de personas con disartria. Esto llevó a identificar los siguientes problemas relacionados con el desarrollo de un SRAH para voz disártrica [15, 51, 24, 34]:

- tasas variables de precisión en el reconocimiento de voz (25% - 95%) para usuarios con niveles alto-medio de disartria;
- el rango de anormalidades en la voz disártrica es muy amplio, variando entre personas afectadas;
- conforme el tamaño del vocabulario del sistema aumenta ($>$ 30 palabras), el nivel de precisión de reconocimiento disminuye;
- desempeño poco significativo de técnicas de adaptación de usuario;
- los síntomas asociados a la disartria dificultan la recopilación de muestras de voz (corpus) para un entrenamiento supervisado robusto del sistema;

[3]Sistema Nacional para el Desarrollo Integral de la Familia, Blv. Tierra del Sol Esq. Calle Pedro Sepulveda. Agencia del Carmen, Huajuapan de León, Oax.

- no hay un corpus de voz disártrica en español mexicano para la realización de análisis o modelado acústico para la construcción de un SRAH;
- no hay proyectos en RAH para el idioma mexicano similares que sirvan como base de comparación, la mayoría están desarrollados para el idioma Inglés.

Es por esto que, para el desarrollo del sistema propuesto, que consiste de una interfaz integrada por los elementos mostrados en la Figura 1.1, se comenzó desde el nivel básico (p.e., hacer un corpus de voz) hasta llegar a un nivel avanzado de programación de la interfaz. Basado en el conocimiento y experiencia de los terapeutas, se definió un perfil de usuario para la interfaz. De esta manera, la interfaz se delimitó, enfocándose hacia las personas que presentan disartria causada por enfermedades neuronales no degenerativas, que mantengan un coeficiente mental coherente y de un nivel bajo-medio de severidad.

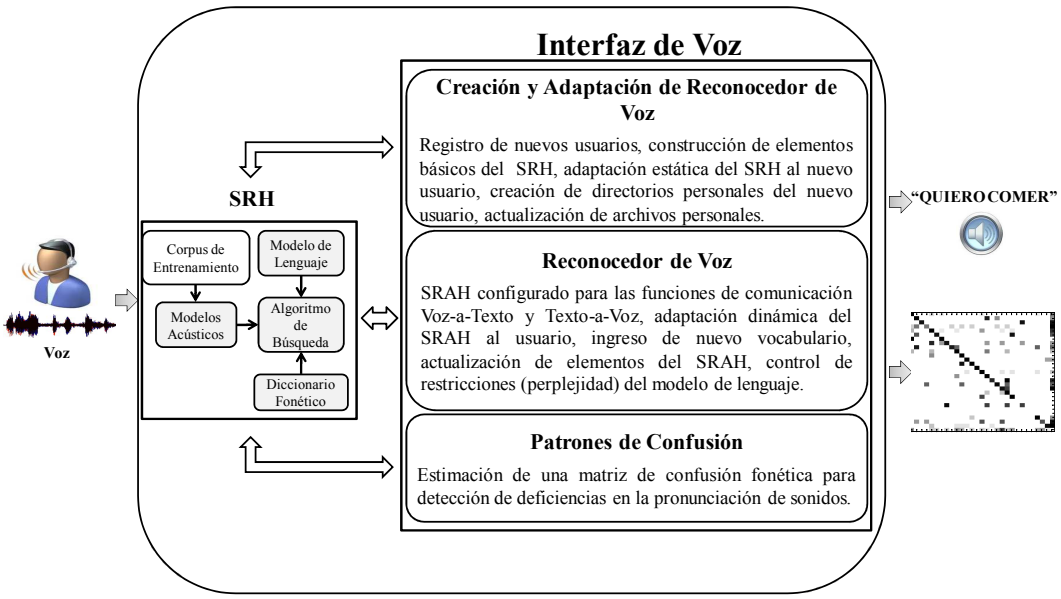


Figura 1.1: Elementos de la Interfaz de Voz Propuesta.

De igual manera, con el apoyo de los terapeutas se encontraron candidatos con diagnóstico de disartria que cubrían el perfil requerido. Es importante mencionar que en este proyecto se trabajó con personas en condiciones reales, a diferencia de otros proyectos que utilizan corpus para la evaluación del sistema.

Una vez lograda la selección de usuarios, se inició con la obtención de las bases

teóricas y prácticas para el desarrollo de la interfaz propuesta (un SRAH y su respectiva interfaz gráfica, ver Figura 1.1). Para esto, se uso la técnica de los Modelos Ocultos de Markov (Hidden MarKov Models, HMM's). En la actualidad, la mayoría de los SRAH se basan en esta técnica estocástica de modelado [27, 28, 49, 50].

Al construir un SRAH, uno de los temas de importancia es el procesamiento de información acústica, el cual requiere algoritmos complejos. Para esto se han desarrollado herramientas que agilizan este proceso, como la biblioteca del HTK Toolkit [61]. Para propósitos de este proyecto, HTK tiene una estructura autosuficiente, se puede utilizar en varias plataformas (DOS, Linux, etc.), y permite al investigador el trabajar con HMM's desde el nivel básico de diseño hasta el de evaluación. Muestra de ello son los diversos trabajos de investigación en SRAH que han utilizado esta biblioteca [19, 24, 25, 37].

Sin embargo, al ser HTK un conjunto de blibliotecas cuyo uso se hace por medio de línea de comandos [6, 52, 59], el manejo de archivos y comandos para usar HTK resulta bastante hostil para los usuarios principiantes [59]. Por lo tanto es necesaria la programación de un sistema que configure y administre en tiempo de ejecución esta herramienta para realizar aplicaciones de RAH requeridas [59]. Esto es, una interfaz gráfica que permita la interacción del usuario con el sistema de una manera útil y flexible sin necesidad de conocimientos técnicos de computación o de RAH, habilitándolo para llevar a cabo las tareas deseadas (p.e., comunicación y diagnóstico, ver Figura 1.1).

Dentro del trabajo inicial, una vez que se obtuvieron muestras de voz disártrica, se construyó un SRAH de manera manual usando HTK. Este SRAH, al ser usado por el mismo usuario con disartria que proporcionó las muestras de voz, se definió como Dependiente del Usuario (DU). Este reconocedor sólo consistió de archivos y directorios con los elementos mostrados en la Figura 1.1, sin ninguna interfaz y sin realizar RAH en tiempo de ejecución (la prueba se hizo con muestras grabadas de voz). Haciendo uso de lenguaje de comandos de HTK y edición manual de texto, se realizó parametrización de archivos de voz, entrenamiento supervisado de modelos acústicos (HMM's), creación de diccionario y modelo de lenguaje, y evaluación de desempeño del SRAH. El tiempo de construcción para este sistema fue de aproximadamente tres semanas, obteniendo un porcentaje de reconocimiento del 75%.

Debido a este bajo porcentaje, se tomó la decisión de utilizar una metodología de diseño diferente, la de un sistema Independiente de Usuario (IU). En este caso, el sis-

tema RAH se entrenó con muestras de voz diferentes a las del usuario con disartria. Al programar un sistema de administración de los diferentes elementos del SRAH, las bibliotecas correspondientes de HTK (incluyendo adaptación de usuario), e integrar la configuración de los mismos en tiempo de ejecución (facilitado mediante la interfaz gráfica), se logró incrementar el RAH hasta niveles comparables al del reconocimiento de voz humano: 95%.

Como se muestra en la Figura 1.1, la interfaz de voz desarrollada usa el SRAH para realizar las siguientes funciones, cada una de ellas integradas en módulos especiales:

- la adaptación continua de voz para el español mexicano y la variación del modelado acústico del SRAH;
- reconocimiento de frases continuas, con la opción a agregar en línea nuevo vocabulario y actualizar, de manera automática, todos los componentes del SRAH;
- obtener visualmente una matriz de confusión fonética para identificar deficiencias en la pronunciación de fonemas del usuario con disartria.

Las características de la interfaz tienen su base en aspectos fundamentales de RAH. El añadir vocabulario en tiempo de ejecución al sistema, a parte de ofrecer flexibilidad al usuario y al terapeuta, tiene una razón más técnica. Se encontró que para usuarios con disartria, SRAHs funcionan mejor cuando la perplejidad del componente de Modelo de Lenguaje es menor [57]. Esto está relacionado con el conocimiento previo que tiene el sistema del vocabulario de uso, y la interfaz permite el control de este factor. De igual manera, un mejor modelado de la voz se puede obtener variando los componentes de los HMM's del sistema, conocidos como gaussianos, lo cual también se puede modificar con la interfaz. Por otro lado, técnicas de adaptación usualmente tienen bajo desempeño cuando se utilizan para adaptar sistemas para voz normal (comerciales o de investigación) a usuarios con disartria [18]. En este trabajo, se encontró que, mediante la aplicación progresiva (dinámica) de estas técnicas, se pueden obtener mejoras significativas para voz disártrica. Finalmente, la matriz de confusión fonética mostró información de deficiencias consistentes con las detectadas por los terapeutas. Esto se obtuvo mediante un alineador de cadenas genéticas [5], el cual fue adaptado para fonemas del español mexicano. Para este alineador se consideraron similitudes acústicas para la clasificación de errores de pronunciación (ver Anexo B).

Es así que se desarrolló esta interfaz como un medio tecnológico de apoyo para la comunicación de personas con disartria, u otra posible deficiencia en el habla. En la práctica, pudo ser utilizada por la persona discapacitada, sus familares y terapeutas. Especialmente estos últimos, pudieron añadir palabras para realizar ejercicios de pronunciación (terapia de lenguaje) con la persona afectada sin requerir conocimiento técnico especializado. De igual manera, escuchar al usuario con una voz más entendible, y apoyarse en la información de confusiones fonéticas para corroborar sus diagnósticos. En esta tesis se muestran los detalles de diseño e implementación (programación) de cada módulo de la interfaz de voz, al igual que de los resultados obtenidos.

1.1 Motivación

El sector de la población con discapacidades se encuentra muchas veces aislada de la sociedad. Y es común la discriminación, independientemente si la discapacidad es del tipo auditiva, del habla, física, visual, o alguna combinación de las anteriores.

El enfoque de esta tesis es sobre la discapacidad del habla, específicamente hacia la disartria. Sin embargo, aunque en otros países se han desarrollado herramientas tecnológicas para apoyar a personas con esta discapacidad, en México los trabajos relacionados a este trastorno son limitados.

Por lo tanto se considera el desarrollo de aplicaciones que permitan ser utilizados como herramientas para la disartria en el español mexicano. Esta herramienta no solo pretende apoyar a la persona disártrica para interactuar de mejor manera con sus familiares, sino también con su entorno.

En específico, la herramienta de apoyo es una interfaz de voz que consiste de: (1) un modulo de adaptación a usuario, (2) un SRAH y sintetizador de voz para comunicación, y (3) un estimador de matriz de confusión fonética para apoyo de diagnóstico de disfunciones en la voz (ver Figura 1.1). Para ello, la interfaz además de cubrir la parte de interacción con el usuario, cubre la parte de administración y configuración de bibliotecas de una herramienta, HTK [61], para llevar a cabo las tareas deseadas.

HTK [6, 52, 59, 61] es un conjunto de bibliotecas especializadas para reconocimiento de patrones usando HMM's. Esta herramienta es ampliamente usada en el desarrollo de sistemas de reconocimiento de voz [19, 24, 25, 37] e investigación. Para un sistema que

vaya a ser usado por un usuario sin conocimientos de computación, o de reconocimiento de voz, HTK tiene sus desventajas [6, 59]:

- las llamadas a bibliotecas y configuración de parámetros que requiere HTK es compleja al realizarse mediante líneas de comandos;
- el manejo de archivos resulta bastante hostil para los usuarios principiantes de HTK.

Por lo tanto, el desarrollo de la interfaz gráfica para utilizar la biblioteca HTK no fue una tarea fácil. Esto porque, aunque HTK realiza cálculos complejos para la decodificación de voz a texto, requiere de:

- creación y configuración de archivos funcionales (p.e., modelo de lenguaje, modelos acústicos, diccionario fonético, etc.) para su operación;
- la coordinación de la ejecución de la biblioteca y los archivos generados para llevar a cabo funciones de re-estimación, entrenamiento y adaptación de usuario.

Estos requerimientos deben cubrirse en tiempo de ejecución, con retroalimentación del usuario, para los propósitos del sistema. El desarrollo de esta interfaz involucró diferentes ámbitos además de la programación de la interfaz: teoría matemática de reconocimiento de voz, probabilidad, conocimiento de lingüística del español mexicano, e Interacción Humano-Computadora.

1.2 Objetivos

1.2.1 Objetivo General

Desarrollo de una Interfaz de Reconocimiento de Voz para la comunicación, para apoyo a diagnóstico, y/o práctica de personas con problemas de disartria de un nivel medio-bajo para el español mexicano.

1.2.2 Objetivos Particulares

Se describen los siguientes objetivos realizados para el desarrollo de la interfaz de voz.

- Evaluar el desempeño de dos metodologías de diseño de un SRAH para voz disártrica: (1) dependiente del usuario, DU (construído con la voz del usuario objetivo con disartria), (2) independiente del usuario, IU (construído con la voz de un usuario normal, pero adaptado para ser usado por el usuario objetivo con disartria).

- **Desarrollo de Interfaz Gráfica de Adaptación de Usuario**:

  - Programación de un submódulo para crear un prototipo de HMM para cada fonema en el español mexicano con un número X de componentes gaussianos, y coordinar HTK para hacer el entrenamiento supervisado del conjunto total de HMM's. El número de componentes gaussianos se considera variable (1-8) y definible por el usuario.
  - Programación de un submódulo de captura y almacenamiento de voz para propósitos de adaptación estática y dinámica. De igual manera, se programó la creación de directorios y archivos personalizados de cada usuario para el uso del reconocedor y adaptación.
  - Diseño de vocabulario para adaptación balanceada estática de usuario.
  - Programación de la rutina de las bibliotecas HCopy y HERest de HTK para parametrización de voz y adaptación de usuario.

- **Desarrollo de Interfaz Gráfica de Reconocimiento de Voz**:

  - Programación de submódulo para controlar el nivel de influencia del modelo de lenguaje sobre la ejecución del reconocedor de voz (ejecutada por la biblioteca HVite de HTK). Esto es, control de nivel de perplejidad del modelo de lenguaje para disminución de la tasa de error del reconocedor. Esto se lleva a cabo mediante un programa para controlar el factor de gramática (Modelo de lenguaje) y el ingreso y actualización de vocabulario del modelo del lenguaje del sistema.
  - Programación de un submódulo para construir automáticamente el diccionario fonético del reconocedor de voz y el modelo de lenguaje dado un vocabulario.

  - Programación de un submódulo para añadir nuevas palabras o frases al vocabulario de la interfaz, y actualizar los componentes del diccionario fonético y modelo de lenguaje del sistema. De igual manera, actualizar el listado de palabras / frases disponibles para ser usadas por el sistema.
  - Programación de submódulo para enlazar un sintetizador de voz para leer el texto decodificado por el reconocedor de voz.
  - Programación del submódulo de administración de archivos y bibliotecas para realizar la adaptación dinámica del usuario con nuevo vocabulario.

- **Desarrollo de Interfaz Gráfica de Patrones de Confusión Fonética**:

  - Diseño y programación de un submódulo alineador de fonemas del español mexicano para estimar patrones de confusión fonética. El sistema muestra los fonemas que el usuario pronuncia y que reconoce de acuerdo a los modelos acústicos base del mismo.

1.3 Estructura de la Tesis

A continuación se presentan los detalles de los capítulos del documento de tesis.

- **Capítulo 2: Marco Teécnico**: Presentación de información técnica concerniente a RAH y software para el diseño de los elementos funcionales de un SRAH. Este fondo técnico es importante para el desarrollo de la interfaz propuesta.

- **Capítulo 3: La Disartria y las Tecnologías de Asistencia**: Presentación de información relevante a la disartria, detalles acerca de sistemas similares al propuesto.

- **Capítulo 4: Desarrollo de la Interfaz de Voz**: Descripción de los pasos seguidos para el desarrollo de la interfaz, esto es:

  - definición de variables de control;
  - descripción del proceso de creación del corpus de entrenamiento del SRAH (selección del corpus textual);

  - descripción del desarrollo e integración de cada sub-módulos de la interfaz de voz para realizar las tareas de adaptación estática y dinámica de usuario, reconocimiento (comunicación), y estimación de patrones fonéticos para terapia.

- **Capítulo 5: Presentación de Resultados**: Presentación de las pruebas en tiempo de ejecución y fuera de condiciones de laboratorio de la interfaz propuesta con los usuarios finales. Se describen los siguientes puntos:

  - búsqueda y selección de usuarios con disartria;
  - selección del vocabulario de uso para la evaluación del sistema;

  De igual manera se presentan comparaciones y una discusión acerca de las aportaciones del presente proyecto.

- **Capítulo 6: Conclusiones y Trabajo a Futuro**: Discusión y comentarios finales acerca de los logros obtenidos y propuestas de mejoras para el sistema y el proyecto en general.

1.4 Publicaciones

El trabajo en esta tesis fue presentado en las siguientes publicaciones:

- Bonilla-Enríquez, G., Caballero-Morales, S.O., "Reconocimiento de Voz para Comunicación y Diagnóstico de Personas con Disartria en México", VII Semana Nacional de Ingeniería Electrónica (SENIE 11), p. 431-440, Tapachula, Chiapas, 28 Octubre de 2011 (ISBN 968-607-477-588-4).

- Bonilla-Enríquez, G., Caballero-Morales, S.O., "Communication Interface for Mexican Spanish Dysarthric Speakers", Mexican International Conference on Computer Science, ENC 2012. *Acta Universitaria*, Vol. 22 (NE-1), p. 98-105, Salamanca, Guanajuato, 28 de Marzo de 2012 (ISSN: 0188-6266).

Capítulo 2

Marco Teórico

En este capítulo se presentará el marco teórico relacionado con el desarrollo de Sistemas de Reconocimiento Automático del Habla (SRAHs), explicándose procesos como la creación de corpus de entrenamiento, modelado acústico, estimación de modelos de lenguaje, evaluación del sistema, e implementación del RAH. Esta información se complementa con la presentación de la biblioteca HTK, explicando cómo se puede utilizar esta herramienta para la realización de estos procesos.

2.1 Sistema de Reconocimiento Automático del Habla (SRAH)

De manera general un SRAH se puede catalogar como:

- Dependiente del Usuario (DU): Aplicación sólo para un usuario. Se construye tomando en cuenta sus características acústicas particulares, es un sistema personalizado.

- Independiente del Usuario (IU): Aplicación para más de un usuario. Se construye tomando en cuenta las características de muchos usuarios y después se personalizan mediante técnicas de adaptación de usuario.

En la Figura 2.1 se presentan los componentes base de un SRAH, los cuales son independientes del tipo de sistema. El proceso de reconocimiento se considera estocástico

(no determinístico) y se basa en el método de Bayes para estimar la secuencia de palabras más probable $\hat{W}$ (de entre todas las posibles secuencias permisibles por un **Modelo de Lenguaje** L) dada una señal acústica de entrada O. $\hat{W}$ es estimada como:

$$\hat{W} = argmax_{W \in L} \frac{P(O|W)P(W)}{P(O)} \tag{2.1}$$

en donde $P(W)$ es la Probabilidad *A-Priori*, obtenida de un **Modelo de Lenguaje**, y $P(O|W)$ es la Probabilidad de Observación, obtenida de los **Modelos Acústicos**. $P(W)$ y $P(O|W)$ son usualmente estimados por medio de N-gramas y Modelos Ocultos de Markov (HMM's)[27]. En las siguientes secciones se describen cada uno de estos elementos.

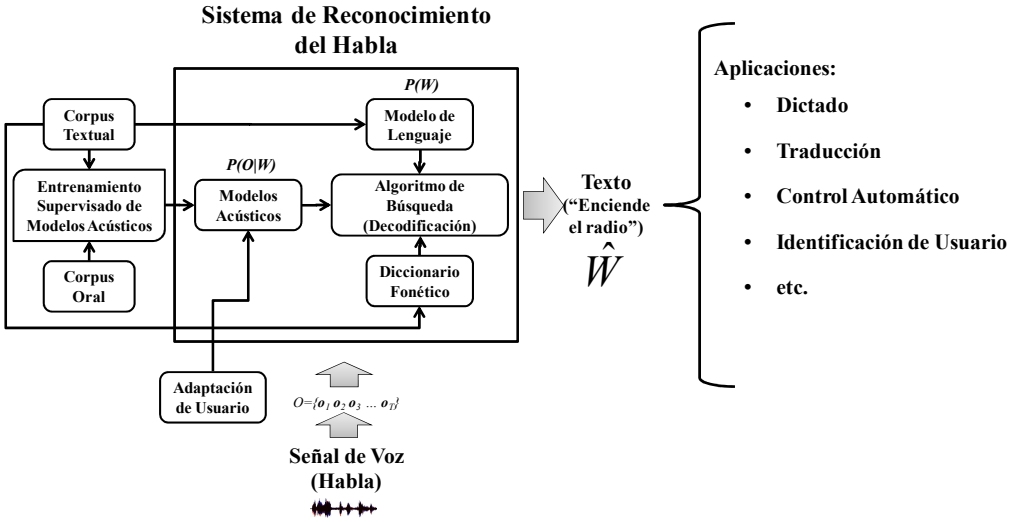


Figura 2.1: Elementos fundamentales de un Sistema de Reconocimiento Automático del Habla (SRAH).

2.1.1 Corpus Textual y Oral, Diccionario Fonético

Como elemento previo importante es el Corpora del Habla (Corpus en singular) para la creación y evaluación de SRHs. Un corpus del habla es una base de datos, una colección de archivos de voz (audio) y transcripciones textuales de los mismos en un formato que puede ser usado para la creación y refinación de modelos acústicos para SRH's. Dentro de estos corpora, se pueden diferenciar dos tipos:

- Textual: Consiste de una colección de textos representativos de un lenguaje. Estos se pueden obtener de extractos de libros, reportes de noticias, etc.

- Oral: Consiste en una colección de archivos de audio (voz) los cuales se pueden obtener de la siguiente manera:

  - de la lectura de Texto Representativo (por ejemplo, de un Corpus Textual).
  - de pláticas espontáneas (por ejemplo, de narraciones acerca de sucesos personales, diálogos entre personas, etc.)

En la Figura 2.2 se muestran los pasos a seguir para la construcción de un corpus para desarrollo de un SRAH. Inicialmente es necesario definir un contexto de uso (p.e., palabras de asistencia en el hogar, frases de control, conversaciones sobre política, etc.). Esto es importante para definir el texto representativo (corpus textual) que se usará como estímulo para obtener las muestras de voz correspondientes (corpus oral).

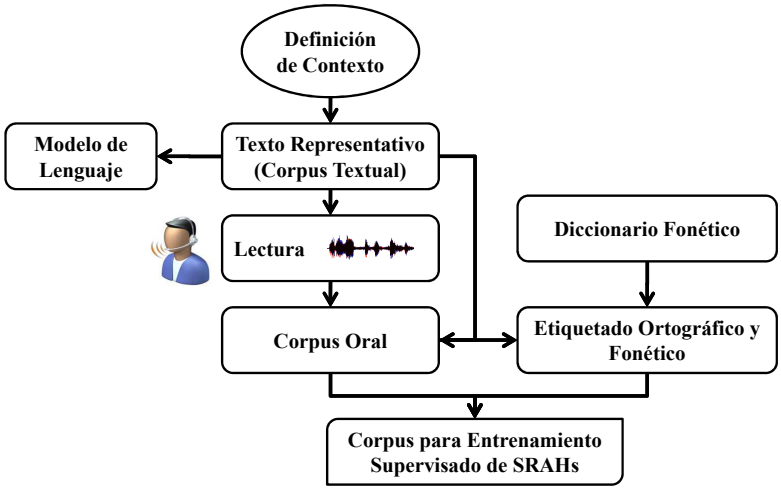


Figura 2.2: Pasos para la construcción de un corpus de voz.

El corpus textual y el oral deben tener una relación en el tiempo para que puedan ser utilizados para el desarrollo de SRAH. Esto es, que se identifique claramente los segmentos de audio que corresponden a una palabra o sonido en específico. Esto es vital para el entrenamiento supervisado y/o adaptación de un SRAH.

Al proceso de relacionar ambos corpus en el tiempo se le conoce como transcripción o etiquetado [61]. En la Figura 2.3 se muestra un ejemplo de etiquetado ortográfico y fonético de una muestra de voz.

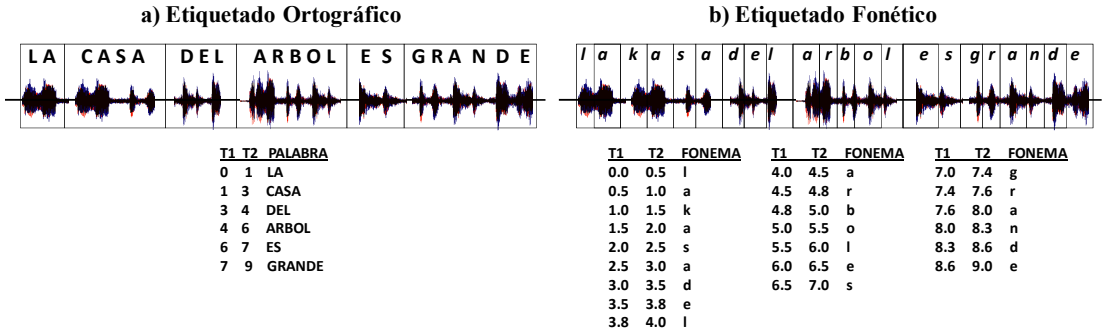


Figura 2.3: Ejemplo de etiquetado ortográfico (palabras) y fonético (fonemas).

Como se muestra en la Figura 2.3, el etiquetado ortográfico consiste en identificar los segmentos de la señal de voz que corresponden a palabras. En cambio, el etiquetado fonético consiste en identificar los sub-segmentos que forman una palabra, los cuales se definen como fonemas. Como ejemplo se tiene la palabra ARBOL, la cual se forma de la secuencia de fonemas $/a/$ $/r/$ $/b/$ $/o/$ $/l/$.

Actualmente los SRAH se modelan a nivel fonema, de tal manera que con un conjunto finito de sonidos se pueden formar una amplia variedad de palabras. El elemento que define las secuencias que forman cada palabra se conoce como **Diccionario Fonético**. Por lo tanto, antes de comenzar con el etiquetado fonético es necesario contar con este recurso. Como ejemplos de estos se tienen el CMU Pronouncing Dictionary[1] y BEEP[2] para el inglés americano y británico respectivamente, con definiciones para aproximadamente 250,000 palabras. Estos diccionarios se han utilizado para el etiquetado de corpus como el WSJ y WSJCAM0. Una vez terminado el corpus se procede al entrenamiento de los componentes del SRAH.

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[2] http://mi.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html

2.1.2 Modelo de Lenguaje

Un **Modelo de Lenguaje (ML)**, o **Gramática**, representa un conjunto de reglas o probabilidades que determinan las secuencias de palabras permisibles en un lenguaje. Esto incrementa el desempeño del SRAH ya que el ML guía el proceso de reconocimiento mediante la restricción de secuencias reconocidas a secuencias que son estadísticamente más probables que otras. Por ejemplo, es común (y correcto) decir *"la casa de mi madre es azúl"*, pero no es correcto decir *"la casa azúl es mi madre de"*. En este caso un ML le asignaría una mayor probabilidad a la primera secuencia que a la segunda.

Las probabilidades y reglas gramaticales de un ML generalmente se estiman a partir de **Corpora Textual**. Un ML frecuentemente utilizado es el conocido como **bigrama** el cual denota un N-grama con contexto $N = 2$, esto es, usa las $N - 1 = 1$ palabras previas para predecir la siguiente [27]. Del ejemplo anterior, la palabra *"es"* ayuda a predecir que la siguiente palabra es *"azúl"* y no *"mi"*. Otro ML utilizado es el **trigrama** el cual está denotado por $N = 3$, es decir, usa las $N - 1 = 2$ palabras previas para predecir la siguiente. La mayoría de los SRH's comerciales usan trigramas, los cuales son estimados usando millones de palabras de textos representativos (Corpora Textual).

Matemáticamente, una secuencia (o frase) de m palabras puede estimarse como el producto de las probabilidades condicionales de cada N-grama de la siguiente manera:

- N=1, **Unigrama**: $Pr(w_1, .., w_m) = Pr(w_1)Pr(w_2)...Pr(w_m)$.
- N=2, **Bigrama**: $Pr(w_1, .., w_m) = Pr(w_1)Pr(w_2|w_1)...Pr(w_m|w_{m-1})$.
- N=3, **Trigram**: $Pr(w_1, .., w_m) = Pr(w_1)Pr(w_2|w_1)Pr(w_3|w_1, w_2)...Pr(w_m|w_{m-2}, w_{m-1})$.

2.1.3 Modelado Acústico

El modelado acústico consiste en el proceso de establecer representaciones estadísticas para las características espectrales de la señal de voz. Para esto, los Modelos Ocultos de Markov (Hidden Markov Models, HMM's) [27, 50] son los más utilizados, aunque se han utilizado también Redes Neuronales Artificiales (Artificial Neural Networks, ANNs) [26].

Un HMM es un modelo estocástico en el cual el sistema modelado se asume que es un proceso de Markov, en donde los estados no son directamente visibles, pero variables

influenciadas por los estados (las observaciones, por ejemplo, los vectores de atributos espectrales de las señales acústicas $\mathbf{o}_t$) son visibles [6, 27, 49, 59, 61]. Un ejemplo de un HMM se muestra en la Figura 2.4.

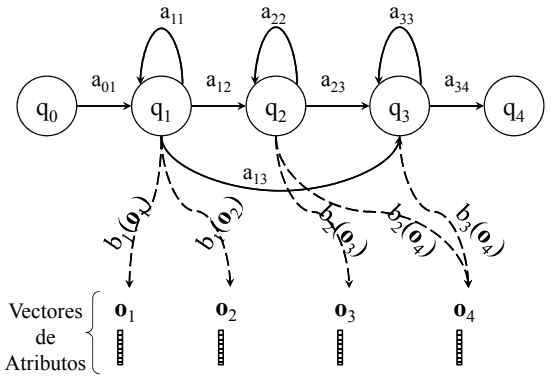


Figura 2.4: Estructura estándar izquierda-a-derecha de 3-estados de un HMM.

Generalmente la notación $\lambda = (A, B, \pi)$ es usada para definir al conjunto de parámetros de un HMM [27], en donde:

- $Q = \{q_0,\ q_1,\ ...,\ q_N\}$, un conjunto de estados, en donde q_0 y q_N son estados no-emisores (no asociados con observaciones). Cada estado tiene asociado una función de probabilidad que modela la emisión/generación de ciertas observaciones (véase $B = \{b_i(\mathbf{o}_t)\}$).

- $A = \{a_{01},\ a_{02},\ ...,\ a_{NN}\}$, una matriz de probabilidades de transición A, en donde cada a_{ij} representa la probabilidad de moverse del estado i al estado j. $\sum_{j=1}^{N} a_{ij} = 1\ \forall i$.

- $B = \{b_i(\mathbf{o}_t)\}$, un conjunto de *Probabilidades de Observación*. Cada término representa la probabilidad de que un vector observado $\mathbf{o}_t$ sea generado o emitido por un estado i. El modelado de $b_j(\boldsymbol{o}_t)$ se hace por medio de una **Mixtura (o Mezcla) de Gaussianas** [27, 61]:

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{K} C_{jk} N(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \tag{2.2}$$

en donde K denota el número de componentes gaussianos, C_{jk} es el peso para la k-ésima mixtura que satisface $\sum_{k=1}^{K} C_{jk} = 1$, y $N(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})$ denota a una sola función de densidad gaussiana con vector de media $\boldsymbol{\mu}_{jk}$ y matriz de covarianza $\boldsymbol{\Sigma}_{jk}$ para el estado j. Esta gaussiana puede ser expresada como:

$$N(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) = \frac{1}{\sqrt{(2\pi)^n \left|\boldsymbol{\Sigma}_{jk}\right|}} e^{-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_{jk})' \boldsymbol{\Sigma}_{jk}^{-1} (\mathbf{o}_t - \boldsymbol{\mu}_{jk})} \tag{2.3}$$

en donde n es la dimensionalidad de $\mathbf{o}_t$, y (') denota la transpuesta del vector.

- $\pi = \{\pi_i\}$, una distribución inicial de estados, en donde $\pi_i = Pr(q_0 = i),\ 1 \leq i \leq N$, y $\sum_{i=1}^{N} \pi_i = 1$.

Para el reconocimiento de amplio vocabulario, HMM's de izquierda-a-derecha (ver Figura 2.4) son generalmente utilizados para modelar sub-unidades de palabras (fonemas), que pueden ser concatenadas para formar palabras [61]. Un Léxico o Diccionario de pronunciaciones entonces es usado para definir las secuencias de fonemas que pueden formar una palabra. Para este caso, un HMM se entrenaría para cada fonema, y mientras que las secuencias legales de fonemas están determinadas por el diccionario, las secuencias permisibles de palabras están restringidas por un N-grama (Modelo de Lenguaje).

Tres problemas concernientes a los HMM's son de interés para los SRH's:

- **Problema de Decodificación (Búsqueda)**. Dada la secuencia observada $\boldsymbol{O}$ y el modelo $\lambda = (A, B, \pi)$, estimar la secuencia de estados Q que mejor describa las observaciones. Note que éste es el problema de reconocimiento del habla, para el cual el algoritmo de **Viterbi** es ampliamente usado.

- **Problema de Evaluación.** Dada la secuencia observada $\boldsymbol{O}$ y el modelo (λ), estimar de manera eficiente la probabilidad de observar dicha secuencia dado el modelo ($Pr(\boldsymbol{O}|\lambda)$).

- **Problema de Aprendizaje.** Dada una secuencia de observaciones $\boldsymbol{O}$ de un conjunto de entrenamiento, estimar/ajustar las probabilidades de transición (A) y emisión (B) de un HMM para describir con más precisión dicha información. Esto es, maximizar $Pr(\boldsymbol{O}|\lambda)$.

Los problemas de **Evaluación** y **Aprendizaje** se trabajan mediante algoritmos estándar como los de **Baum-Welch** y **Viterbi**. Para el caso de **Aprendizaje Supervisado** de

HMM's es necesario utilizar **Corpora Oral** etiquetado a niveles ortográfico (palabra) y fonético.

2.1.4 Algoritmo de Búsqueda

Un algoritmo eficiente es necesario para buscar a través de todas las secuencias en L aquellas que sean más probables que correspondan a las observaciones O. El algoritmo de **Viterbi** es ampliamente utilizado para la tarea de encontrar la secuencia más probable de estados $Q^* = \{q_1, q_2, ..., q_n\}$ que pudiera haber generado una secuencia de observaciones $O = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_t\}$ dado un modelo λ.

Como se muestra en la Figura 2.5, este algoritmo puede visualizarse como encontrar el mejor camino a través de una matriz, o "Malla de Viterbi" (Viterbi trellis), en donde el eje vertical representa los estados de un HMM y el eje horizontal representa los segmentos de voz (p.e., los vectores de características espectrales). Cada celda o casilla de la malla de Viterbi, $v_t(j)$, almacena la probabilidad de que el HMM esté en el estado j después de ver las primeras t observaciones y pasando a través de la secuencia de estados $q_1, ..., q_{t-1}$ más probables de acuerdo al modelo λ. De esta manera, la celda contiene la probabilidad acumulada del "mejor" (más probable) camino para las primeras t observaciones y que termina en el estado j del HMM.

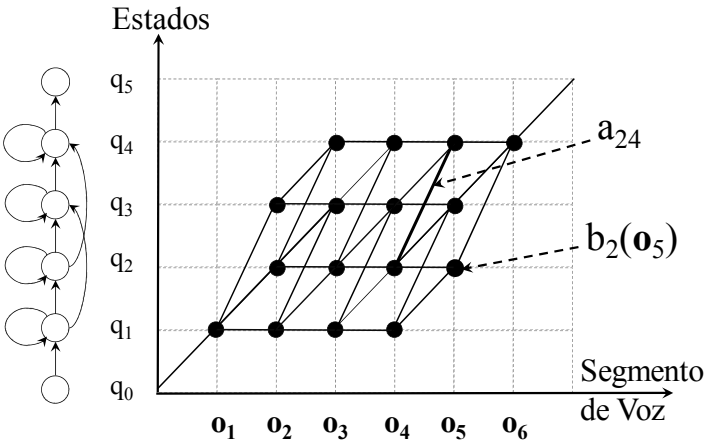


Figura 2.5: El algoritmo de Viterbi para reconocimiento de palabras.

Formalmente, el camino más probable de todas las posibles secuencias de estados

de longitud $t-1$ puede ser expresado como:

$$v_t(j) = argmax_{q_1,q_2,...,q_{t-1}} Pr(q_1 q_2 ... q_{t-1}, q_t = j, \boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_t | \lambda) \tag{2.4}$$

El cálculo de la Ec. 2.4 puede ser optimizado mediante la siguiente recursión para un determinado estado q_j en el tiempo t:

$$v_t(j) = argmax_{1 \leq i \leq N-1} \{v_{t-1}(i) a_{ij} b_j(\boldsymbol{o}_t)\} \tag{2.5}$$

en donde $v_1(1) = 1$ y $v_1(j) = a_{1j} b_j(\boldsymbol{o}_1)$, $1 < j < N$. $v_t(j)$ representa la máxima probabilidad de las observaciones $\boldsymbol{o}_1$ a $\boldsymbol{o}_t$ de estar en el estado j en el tiempo t.

Si un modelo de languaje de bigramas se utiliza, la malla se expande como se muestra en la Figura 2.6. Las transiciones dentro de las palabras se mantienen igual como en la Figura 2.5. Entre palabras una transición se añade (mostrada en líneas punteadas) desde el estado final de una palabra al estado inicial de la siguiente, la cual está ponderada con la probabilidad del bigrama (par de palabras). Más información acerca de este algoritmo se puede encontrar en [27, 49, 61]

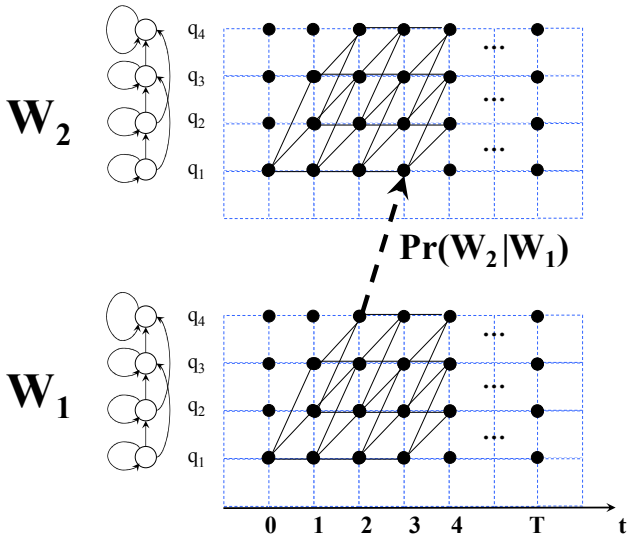


Figura 2.6: La malla de Viterbi para un bigrama.

2.1.5 Adaptación

Cuando las muestras de voz de entrenamiento se encuentran disponibles, los parámetros λ de los HMM's descritos en la Sección 2.1.3 son estimados de manera eficiente, fa-

voreciendo que el algoritmo de Viterbi produzca buenos resultados. Sin embargo este desempeño depende de las muestras de voz utilizadas para el entrenamiento, y su desempeño puede ser deficiente con usuarios distintos cuyas voces no se usaron para entrenar el SRAH.

En este caso, la técnica de adaptación de usuario conocida como Maximum Likelihood Linear Regression (MLLR) [31] y Maximum A-Posteriori (MAP)[61], se han desarrollado para ajustar los parámetros de los HMM's de un sistema Independiente de Usuario (IU), o Dependiente de Usuario (DU), a las características acústicas de un usuario en particular. Estas técnicas normalmente requieren de algunas muestras de voz del usuario (datos de adaptación) para estimar "transformaciones" que ajusten los parámetros de los HMM's a su voz. La adaptación es supervisada cuando hay conocimiento de las palabras pronunciadas por el usuario, y es no supervisada cuando no se tiene dicha información.

En el presente proyecto, MLLR se utilizará como técnica de adaptación, la cual se basa en el supuesto de que un conjunto de transformaciones lineales se puede usar para reducir la diferencia entre los modelos de un SRAH y los datos de adaptación. Estas transformaciones son aplicadas sobre la media y varianza de las mixturas de gaussianas de los HMM del sistema base (ver Sección 2.1.3, Ec. 2.2 y Ec. 2.3), teniendo el efecto de ajustar dichos parámetros de tal manera que aumente la probabilidad de que los HMM's del sistema generen los datos de adaptación.

MLLR se realiza en dos pasos:

- **Adaptación Global**. El primer requisito para permitir la adaptación es especificar el conjunto de estados (componentes de los HMM's) que comparten la misma transformación. Esto se realiza mediante una "clase base global". En este paso, una transformación global es generada y es aplicada a cada componente Gaussiano de los HMM's del sistema base.

- **Adaptación Dinámica**. En el segundo paso se utiliza la *transformación global* como transformación de entrada para adaptar los modelos, produciendo un mejor alineamiento para la estimación de transformaciones más específicas a ciertos componentes gaussianos mediante el uso de un *árbol de regresión de clases*. Este proceso se considera como dinámico ya que las transformaciones son estimadas

de acuerdo a la "cantidad" y "tipo" de datos de adaptación disponibles. La Figura 2.7 muestra la estructura de un ábol de regresión de classes, el cual es construído para agrupar componentes que sean similares acústicamente, pudiendo ser transformados de manera similar. Cada componente gaussiano de un HMM pertenece a una clase en particular, y el asociar cada transformación a un conjunto de mixturas favorece la adaptación de modelos para los cuales no hay datos disponibles. De esta manera, todos los modelos pueden ser adaptados de manera dinámica cuando más datos de adaptación se encuentran disponibles.

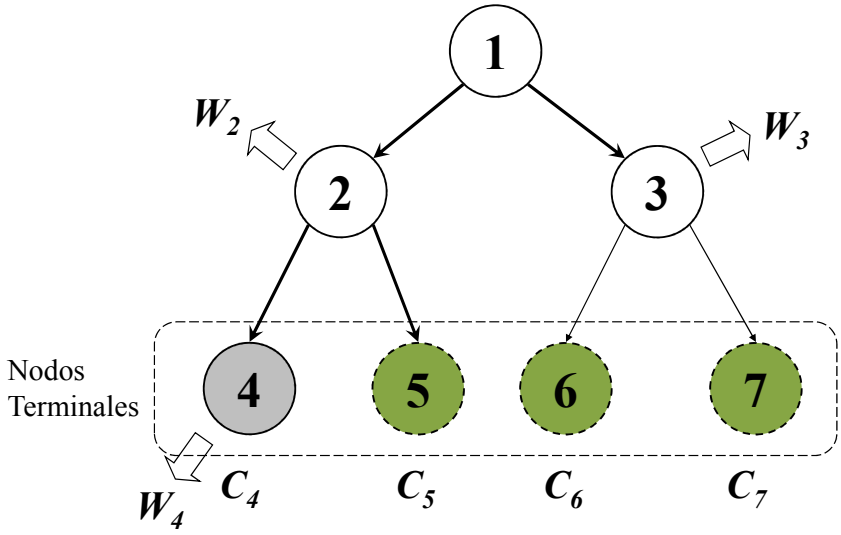


Figura 2.7: Árbol de Regresión Binario.

En la Figura 2.7 hay cuatro nodos terminales, o clases base, identificadas como $\{C_4, C_5, C_6$ y $C_7\}$. Nodos y flechas con lineas continuas indican que hay suficientes datos en esa clase para generar una matriz de transformación, y aquellos con lineas punteadas indican que no hay suficientes datos. Durante la adaptación dinámica, los componentes de las mixturas de los HMM que pertenecen a los nodos 2, 3 y 4 son usados para construir un conjunto de transformaciones definidas como W_2, W_3 y W_4. Cuando el modelo transformado es requerido, las matrices con las transformaciones lineales (para la media y covarianza) son aplicadas de la siguiente manera sobre los componentes gaussianos en cada clase base: $W_2 \rightarrow \{C_5\}$, $W_3 \rightarrow \{C_6, C_7\}$, y $W_4 \rightarrow \{C_4\}$. De esta forma se adaptan las

distribuciones de las clases con datos insuficientes (nodos 5, 6, y 7) y de aquellas con suficientes datos.

2.2 SRAHs con la Biblioteca HTK Toolkit

Al trabajar con SRAHs, uno de los temas que lo complica es el procesamiento de información, por ejemplo, para la búsqueda Viterbi, el análisis de vectores de voz, estimaciones de probabilidades, etc. Como se presentó en la sección anterior, los SRAH involucran muchos procesos estocásticos en su operación. Para agilizar algunos de estos procesos se han desarrollado bibliotecas como las de HTK Toolkit [61].

HTK es un conjunto de herramientas para el diseño y desarrollo de Modelos Ocultos de MarKov (HMM's) que fue creado para el área de RAH. En la actualidad es aplicable a muchas areas del conocimiento, siempre que el problema a solucionar pueda ser planteado como un Modelo Estocástico Markoviano [6, 52]. El uso de esta herramienta depende de dos aspectos: la línea de comando como interfaz con el sistema operativo, y módulos operacionales independientes [6, 52].

HTK es ampliamente utilizado en el ámbito de investigación para el diseño y desarrollo de SRAHs [6, 18, 24, 52, 59] dadas las siguientes características:

- es un software de libre distribución;
- HTK tiene una estructura robustas, autosuficiente, y permite diseñar HMM's desde su nivel fundamental [6];
- utilizable en diversas plataformas como Windows, Linux, Unix, y DOS [61].

En la Figura 2.8 se muestran los diferentes módulos de HTK involucrados en el desarrollo de los elementos de un SRAH mostrados en la Figura 2.1. En las siguientes secciones se presentan los detalles de construcción de cada uno de estos elementos con herramientas estándar incluyendo el HTK toolkit.

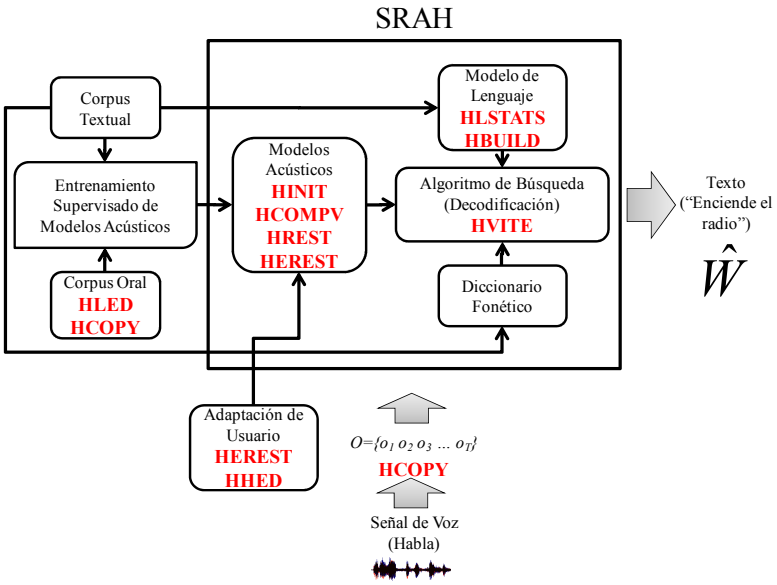


Figura 2.8: Módulos o bibliotecas de HTK usados para el diseño y desarrollo de cada uno de los elementos de un SRAH.

2.2.1 Corpus Oral, Etiquetado, y Diccionario Fonético

Grabación de Voz y Etiquetado Ortográfico

Como se presentó en la sección 2.1.1, Figura 2.2, incialmente un corpus textual es necesario para comenzar a grabar el corpus oral. En su forma más sencilla estas muestras de voz se pueden grabar en formato WAV monoaural con velocidad de muestreo de 8 kHz [61]. Actualmente hay diversidad de herramientas de uso libre para la grabación de muestras de voz, sin embargo para este proyecto se hará referencia al software *WaveSurfer* [2]. Con este programa, además de poderse grabar voz, se puede hacer el etiquetado de la misma.

En la Figura 2.9 se muestra el WaveSurfer una vez que se ha grabado la frase "el texto es significativo" presionando el botón rojo (controles superior-derecho). Al terminar de grabar se puede visualizar la forma de onda al fondo de la interfaz, la cual se puede complementar con su respectivo espectrograma de frecuencia. Para hacer el etiquetado de archivos de voz, WaveSurfer maneja varios estándares los cuales se muestran en la Figura 2.9. El más usado es el TIMIT [17][3], el cual considera etiquetas ortográficas

[3]Este nombre viene del corpus en inglés americano TIMIT, que está etiquetado en ambos niveles y es de los más extensos en dicho idioma para el desarrollo de SRAHs.

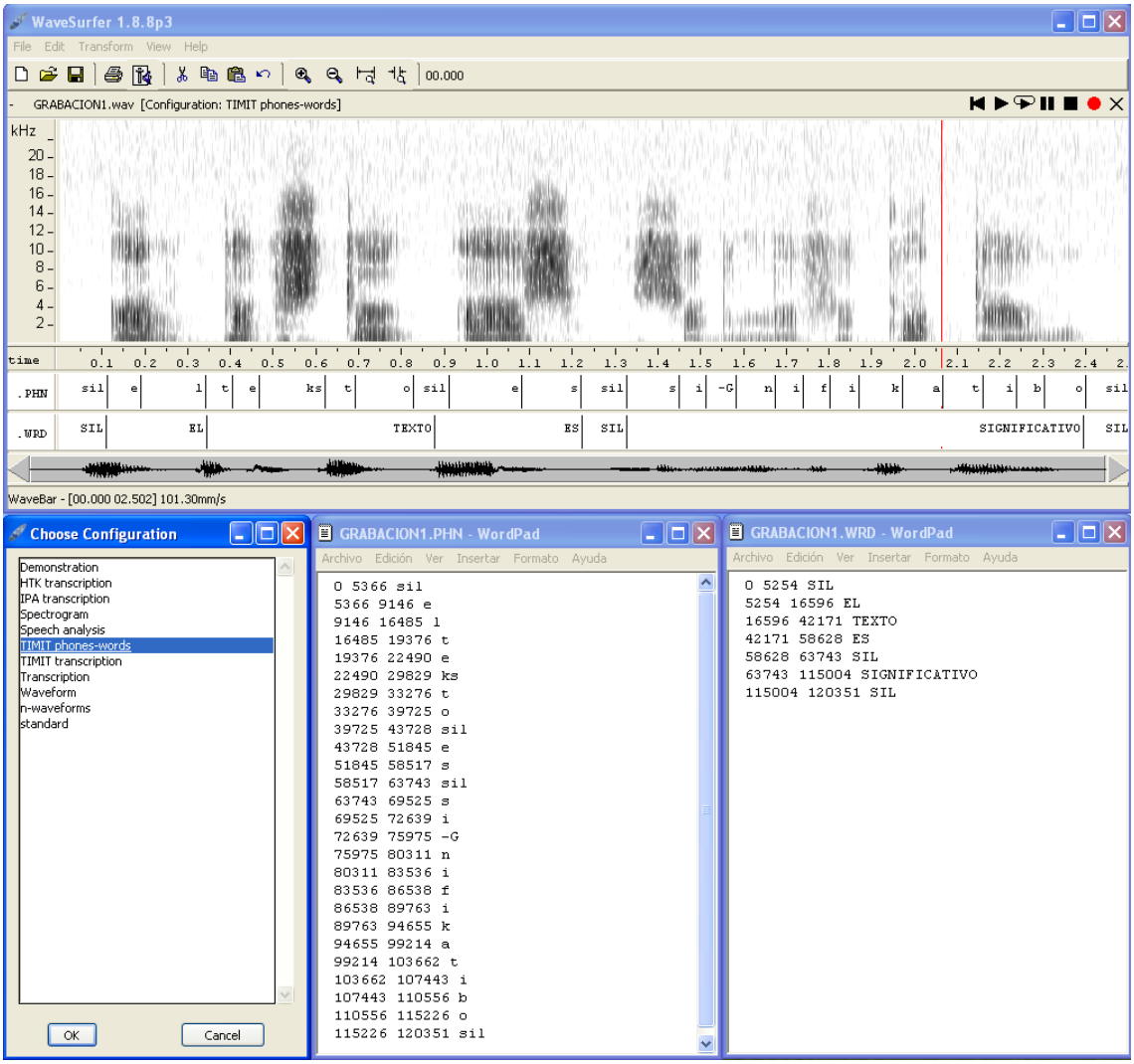


Figura 2.9: Grabación de voz y etiquetado manual en WaveSurfer.

(.WRD) y fonéticas (.PHN). Mediante el cursor de la interfaz se delimitan los segmentos correspondientes a cada palabra o fonema, pudiendo ingresar el nombre (etiqueta) de cada uno. Sobretodo para el etiquetado fonético, el uso del espectrograma facilita la diferenciación de fonemas. Al guardar el proyecto de WaveSurfer, éste crea los archivos correspondientes a las etiquetas fonéticas y ortográficas del archivo de voz asignándoles el mismo nombre. En la Figura 2.9 se muestran estos archivos de texto, observándose una similitud con los conceptos presentados en la Figura 2.3. Los números que aparecen a la izquierda de las etiquetas corresponden a los rangos de tiempo (o longitud) del segmento de voz que corresponden a dichas etiquetas.

Diccionario Fonético

Es importante recordar que para realizar el etiquetado fonético es necesario conocer acerca de la fonética del idioma en cuestión, en este caso del español mexicano. Recursos como los **diccionarios fonéticos** han facilitado esta tarea al proveer de las secuencias de fonemas que corresponden a una palabra en particular. En este proyecto se utilizó el *TranscribEMex*, que es una utilidad en lenguaje *perl* desarrollada para el etiquetado fonético[4] de corpus orales para el español de la ciudad de México [46]. Este herramienta fue desarrollada por el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS) de la Universidad Autónoma de México (UNAM), y fue usado para el etiquetado del corpus DIMEx100[5][47]. En la Figura 2.10 se muestra una ejecución de este programa.

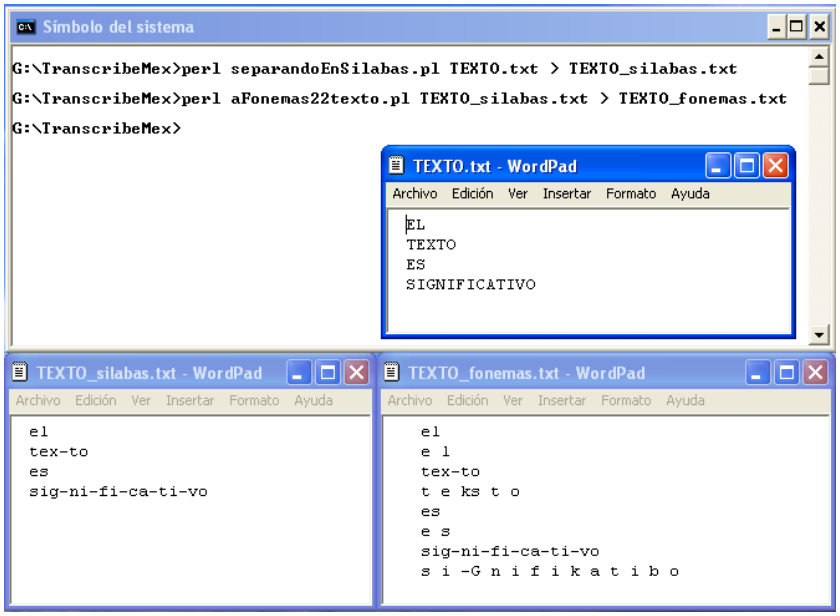


Figura 2.10: Transcripción fonética de texto usando TranscribEMex.

Para obtener la transcripción fonética de una palabra (en TEXTO.txt) primero se separa en sílabas con el programa *separandoEnSilabas2.pl*. El archivo generado por este programa (TEXTO_silabas.txt), mostrado en la Figura 2.10 ahora se convierte en

[4]La definición de fonemas usa la codificación propuesta por el Mtro. Javier Cuétara [11].

[5]Este corpus fue considerado inicialmente para realizar el presente proyecto. Sin embargo, actualmente este recurso se encuentra en procedimientos legales para su licencia de uso, y por lo tanto no disponible.

el archivo de entrada para la separación en fonemas, la cual es implementada con el programa *aFonemas22texto.pl*. El archivo resultante (TEXTO_fonemas.txt) incluye (1) la palabra original seguida de (2) su representación en sílabas, y (3) la secuencia fonética respectiva. *aFonemas22texto.pl* utiliza 22 fonemas principales para la transcripción fonética, y 5 más para definir fonemas con particularidades en su pronunciación (variación en coarticulación). Adicionalmente se consideran fonemas estándar en RAH, denominados como $/sil/$ y $/sp/$, que se utiliza para identificar *silencio* y *pausa corta* entre palabras. Por lo tanto, para el desarrollo de un SRAH para el español mexicano se consideran 29 fonemas en total, los cuales se muestran en la Tabla 2.1.

Tabla 2.1: Fonemas para el español mexicano definidos por TranscribEMex.

Fonemas del Español Mexicano			
/a/	/m/	/r(/	/sil/
/b/	/n/	/_D/	/sp/
/tS/	/ñ/	/_G/	
/d/	/o/	/_N/	
/e/	/p/	/_R/	
/f/	/r/		
/g/	/s/		
/i/	/t/		
/x/	/u/		
/k/	/ks/		
/l/	/Z/		

De esta manera, al tener el texto original y su transcripción fonética, se puede llevar a cabo el etiquetamiento completo el corpus oral con WaveSurfer. En la Figura 2.11 se muestra un ejemplo del diccionario fonético para el sistema propuesto, el cual sigue el formato requerido por la biblioteca HTK.

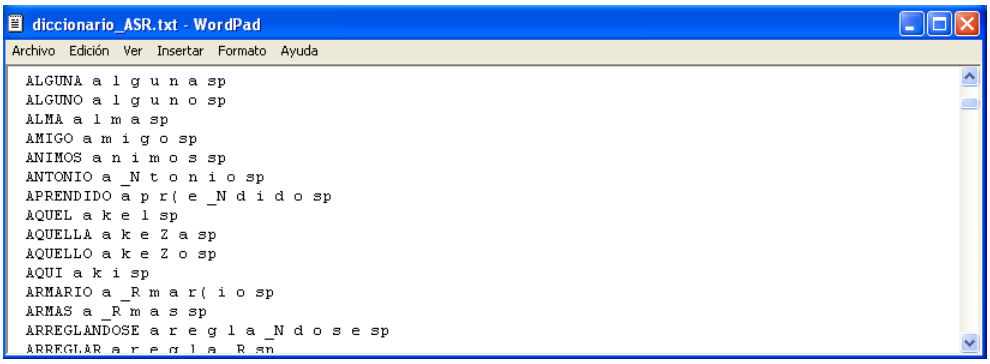


Figura 2.11: Ejemplo de diccionario fonético usando TranscribEMex.

Etiquetado Fonético

En este momento se presentará la primera biblioteca de HTK, **HLEd**, que puede ser utilizada para agilizar el etiquetado fonético si ya se cuenta con el ortográfico (con los tiempos definidos para cada palabra) y el diccionario fonético. En la Figura 2.12 se presenta la ejecución de HLEd, la cual genera un etiquetado fonético (GRABACION1.PHN) a partir de etiquetas ortográficas (GRABACION1.WRD) y el diccionario fonético (DICT.txt). Note que, en comparación con el etiquetado fonético de la Figura 2.9, el cual fue hecho manualmente, las etiquetas proporcionadas por HLEd no son tan precisas. Esto se debe a que HLEd sólo divide el tiempo correspondiente a cada palabra entre el número de fonemas que la forman en el diccionario fonético, quedando asignado un segmento constante a cada fonema.

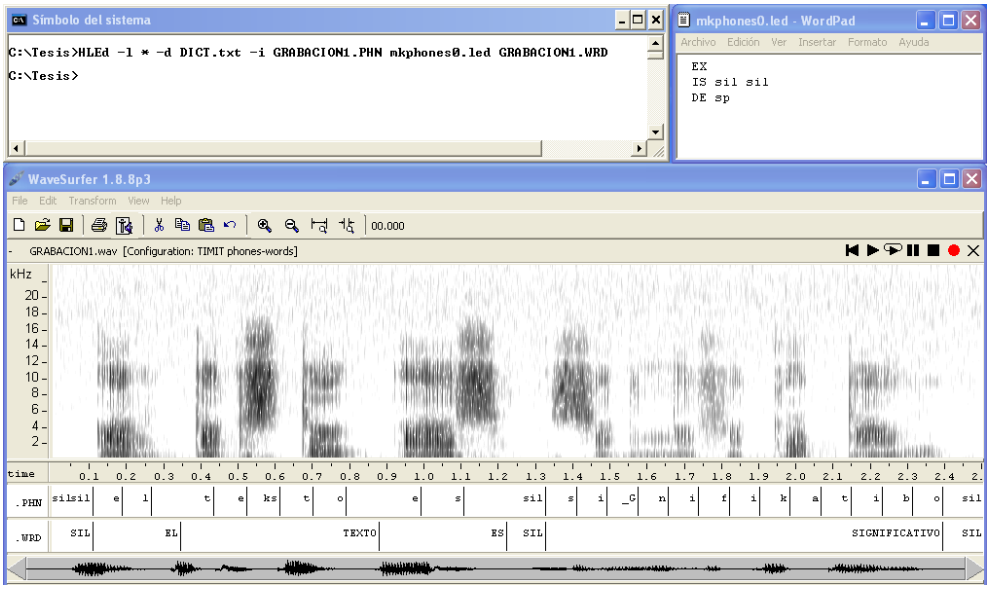


Figura 2.12: Etiquetado fonético usando la biblioteca HLEd de HTK.

Otra desventaja de HLEd es que la división se hace entre palabras separadas por la etiqueta de silencio (SIL). Es por esto que las etiquetas fonéticas para la palabra SIGNIFICATIVO se encuentran equidistantes, en tanto que las de TEXTO se encuentran más dispersas, habiendo segmentos de fonemas asignados fuera de los límites de esta palabra. En este proyecto se hizo un programa para hacer la segmentación, independientemente de la existencia de una etiqueta de silencio entre palabras. Sin embargo, para propósitos de práctica, este procedimiento con HLEd es aceptable [61].

Codificación de Corpus Oral

Ya que se tiene el corpus completo, es necesario extraer información de las muestras de voz de tal manera que se optimize el proceso de RAH. Para esto, la señal se codifica en formatos específicos, siendo los Coeficientes Cepstrales en las Frecuencias de Mel (Mel Frequency Cepstral Coefficients, MFCCs) [27, 61] el formato más utilizado para SRAHs. Los MFCCs se derivan de la Transformada de Fourier (FT) o de la Transformada de Coseno Discreta (DCT). La diferencia básica entre ambas y los MFCCs es que en estos últimos las bandas de frecuencia se sitúan logarítmicamente según la escala de Mel, la cual modela la respuesta (percepción) auditiva humana más apropiadamente que las bandas espaciadas linealmente de FT o DCT. Esto permite la compresión de audio y el procesamiento más rápido de la señal de voz para RAH [13]. Los MFCCs se consideran *vectores de información espectral* $\mathbf{o}_t$ (véase Figura 2.4) de la señal de voz, los cuales son la entrada para el SRAH.

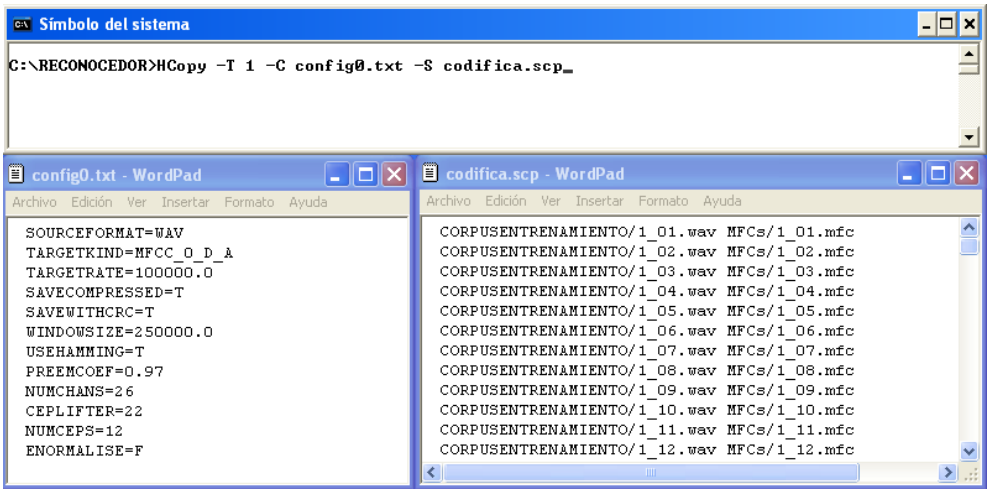


Figura 2.13: Codificación de voz en MFCCs usando la biblioteca HCopy de HTK.

Aquí es en donde se introduce otra biblioteca de HTK, **HCopy** para codificar el corpus oral en MFCCs. En la Figura 2.13 se presentan los elementos involucrados en el proceso y la ejecución de la codificación. HCopy necesita los siguientes archivos:

- Un archivo de configuración, config0.txt, en el cual se indiquen las características de los MFCCs (TARGETKIND=MFCC_0_D_A) y el formato de origen del archivo de voz en WAV (SOURCEFORMAT=WAV). Usualmente se especifica que la

señal se muestree cada 10 milisegundos (TARGETRATE=100000.0) con una ventana Hamming de 25 milisegundos para su codificación (WINDOWSIZE=250000.0, USEHAMMING=T). Adicionalmente se especifica que se usarán 12 MFCCs (NUMCEPS=12), con un coeficiente adicional que representa la energía de la señal acústica [27, 61]. A cada uno de estos coeficientes (13 en total) se les añadió coeficientes delta (_D) y de aceleración (_A), dando una codificación de 39 coeficientes. Esto mejora la detección de fonemas [27, 61].

- Un archivo de registro, codifica.scp, en donde se especifiquen los archivos de audio originales (entrada para codificar) y los codificados en MFCCs (salida codificada). Como se muestra en la Figura 2.13, la dirección de los archivos se incluye y se define con respecto al directorio raíz.

Una vez codificado el corpus se procede a entrenar de manera supervisada los modelos acústicos del SRAH. Esto se presenta en la siguiente sección.

2.2.2 Entrenamiento Supervisado de los Modelos Acústicos

El realizar el entrenamiento de un modelo acústico involucra dos procesos: inicialización y re-estimación. A continuación se presenta, de manera general, la secuencia de ejecución y configuración de bibliotecas de HTK para el entrenamiento de HMM's. Para mayor información se recomienda consultar el manual de HTK [61].

Como se presentó en la Sección 2.1.3, HMM's son la técnica más común para el desarrollo de SRAHs. Los modelos acústicos para fonemas se construyeron siguiendo una topología estándar de tres estados (con dos estados no emisores) con secuencia izquierda-a-derecha (ver Figura 2.4).

En la Figura 2.14 se muestra la declaración de un HMM prototipo en HTK con la topología antes mencionada. Note que cada estado emisor (2, 3, y 4) tiene un vector de media (Mean) y varianza (Variance). Estos corresponden a las distribuciones de probabilidad de observación, definidas por las Ec.'s 2.2 y 2.3. El prototipo de HMM (nombrado *proto*) se inicializa con 0's para la media, y 1's para la varianza, que son los valores estándar de una distribución normal $N(0, 1)$ [61]. Ya que el SRAH se hará a nivel fonético, se tendrá un prototipo para cada uno de los fonemas del español mexicano, con excepción del fonema $/sp/$.

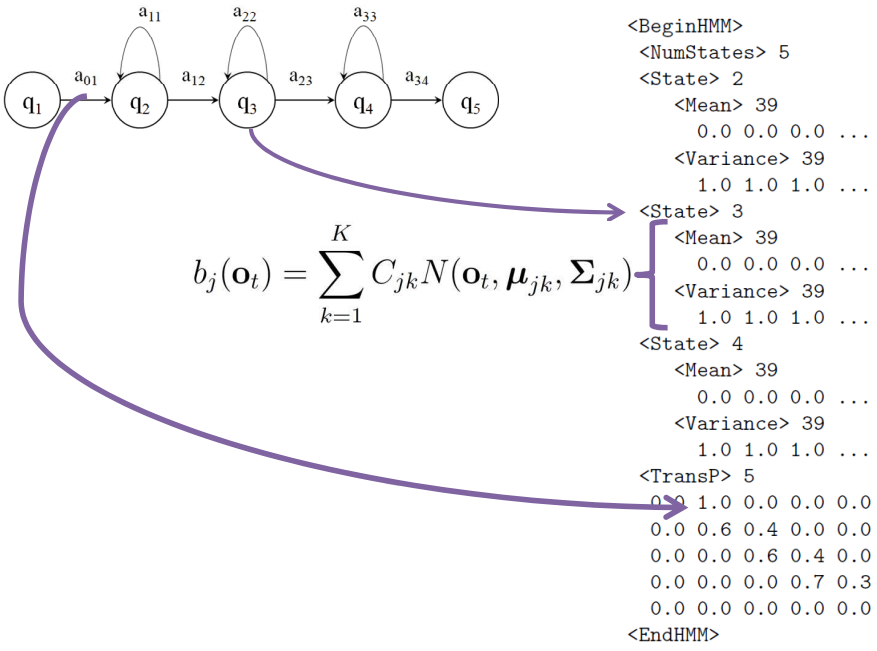


Figura 2.14: Declaración de un HMM en HTK con un solo componente gaussiano. Archivo *proto*.

La **inicialización** de los modelos acústicos con el corpus se realiza mediante la biblioteca **HInit** de la siguiente manera:

HInit -A -D -T 1 -C config.txt -S entrena.scp -M **protobase** -H *proto* -l *fonema* -L **etiquetas** *proto* -X phn

en donde:

- config.txt es un archivo de configuración igual a config0.txt pero sin la línea SOURCEFORMAT=WAV. Esto dado que el archivo se usará para configurar la lectura de archivos en lugar de su codificación a otro formato.
- entrena.scp, que es un archivo de texto en donde se especifica la lista de archivos de sonido para el entrenamiento supervisado. Similar a codifica.scp, pero sin la columna de datos en donde se especifican los archivos de origen a codificar.
- **etiquetas**, que es un directorio en donde se encuentran las etiquetas ortográficas

y fonéticas de los archivos de sonido (.WRD, .PHN). Con "-X phn" se le indica a HInit que utilice las etiquetas fonéticas para el entrenamiento.

- **protobase**, el directorio en donde se almacena el nuevo HMM inicializado con las características del *fonema* indicado. Note que este archivo de HMM se sigue llamando *proto*, por lo que es necesario renombrarlo con el nombre del *fonema* seleccionado.

Este proceso se repite para cada uno de los fonemas del español mexicano, con excepción de $/sp/$. Después de este paso se estima un valor global inicial para la varianza de los HMM's, el cual se guarda en el archivo *vFloors*. Esto se realiza con la biblioteca **HCompV** de la siguiente manera:

HCompV -A -D -T 1 -C config.txt -S entrena.scp -M **flat** -H **protobase/***fonema* -f 0.01 *fonema*

en donde el archivo *vFloors* se guarda en el directorio **flat**. Para esto sólo es necesario un HMM, ya sea el de $/a/$, $/b/$, $/Z/$, etc. Una vez que se tienen los modelos inicializados para cada fonema, éstos se copian del directorio **protobase** a **hmm0** y se *re-estiman* usando el comando HREst:

HRest -A -D -T 1 -C config.txt -S entrena.scp -M **hmm1** -H *vFloors* -H **hmm0/***fonema* -l *fonema* -L **etiquetas** *fonema* -X phn

Este proceso se repite tres veces, en donde los modelos re-estimados de **hmm0** se guardan en **hmm1**, de **hmm1** a **hmm2**, y de **hmm2** a **hmm3**. Esta biblioteca utiliza el algoritmo de Baum-Welch o Forward-Backward [27]. Este modo de entrenamiento se realiza de modo iterativo hasta que hay convergencia de los parámetros del HMM para cada fonema del conjunto.

En este punto se tienen en **hmm3** un archivo con un HMM para cada fonema (28 modelos en total). Para hacer más manejable el uso de modelos, éstos se unen en un solo archivo de definición de modelos acústicos llamado *hmmdefs*, y *vFloors* se renombra como *macros*. Esto se puede hacer con la biblioteca HLEd de la siguiente manera:

HLEd -d **hmm3** -w *hmmdefs mixtures* lista_fonemas.txt

Note que ya no es necesario llamar a cada *fonema* para su re-estimación, ya que ahora todos los harán al mismo tiempo. Es por esto que ahora hay un archivo lista_fonemas.txt, el cual contiene la lista de los 28 fonemas (HMM's). Adicionalmente, mediante el archivo de configuración *mixtures* se puede especificar el número de mixturas gaussianas que se utilizarán para modelar las distribuciones de probabilidad de cada estado de los HMM's. Mediante este procedimiento se puede mejorar el modelado acústico.

El archivo de HMM's final *hmmdefs*, y el que lleva información global de la varianza de los datos *macros*, se copian al directorio **hmm4** para su re-estimación con la biblioteca HERest. Este módulo es más eficiente para HMM's con componentes gaussianos, y se usa de la siguiente manera:

HERest -C config.txt -L **etiquetas** -X phn -t 250 150 1000 -S entrena.scp -H **hmm4/***macros* -H **hmm4/***hmmdefs* -M **hmm5** lista_fonemas.txt

Este proceso se repite dos veces, terminando los modelos re-estimados en el directorio **hmm6**. En este punto es cuando a los modelos de los fonemas en *hmmdefs* se les añade el fonema $/sp/$, el cual representa a las pausas cortas existentes entre palabras.

Este modelo se crea manualmente a partir del modelo del silencio $/sil/$ y se pega dentro del archivo *hmmdefs* el cual se guarda en el directorio **hmm7**. Finalmente se integra dentro de los modelos mediante la biblioteca HHEd de la siguiente manera:

HHEd -H **hmm7/***macros* -H **hmm7/***hmmdefs* -M **hmm8** *sil.hed* lista_fonemas.txt.

en donde *sil.hed* es un archivo de configuración en donde se indica qué mixturas de gaussianas del modelo $/sil/$ se van a relacionar con el modelo $/sp/$. En el archivo *lista_fonemas.txt* se añade *sp*. El *hmmdefs* integrado se guarda en el directorio **hmm8**, y se re-estima dos veces para terminar en **hmm10**:

HERest -C config.txt -L **etiquetas** -t 250 150 1000 -S entrena.scp -H **hmm8/***macros*

-H **hmm8/***hmmdefs* -M **hmm9** lista_fonemas.txt

Finalmente se realiza un re-alineamiento de los patrones de los HMM's con los etiquetados fonéticos extraídos directamente de los etiquetados ortográficos (guardados en un único archivo *e_ortografico.mlf*). Para esto se utiliza el algoritmo de Viterbi, el cual es implementado por la biblioteca HVite y se ejecuta de la siguiente manera:

HVite -l '*' -o SWT -b silence -C config.txt -a -H **hmm10/***macros* -H **hmm10/***hmmdefs* -i *alineado_ortografico.mlf* -m -t 250 -y lab -I e_ortografico.mlf -S entrena.scp diccionario.txt lista_fonemas.txt

Esta instrucción genera la salida de fonemas (*alineado_ortografico.mlf*) comparando las señales acústicas y los etiquetados ortográficos, la cual servirá para la re-estimación final de los HMM's del SRAH:

HERest -C config.txt -i *alineado_ortografico.mlf* -t 250 150 1000 -S entrena.scp -H **hmm10/***macros* -H **hmm10/***hmmdefs* -M **hmm11** lista_fonemas.txt

Los modelos acústicos finales re-estimados se guardan en el diccionario **hmm12**. Note que en las etiquetas ortográficas inicialmente al silencio se le asoció el identificador SIL (ver Figura 2.9), el cual sirvió para identificar el fonema $/sil/$ en los etiquetados fonéticos. Subsecuentemente para el re-alineamiento y estimación del Modelo de Lenguaje (ver Sección 2.2.4) este identificador se eliminó ya que gramaticalmente no tiene significado.

El proceso de reconocimiento se realiza con la siguiente instrucción:

HVite -C config.txt -H **hmm12/***macros* -H **hmm12/***hmmdefs* -S *prueba.scp* -l '*' -i *salida.mlf* -w *ML* -p 0 -s 5 diccionario.txt lista_fonemas.txt

en donde *ML* es el archivo del modelo de lenguaje del corpus textual de entrenamiento, *diccionario.txt* el diccionario fonético que incluye todas las palabras del corpus textual, y *entrena.scp* la lista de archivos de sonido que se reconocerán. La salida

del reconocedor se guarda en el archivo *salida.mlf*.

Hay un parámetro en especial que se conoce como el *factor de gramática, -s*, el cual regula o penaliza la influencia del modelo de lenguaje sobre la evidencia acústica en el proceso de reconocimiento. Por ejemplo, si $-s$=0 el algoritmo de Viterbi no utiliza la información del ML para estimar la secuencia de palabras correspondientes a los archivos de sonido dados, conforme $-s$ aumenta esta información es considerada. Si $-s$=100, el reconocedor usaría sólo la información del ML y muy poco de los archivos de sonido. Usualmente este parámetro se ajusta manualmente hacia valores dentro del rango 5-15 para voz normal.

2.2.3 Adaptación de Usuario

Como se presentó en la sección 2.1.5, un SRAH se puede adaptar a los patrones de voz de usuarios diferentes a aquellos con los que se entrenó. La adaptación MLLR, que se realiza en dos etapas (global y dinámica), se realiza en HTK de la siguiente manera:

HHEd -H **hmm12/***macros* -H **hmm12/***hmmdefs* -M **classes** regtree.hed lista_fonemas.txt

A partir de los fonemas descritos en los archivos de definición *hmmdefs* se estiman las relaciones entre fonemas para estimación de transformaciones de adaptación. Esto es, la creación de un árbol de regresión en donde cada nodo se asocia a conjuntos de fonemas. Mediante el archivo de configuración *regtree.hed* se define el número de nodos a 32. La ejecución de la instrucción por lo tanto genera *rtree.tree* que enumera los nodos y diferentes ramificaciones del árbol, y *rtree.base* la relación de componentes gaussianos de cada fonema que compartirán las mismas transformaciones (ver Sección 2.1.5).

Teniendo estos parámetros se procede a estimar la transformación global:

HERest -C config.txt -C *config.global* -S adaptar.scp -I *fonemas_adaptacion.mlf* -H **hmm12/***macros* -u a -H **hmm12/***hmmdefs* -K **adaptacion** mllr1 -J **classes** -z TMF lista_fonemas.txt

en donde *config.global* es un archivo de configuración en donde se indica que la

transformación se aplicará sobre la media de los componentes gaussianos, *fonemas_adaptacion.mlf* es la transcripción fonética correcta de las frases de adaptación, *adaptar.scp* la lista de los archivos de sonido del nuevo usuario para adaptación (lecturas de las frases de adaptación), **adaptacion** es el directorio en donde se va a guardar la transformación global con extensión mllr1, y **classes** el directorio en donde se encuentran los archivos del árbol de regresión. Finalmente la siguiente instrucción:

HERest -a -C config.txt -C config.rc -S adaptar.scp -I *fonemas_adaptacion.mlf* -H **hmm12/***macros* -u a -H **hmm12/***hmmdefs* -J **adaptacion** mllr1 -K **adaptacion** mllr2 -J **classes** -z TMF lista_fonemas.txt

implementa la adaptación dinámica, en donde transformaciones más específicas se generan usando la transformación global y el árbol de regresión, guardándolas en el directorio **adaptacion** con extensión mll2. El reconocedor adaptado al nuevo usuario se ejecuta con la siguiente instrucción:

HVite -C config.txt -S prueba.scp -J **adaptacion** mllr2 -k -J **classes** -H **hmm12/***hmmdefs* -H **hmm12/***macros* -l '*' -i *salida.mlf* -w *ML* -p 0 -s 5 diccionario.txt lista_fonemas.txt

en donde:

- prueba.scp, es el archivo en donde se encuentran listados los archivos de sonido a reconocer (prueba).
- *salida.mlf*, el archivo de texto de salida, en donde se encontrarán las palabras reconocidas por el sistema.
- *ML*, el modelo de lenguaje del sistema.
- diccionario.txt, el diccionario fonético del sistema.
- -s 5, el valor estándar para la penalización del modelo de lenguaje sobre la evidencia acústica al implementar el algoritmo de Viterbi para RAH.

2.2.4 Modelo de Lenguaje

Para ejecutar el reconocedor mediante el modulo HVite (que implementa el algoritmo de Viterbi) es necesario construir el Modelo de Lenguaje del SRAH. Este es estimado apartir del corpus textual. Representa un conjunto de reglas o probabilidades que determinan las secuencias de palabras permisibles en un lenguaje (ver sección 2.1.2).

Para la creación del ML con HTK se utilizaron los etiquetados ortográficos sin el identificador SIL para silencio. Primero se ejecutó la biblioteca HLStats el cual estima información estadística concerniente a la frecuencia (número de ocurrencias) de aparición de palabras en el corpus textual. Ya que se usarán bigramas, también se estima la frecuencia de los diferentes pares de palabras en el corpus:

HLStats -b *bigrama* -o lista_de_palabras.txt *e_ortografico.mlf*

Estas estadísticas se guardan en el archivo *bigrama*, las cuales se utilizan para construir una red de palabras que represente las diferentes secuencias de palabras posibles. Para esto se utiliza la biblioteca HBuild:

HBuild -n *bigrama* lista_de_palabras.txt *ML*

HBuild utiliza las estadísticas del archivo *bigrama* y construye el modelo de lenguaje *ML* el cual ya puede ser utilizado por HVite para ejecutar el proceso de reconocimiento de voz.

2.2.5 Métricas de Desempeño

La métrica de desempeño para un SRAH es el Porcentaje de Precisión de Reconocimiento de Palabras (Word Accuracy, WAcc) [61], la cual se calcula como:

$$WAcc = \frac{N - D - S - I}{N} \tag{2.6}$$

Esta métrica se calcula mediante un alineamiento entre la transcripción correcta W (referencia) de la señal de voz, y la secuencia de palabras decodificada $\hat{W}$ (reconocida) por el SRAH para la misma señal. De esta manera se tiene que, para la Ec. 2.6, N es

el número de palabras en W, D el número de palabras en W que no aparecen en $\hat{W}$ (eliminaciones), S el número de palabras en W que fueron confundidas con otras en $\hat{W}$ (sustituciones), e I el número de palabras extra que aparecen en $\hat{W}$ pero no tienen correspondencia con ninguna en W (inserciones).

Una métrica paralela al WAcc que se utiliza ampliamente para medir el desempeño de estos sistemas es la Tasa de Error de Palabras (Word Error Rate, WER) [61], la cual se expresa como:

$$WER = 1 - WAcc = 1 - \frac{N - D - S - I}{N} \tag{2.7}$$

La biblioteca HResults de HTK puede calcular el WAcc siempre y cuando se tengan las transcripciones de referencia W de la siguiente manera:

HResults -I *e_ortografico.mlf* lista_fonemas.txt *salida.mlf*

En la Figura 2.15 se muestran las estadísticas generadas por HResults, en donde:

- SENT identifica el porcentaje de frases reconocidas de manera completa.
- WORD identifica el número de palabras reconocidas correctamente, considerando sustituciones y eliminaciones. Esto es: WORD=(N-D-S)/N.
- Acc es la precisión del reconocedor (WAcc), que se estima igual que WORD pero considerando también las inserciones (ver Ec. 2.6).

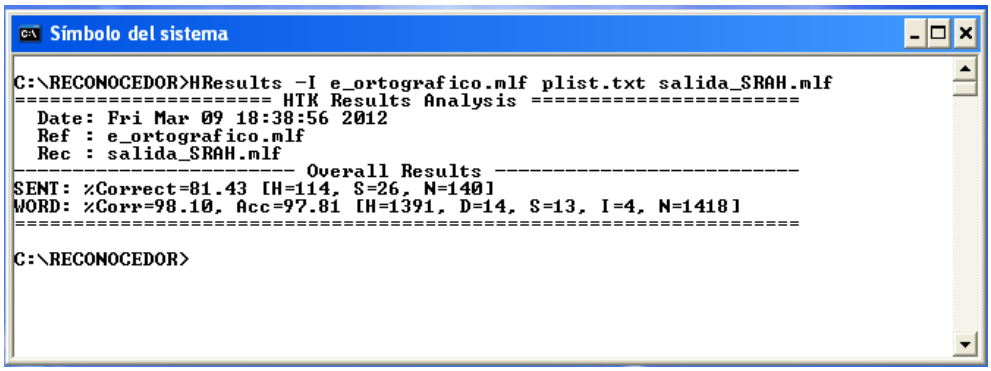


Figura 2.15: Ejecución de HResults para estadísticas de desempeño.

Capítulo 3

La Disartria y las Tecnologías de Asistencia

3.1 Disartria

De entre las diferentes formas que el ser humano tiene para comunicarse, el habla es la más significativa. El habla se utiliza para llevar acabo la comunicación de forma verbal, y dar o recibir un mensaje.

Por comunicación se entiende cualquier interacción que transmite información. Relatar, informar, explicar, y expresarse son funciones de una comunicación, lo cual implica enviar y recibir mensajes con significado [43]. En la Figura 3.1 se muestra la manera en que normalmente se establece una adecuada comunicación verbal con otras personas, lo cual permite el interactuar con el medio ambiente y el social. Como se presentó en el Capítulo 1, muchas personas en México sufren de alguna discapacidad que nos les permite llevar a cabo este proceso.

La disartria se refiere a un grupo de desórdenes motores del habla que resulta del déficit en el control muscular de los mecanismos del habla debido a un daño en el sistema nervioso periférico o central [33]. Este trastorno del habla es el más común, afectando 170 de cada 100,000 personas en países desarrollados [16]. En este caso, el problema del habla es debido a un estado neuromuscular anormal (parálisis, atrofia, espasticidad) o el resultado de la disrupción de los movimientos de esos músculos (debilidad o falta de coordinación). Por lo tanto, a menudo hay un rango reducido en los movimientos y flex-

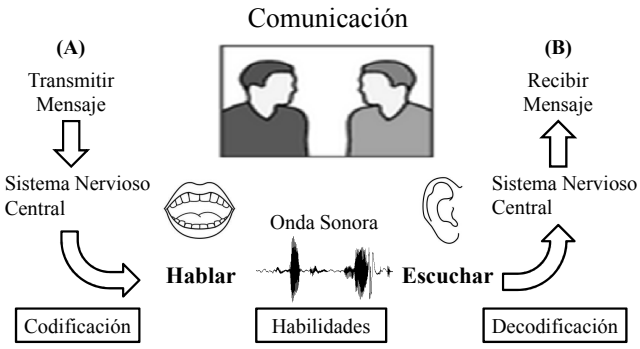


Figura 3.1: Interacción por medio de comunicación verbal.

ibilidad, siendo las características que generalmente se observan en el habla disártrica: hipernasalidad, consonantes imprecisas, distorsiones vocales, y problemas en el control de la velocidad [54].

Es por esto que la disartria comprende las disfunciones motoras de la respiración, fonación[1], resonancia, articulación[2] y prosodia.

Esta discapacidad puede ser causada por lesiones o enfermedades, como las que se mencionan a continuación [8, 58]:

- Por un daño cerebral debido a un tumor, accidente cerebrovascular, o lesión traumática.
- Por un daño a los nervios que inervan los músculos faciales como un traumatismo facial o cervical, cirugía para cáncer de cabeza y cuello (extirpación parcial o total de la lengua o la laringe).
- Por enfermedad que afecte a nervios y músculos (enfermedades neuromusculares) como son la parálisis cerebral, esclerosis múltiple, distrofia muscular, Mal de Parkinson.
- Por otras causas: Intoxicación con alcohol, prótesis dentales mal ajustadas, efectos secundarios de medicamentos que actuán sobre el sistema nervioso central, como narcóticos, fenitoína o carbamazepina.

[1]La producción de un sonido por la laringe que se origina por medio del flujo de aire expulsado que hace vibrar pliegues vocales o "cuerdas vocales" (Douglas,2002).

[2]El control y modelamiento del sonido producido en la laginge por las cavidades nasal y oral, y por órganos que cumplen la función valvular (labios, dientes, lengua, mejillas, paladar, velo del paladar y movimientos mandibulares) (Marchesan,2004).

De entre las secuelas que pueden dejar las lesiones en el sistema nervioso (y que afectan a la expresión del lenguaje del sujeto disártrico), se pueden citar las siguientes [33]:

- Deformaciones en la articulación por la dificultad motríz que se presenta y la falta de coordinación y control en los movimientos, pudiendo llegar a tener una expresión casi ininteligible.

- Trastornos respiratorios, con falta de sincronía entre la respiración y la fonetización, presentándose en algunos casos contracciones y espasmos que entorpecen el acto de la respiración y de la fonetización.

- Alteraciones en el tono de hipertonía o distonía, dificultando la articulación de la palabra cuando cualquiera de estos síntomas afecta a la zona buco-facial.

Como se muestra en la Tabla 3.1 existen varios tipos de disartria, existiendo personas gravemente afectadas con capacidad limitada o nula en el control de su cuerpo, el cual nos les permite interactuar con su medio ambiente, ser independiente, o utilizar otro medio de comunicación (por ejemplo, teclados, pantallas táctiles, etc.).

Por el sitio de la lesión	Disartria por trastornos de la motoneurona superior.		
	Disartria por trastornos de la motoneurona inferior.		
	Disartria por trastornos cerebelosos.		
	Disartria por trastornos estrapiramidales.		
	Disartria por trastornos de múltiples sistemas motores.		
Por las características sintomatológica	Disartria espática.		
	Disartria flácida.		
	Disartria atáxica.		
	Disartria discinética	Hipercinética	Corea, atetosis, temblor, distonía, discinesia tardía.
		Hipocinética	Parkinson
	Disartria mixta		
Por las características articulatorias	Disartria generalizada (trastornos en la articulación general y aislada.		
	Estado disártrico (evolutiva cambiante).		
	Restos disártricos (trastornos en la articulación general y no en la aislada).		

Tabla 3.1: Clasificación de disartria [12, 43, 48].

3.1.1 Sintomatología

La sintomatología se denomina de los distintos errores que se dan en el habla de la persona, los cuales se definen a partir de la raíz griega del fonema afectado: rota ($/r/$), sigma ($/s/$), lambda ($/l/$), etc. En donde el sufijo "tismo" o "cismo" se utiliza cuando el fonema no se articula correctamente (por ejemplo, sigmatismo o sigmacismo = dificultad para articular la $/s/$). Con la misma raíz y el prefijo "para" se define el error existente cuando el fonema es sustituído por otro (por ejemplo, pararrotacismo o pararrotatismo = cuando se sustituye el fonema $/r/$ por otro, generalmente $/g/$, $/d/$ o $/l/$) [39].

Los síntomas específicos son la sustitución, omisión, inserción y distorsión de los fonemas:

- **Sustitución**: es el error de articulación por el cual un sonido es reemplazado por otro. La persona no puede realizar una articulación y la suple por otra más fácil o, de entrada, percibe mal el sonido y lo reproduce tal como él lo discrimina (como lo emite). Es el error más frecuente dentro de las disartrias funcionales y el que presenta más dificultades para su corrección. Las formas más frecuentes son la sustitución de $/r/$ por $/d/$ o por $/g/$, de $/s/$ por $/z/$, y del sonido $/k/$ por $/t/$.

- **Omisión**: se omite el fonema (se pronuncia "iño" por "niño") o toda la sílaba en que se encuentra dicho fonema (se pronuncia "loj" por "reloj").

- **Inserción**: se intercala un sonido que no corresponde a esa palabra para apoyar y resolver la articulación que se dificulta (se pronuncia "Enerique" por "Enrique").

- **Distorsión**: se articula el sonido de forma incorrecta pero aproximada a la adecuada y sin llegar a ser una sustitución.

De esta manera, las formas y variedades de la sintomatología de la disartria se presentan a continuación:

- Rotacismo, disartria del fonema $/r/$ (vibrante múltiple).

- Lambdacismo, disartria del fonema $/l/$.

- Gammacismo, disartria del los fonemas guturales $/g/$, $/x/$, y $/k/$.

- Deltacismo, disartria de los fonemas dentales $/t/$ y $/d/$.
- Rinoartria, disartria de los fonemas nasales $/m/$, $/n/$, y /ñ/.
- Pararrotacismo, sustitución del fonema $/r/$ por otro como $/g/$, $/t/$, $/d/$, $/l/$, etc.
- Parasigmatismo, sustitución del fonema $/s/$ por otro como $/t/$, $/z/$, etc.
- Paralambdacismo, sustitución del fonema $/l/$ por otro.
- Paragammacismo, sustitución de los fonemas guturales por otros.
- Paradeltacismo, sustitución de los fonemas dentales por otros.

3.1.2 Prognosis

Aunque la disartria en general es una discapacidad causada por transtornos de la salud serios, en la mayoría de los casos las espectativas de mejora son amplias de acuerdo a las mismas causas:

- Las personas que presentan Esclerosis Lateral Amiotrófica (ELA, o enfermedad de Lou Gehring) pierden eventualmente la capacidad del habla.
- Son pocas las personas con Mal de Parkinson o Esclerosis Múltiple que pierden la capacidad del habla.
- La disartria causada por medicamentos o prótesis dentales mal ajustadas se pueden contrarrestar.
- La disartria causada por un accidente cerebrovascular o lesión cerebral no empeora y puede mejorar mediante terapias.
- La disartria después de una cirugía de la lengua o la laringe no empeora y puede mejorar con terapia.

El presentar disartria puede generar complicaciones adicionales como la neumonía causada por inhalación de saliva o alimento, depresión, baja autoestima, problemas sociales, por mencionar algunas.

3.2 SRAHs con Aplicación para Personas con Capacidades Diferentes

Existen tecnologías que han dado apoyo a personas con algunas discapacidades, especialmente de la voz. Sin embargo la mayoría de estas tecnologías pueden ser inaccesibles dado el costo que pueden tener. Por ejemplo, para mejorar la comunicación por voz se tienen el EchoVoice de Saltillo Corporation con un costo de $495 USD [53], o el Speech Enhancer de VoiceWave Technology Inc. con $7500 USD [56].

Es importante señalar que dichos sistemas sólo amplifican la voz pero no hacen reconocimiento. Por lo tanto, estos sistemas no implementan algún proceso como corrección fonética, o incorporación de información estadística, que mejore las anormalidades que presente la voz.

El uso de sistemas comerciales de dictado, como Dragon Naturally Speaking, Microsoft Dictation, VoicePad Platinum, e Infovox RA [15, 30, 34, 51] han mostrado niveles variables de reconocimiento en el rango del 50% al 95% para usuarios con diferentes niveles de disartria y esquemas de uso (palabras discretas o frases continuas), obteniendo los mejores desempeños cuando se usaron vocabularios pequeños (10 - 78 palabras).

Proyectos de investigación se han desarrollado en otros países para mejorar estos sistemas. En [26] se hizo uso de Redes Neuronales Artificiales (ANN's), las cuales tuvieron mejor desempeño que el sistema comercial IntroVoice. Desempeños significativos también fueron obtenidos con Modelos Ocultos de Markov (HMM's) [57]. En [24] se obtuvieron tasas de precisión en el reconocimiento de voz del 86.9% en usuarios con disartria severa y un vocabulario de 7-10 palabras para control de dispositivos electrónicos (Radio, TV, etc.).

3.2.1 Proyecto STARDUST

STARDUST (Speech Training and Recognition for Dysarthric Users of Assistive Technology) [16, 23] es un proyecto llevado a cabo en el Reino Unido enfocado al desarrollo de SRAHs como tecnología de asistencia para personas con disartria. En la Figura 3.2 se muestra el SRAH inicialmente desarrollado, el cual consiste de un programa de capacitación para ayudar a los hablantes con disartria a mejorar la coherencia de sus

vocalizaciones con un pequeño vocabulario. Esta interfaz se implementó utilizando los módulos del HTK Toolkit [61] para el reconocimiento de voz. El entrenamiento y configuración del sistema dependiente del usuario (DU) se hizo a nivel de palabras (sistema discreto) usando un vocabulario de diez palabras (en donde cada una fue repetida seis veces). Un modelo acústico se creó para cada palabra, contrario al modelado a nivel fonético usado para sistemas comerciales para usuarios sin discapacidades en el habla.

Figura 3.2: Interfaz STARDUST.

Como se muestra en la Figura 3.2, el usuario puede reproducir la palabra a través de la computadora, pronunciar la palabra (reconocimiento de voz, o ejercicio de vocalización), o pasar a la siguiente palabra. En el modo de práctica (ejercicio de vocalización) una palabra de "estímulo" aparece (Stimulus). Al pronunciar el usuario la palabra, ésta se compara con aquella que históricamente haya sido la mejor pronunciada (evaluada mediante una probabilidad de reconocimiento). Visualmente una barra muestra qué tan bien la palabra que pronunció se aproxima a la mejor, lo cual sirve de medida de referencia para que practique su pronunciación (el usuario puede tratar de hacer cada enunciado tan cerca de la palabra objetivo como sea posible). En los ensayos con 8 usuarios, todos mostraron un aumento de la precisión en el reconocimiento después de utilizar la interfaz. Se consideró que esto se debió a una mayor coherencia en la pronunciación resultado de la práctica. En general, la tasa de precisión en el reconocimiento de comandos fue del 88.5% con vocabularios de 7-13 palabras.

Del proyecto STARDUST se desarrollaron otros sistemas:

- Sistema de Control Ambiental (Environmental Control System, ECS) [24]. El usuario puede dar un comando como **subir el volumen al televisor** (TV VOLUME UP) y un sistema de control interpreta dicha orden como una señal que active el dispositivo deseado. Pruebas del ECS STARDUST con cinco personas con disartria severa mostró una precisión de reconocimiento del 86.9% en promedio, y una tasa de terminación de la tarea global de 78.6% en el uso normal en el hogar. El ECS STARDUST fue más rápido de operar que un sistema convencional con interruptores. El trabajo fue objeto de seguimiento en el proyecto VIVOCA [60].

- VIVOCA (Voice Input Voice Output Communication Aid) [22, 60]. Consiste del desarrollo de un dispositivo portable de reconocimiento y síntesis de voz para personas con voz ininteligible. Se desarrolló en manera de aplicación para PDAs, obteniendo desempeños aceptables para usuarios con disartria severa.

- STRAPTK (Speech Training Application Toolkit) [19]. El objetivo de esta interfaz fue el de proporcionar al usuario una herramienta personalizada para la mejora de su articulación a partir de estímulos audiovisuales. Integró diversas tecnologías tales como el reconocimiento de voz, herramientas de transcripción, y un sistema de gestión de bases de datos para soporte de múltiples configuraciones para el esquema de entrenamiento continuo del sistema de reconocimiento. En la Figura 3.3 se muestran algunas características de este sistema como:

  - Módulo de grabación y administración de muestras de voz para adaptación. Se muestra como una matriz de 4 por 10, en donde cada celda representa la i=1,...,10 repetición de cada j=1,...,4 palabra de estímulo.

  - Módulo de reconocimiento de voz y evaluación de la pronunciación de palabras. Incluye rutinas de ejercicios (estímulos) para la práctica y mejora en la pronunciación del usuario.

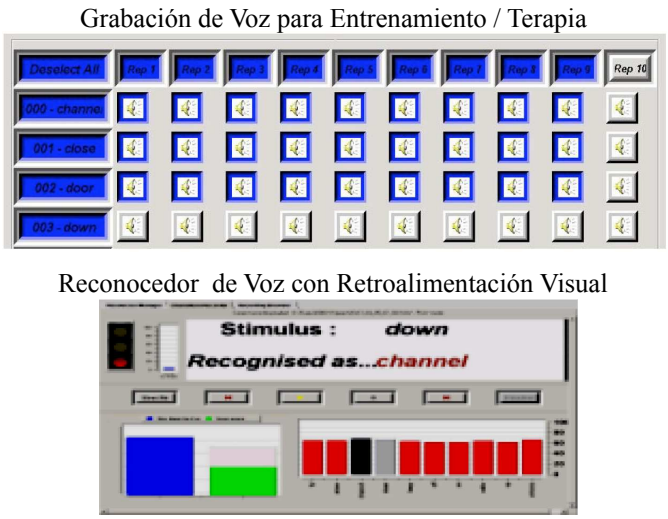


Figura 3.3: Interfaz STRAPTK.

3.2.2 CanSpeak

CanSpeak [21] es una interfaz que usa reconocimiento de voz disártrica para activación de funciones en una computadora. Fue desarrollada en Java utilizando Sphinx-4 como biblioteca para el reconocimiento de voz. Este sistema utiliza una pequeña lista de palabras clave personalizadas para cada usuario, que sean fáciles de pronunciar. Cada una de estas palabras se asocian con una letra del alfabeto, un dígito, o un comando. Por lo tanto, cada vez que se reconoce una palabra clave, el sistema envía una pulsación de tecla o un comando asociado a la aplicación que está utilizando la interfaz. La aplicación interpreta la entrada de acuerdo al contexto de interacción. El número de palabras clave depende de la aplicación y las necesidades del usuario, siendo 47 palabras las usadas para los resultados presentados en [21].

En la Figura 3.4 se muestra la interfaz de CanSpeak (lista de palabras reconocibles) integrada con otro sistema, KeySurf [20], que es una aplicación para navegación en Internet usando el teclado. La aplicación resultante, WebSpeak, es una interfaz de navegación en Internet multimodal que combina voz y teclado. Este tipo de construcción modular permite que el usuario pueda utilizar los insumos actuales.

El estudio se llevo acabo con 4 usuarios con parálisis cerebral y disartria diagnosticada (no se dan detalles del nivel de inteligibilidad). Las pruebas se hicieron de la siguiente manera: se seleccionó un vocabulario original de 47 palabras y se hizo una prueba inicial con los 4 usuarios. Subsecuentemente, a la mitad de los usuarios (Grupo

Figura 3.4: Interfaz WebSpeak integrada con CanSpeak (lista de palabras del lado izquierdo), y KeySurf integrado con un navegador de Internet (lado derecho)

1) se les dió una sesión para seleccionar un vocabulario que pudieran pronunciar mejor. A los usuarios restantes (Grupo 2) no se les dió dicha sesión y usaron el mismo vocabulario para la segunda prueba. La personalizacion de las palabras claves con la ayuda de padres y terapeutas hizo que la tasa de precision se duplicará para los usuarios del Grupo 1: 40.6%-83.3%, y 37.5%-75% para los dos usuarios. Sin embargo no ocurrio lo mismo con los usuarios del Grupo 2: 56.2%-62.5%, y 28%-34.3% respectivamente.

Es importante notar que este sistema no emplea ninguna técnica de reconocimiento de voz. El funcionamiento de CanSpeak depende de una selección previa de palabras, la cual debe hacerse en conjunto con el usuario. Para este sistema se reportó que la sesión del Grupo 1 para personalización de vocabulario tomó cuatro horas, lo cual puede ser exhaustivo para personas con disartria. De igual manera, el SRAH usado, que fue Sphinx, ya tiene sus parámetros predefinidos.

En México existen ya algunos trabajos para personas con discapacidades en el habla, los cuales se presentan en la siguiente sección.

3.2.3 Juego "Gravedad"

Este proyecto fue desarrollado en la Universidad Autonónoma de Yucatan, Merida[37], y consiste de un juego electrónico con reconocimiento de voz para estimular a niños en edad escolar con problemas del lenguaje de dislalia[3]. Se desarrolló utilizando el lenguaje C y las bibliotecas gráficas SDL. Los modelos acústicos se desarrollaron utilizando HTK Toolkit [61].

Este juego es presentado en un entorno local con palabras cotidianas, personajes y ambientes que fueran familiares para niños pequeños (ver Figura 3.5). Los personajes del juego son Rafa (toro), Tony (caballo), Lucy (niña), Miguel (niño vaquero) y Aluxe (un duende travieso).

Figura 3.5: Juego Gravedad para niños con problema de lenguaje de dislalia.

El modo de uso consiste en que el niño pronuncie palabras para cumplir con los objetivos del escenario del juego, siendo estos los siguientes:

- Escenario 1 - Recámara: el niño debe ordenar la recámara para que se le permita salir, pero el Aluxe aparece y encanta las cosas y las convierte en piedras. Una vez que se hacen parejas de los objetos y son mencionados por su nombre éstos se guardan en una caja (huacal).

- Escenario 2 - Cocina: el niño debe pronunciar de forma correcta los nombres de las cosas que el Aluxe ha encantado y suspende en el aire, si no éstas caen y ensuciarán el lugar, será reprendido por la mamá y el Aluxe habrá ganado.

- Escenario 3 - Patio: el niño tiene que ir recogiendo (nombrar) las pistas para encontrar a su mamá.

[3]Dislalia: trastorno del lenguaje, un defecto aislado de la articulación causado por un mal aprendizaje del habla. Se presenta más frecuentemente en los niños [1].

3.2.4 Sistema de Procesamiento de Fonemas para Rehabilitación de Habla

Este es un sistema desarrollado por Ma. de la Paz Copalcua-Pérez del Instituto Tecnológico de Apizaco [10] para rehabilitación de pacientes con problemas de lenguaje. Dicho sistema fue desarrollado como tema de tesis para obtener el título de Maestría en Sistemas Computacionales. El sistema se programó en Java utilizando la paquetería JavaSound para el procesamiento de la señal de voz, y constó de 4 módulos principales:

- Módulo de entrada: da acceso al paciente a cada uno de los módulos de terapia. El terapeuta puede accesar a los mismos módulos que el paciente además de entrar al módulo de entrenamiento del sistema y puede ver el historial de sus pacientes.
- Módulo de entrenamiento del sistema: el terapeuta tiene la posibilidad de entrenar al sistema con nuevos fonemas y realizar la adaptación continua.
- Módulo de terapia: el paciente recibe una terapia de tipo visual a partir de la imitación de movimientos de la boca de cada fonema. Dentro de este módulo se desarrollan 3 distintas terapias: 1) representación fonética visual, 2) entrenamiento visual, y 3) nivel de distorsión visual.
- Módulo de reconocimiento de fonemas por voz: una vez procesada la voz por el módulo de terapia, se procesa la señal para verificar si el sonido se parece al fonema que el paciente intenta aprender, de ahí la respuesta será nuevamente enviada al módulo de terapia para ver el porcentaje de reconocimiento de la señal de voz.

Este sistema es el más cercano, dentro de México, a la interfaz propuesta en este trabajo. Sin embargo su objetivo es más el de terapia que el de herramienta de comunicación. De igual manera hay muy poca información disponible acerca de su implementación o de su posterior seguimiento.

3.2.5 Interfaz para Niños con Problemas de Lenguaje

En [44] se desarrolló un juego interactivo que incorporó reconocimiento de voz para mejorar la pronunciación de niños con problemas de lenguaje. En especial, se abordó la

dislalia, cuya manera de corregirse es mediante la repetición de palabras. Para la construcción del reconocedor de voz se utilizó la biblioteca HTK Toolkit[61] y el lenguaje de programación C+. Adicionalmente se usó el CSLU Toolkit para grabación y etiquetado de voz para el corpus de entrenamiento del reconocedor. Este corpus consistió de las voces de 42 niños de entre 7 y 13 años de la ciudad de Tizimín en Yucatán. Animaciones para el juego se complementaron con las bibliotecas SDL (Simpler DirectMedia Layer). En la Figura 3.6 se muestran algunas ilustraciones del juego interactivo.

Figura 3.6: Juego para niños con problema de lenguaje de dislalia.

La dinámica del juego consiste en que dos personajes compiten por llegar a la meta, uno de los cuales es controlado por medio de la voz del niño. Las palabras que el niño debe de pronunciar corresponden a la respuesta de una pregunta que la interfaz formula. Estas preguntas son adivinanzas que previamente se han determinado dentro del juego, de tal manera que cuando el niño responde correctamente usando el micrófono su personaje avanza. En caso contrario, es el oponente quien avanza. El juego se termina cuando uno de los dos personajes llega a la meta o cuando se pronuncia las palabras "SALIR" o "TERMINAR" emitiendo un mensaje de felicitación o de ánimo para repetir el juego.

3.2.6 Sistemas Comerciales

El uso de SRAHs comerciales ha sido también utilizado para su uso con voz disártrica. Sin embargo han habido diferencias en sus desempeños. A continuación se presentan algunos casos, notando que al momento no se han encontrado estudios similares para el español mexicano:

- En [15] se usó la versión 1.01A de Dragon Dictate (inglés), el cual reconocía palabras discretas. Fue probado con 10 usuarios con alta y baja inteligibilidad (5 en cada grupo). Constante selección de vocabulario para adaptación fue realizada para mejorar el desempeño del sistema para cada usuario. Al final el sistema fue probado con el texto "Pledge of Allegiance" que consta de aproximadamente 24 palabras únicas, y fue repetido 8 veces. Los usuarios con alta inteligibilidad alcanzaron en promedio 98% de precisión al llegar a la última repetición, habiendo mejoría constante conforme se daban las repeticiones. Sin embargo, para el grupo de baja inteligibilidad el promedio fue de aproximadamente 80% al terminar la octava repetición.

- En [29] Dragon Dictate (inglés) fue usado con un usuario con disartria leve cuyo vocabulario fue menor a 70 palabras. A lo largo de diferentes sesiones de entrenamiento y prueba, el desempeño del sistema mejoró de 43% a 90% para el reconocimiento de frases.

- En [34] una versión más reciente de Dragon Dictate, Dragon Naturally Speaking (inglés), fue usada con un usuario con parálisis cerebral y disartria leve. Al igual que en los estudios anteriores, extensas sesiones de entrenamiento fueron llevadas a cabo antes de probar el sistema. 33 frases de entre 5-15 palabras, y una selección de relatos, fueron usadas para entrenamiento del SRAH, en tanto que 20 frases de 5-10 palabras fueron usadas para pruebas. En general, el SRAH mostró un desempeño de 54.17% a 82.20% sobre las 20 frases de prueba.

- En [35] una comparación de desempeño de tres SRAHs con un usuario con disartria leve fue presentada. Los sistemas (para el idioma inglés) fueron Microsoft Dictation, Dragon NaturallySpeaking (ambos reconocen habla continua), y VoicePad Platinum (que reconoce palabras discretas). En general, el usuario tuvo que leer relatos un determinado número de veces, recitando aproximadamente 4326 palabras para entrenamiento de Microsoft Dictation, 2960 para Dragon NaturallySpeaking, y 3880 para VoicePad Platinum. Los sistemas se probaron con 20 frases de 5-15 palabras (como en el caso de [34]), obteniendo un desempeño máximo de 70% con Dragon NaturallySpeaking y Microsoft Dictation (al final de 5 pruebas), y un mínimo de 45% con VoicePad Platinum.

- En [51] se usaron las versiones suizas de Dragon Dictate (reconoce habla continua, modelado a nivel fonema, sistema adaptable a usuario) e Infovox RA (reconoce palabras discretas, modelado a nivel palabra, sistema dependiente de usuario). Estos sistemas se probaron con 4 usuarios con los siguientes niveles de disartria: leve, moderado, severo, y muy severo. Para Dragon Dictate, 2 relatos de una novela fueron usados como material de estímulo para la adaptación del sistema. La prueba final consistió en una sola lectura de un pequeño fragmento de un relato. Al final de las sesiones de entrenamiento, los siguientes resultados se obtuvieron para cada usuario con el texto de prueba: 60% (disartria leve), 55% (disartria moderada), 30% (disartria severa), y 26% (disartria muy severa). Sin embargo, conforme los usuarios continuaban repitiendo el mismo texto, el desempeño del sistema iba mejorando hasta obtener los siguientes niveles: 97% (disartria leve), 97% (disartria moderada), 81% (disartria severa), y 75% (disartria muy severa).

  Infovox fue entrenado con el material disponible por el sistema, y se probó con un conjunto de 43 palabras. Los resultados de desempeño fueron los siguientes: 95% (disartria leve), 83% (disartria moderada), 74% (disartria severa), y 62% (disartria muy severa). Sin embargo, para ambos sistemas, constante supervisión por parte de terapeutas y técnicos fue requerida para re-selección de vocabulario y llevar a cabo las sesiones de entrenamiento. Estas sesiones tuvieron aproximadamente la siguiente duración para cada usuario: 2 horas (disartria leve), 4 horas (disartria moderada), 8 horas (disartria severa), y 6 horas (disartria muy severa).

Capítulo 4

Desarrollo de la Interfaz de Voz

El desarrollo de un SRAH no es una tarea fácil. Para construir un SRAH robusto usualmente se usan corpus de voz extensos. Los sistemas comerciales son entrenados con cientos o miles de muestras de voz de usuarios de diferentes géneros y edades. Estos corpora tienen costos significativos y requieren de mucho tiempo para realizarse ya que deben ser etiquetados a los niveles ortográfico y fonético. Con excepción del corpus DIMEX [47], hay pocos recursos de este tipo para el español mexicano. Adicionalmente, el hecho de crear un corpus es más demandante si se trata de usuarios con disartria, ya que para producir muestras de voz es necesario de tiempo y esfuerzo. Actualmente no hay conocimiento de algún corpus mexicano de voz disártrica para desarrollo de SRAHs.

Para la interfaz propuesta se consideró (1) recursos limitados del corpus de voz, (2) que un SRAH entrenado con voz normal se puede adaptar para un usuario con voz disártrica. Se consideró que mediante el diseño especial de un corpus textual para la producción de muestras para entrenamiento y adaptación, se puede desarrollar un SRAH robusto. También se consideró el efecto de adaptación continua, control de perplejidad y restricciones estadísticas del modelo de lenguaje, para mejorar el desempeño de este SRAH y obtener niveles de precisión similares a los de sistemas comerciales con voz normal.

En cuanto a los requerimientos de hardware y de software para la instalación y fun-

cionamiento de la interfaz se establecieron los siguientes:

- **Hardware:**
  - Procesador Pentium III/IV/Atom, o más reciente, a 1.3 GHz.
  - Tarjeta de entrada/salida de audio para micrófono y bocinas externas.
  - 512 MB o más de memoria RAM.
  - 100 MB de disco duro libre.
- **Software:**
  - Sistema Operativo Windows XP de 32 o 64 bits.
  - Voz "Isabel" de ScanSoft para Windows XP.
  - Matlab 2008.
  - Biblioteca HTK Toolkit [61] y TranscribEMex [11].

En este capítulo se presentan los principios de diseño de la interfaz, esto es, las variables consideradas para obtener el desempeño deseado con los recursos disponibles. Finalmente se presentan los detalles técnicos del diseño de la interfaz y de cada submódulo que la integra.

4.1 Definición de Variables de Control

Los componentes de un SRAH se mostraron en la Figura 2.1. Los modelos acústicos son el núcleo funcional del SRAH y son inicializados y re-estimados con la información del corpus de entrenamiento (para entrenamiento supervisado). Como se presentó en la Sección 2.1.3, se utilizaron HMM's en este proyecto para el modelado acústico. Un HMM está constituído por los parámetros $\lambda = (A, B, \pi)$, en donde B, el conjunto de probabilidades de observación, son modeladas por medio de mixturas o mezclas de gaussianas.

En tanto que es práctica común el usar tres componentes gaussianos [24], el desempeño de un SRAH está relacionado con el número de estos componentes [61]. Para reconocimiento de voz disártrica (y desempeño con pocos recursos de entrenamiento),

éste se considera un factor principal. Por lo tanto, el **número de componentes gaussianos para modelado acústico** se consideró como la **primer variable** de control de la interfaz a poder ser manipulada por el usuario.

Otro componente mostrado de un SRAH es el Modelo de Lenguaje (ML), el cual representa un conjunto de reglas o probabilidades que restringen la secuencia de palabras reconocidas por el SRAH a secuencias más válidas. Comúnmente se usan N-gramas como ML, siendo para este trabajo, bigramas (N=2) el ML usado para reconocimiento de habla continua [27, 61].

Para el ML existen dos métricas para medir su desempeño: (1) Tasa de Error de Palabras (Word Error Rate, WER), y (2) la perplejidad. Como se presentó en la Sección 2.2.5, WER es dependiente del SRAH, y es estimada por la secuencia de palabras generada por el sistema. En algunos casos, una baja WER se correlaciona con una baja perplejidad de ML [57]. Para reconocimiento de voz disártrica, se recomienda baja perplejidad para lidiar con el efecto de la lenta articulación de fonemas [57].

La perplejidad no depende del SRAH, por lo que puede ser estimada más rápido que el WER [7]. La perplejidad aumenta cuando el vocabulario crece en tamaño, y el uso de N-gramas reduce la perplejidad para vocabularios extensos al restringir las secuencias reconocidas a secuencias más probables. Sin embargo, para lograr esto, es necesario que el vocabulario de uso sea conocido por adelantado por el SRAH [7].

Para lidiar con esta situación se consideró el construir el ML en tiempo de ejecución mientras se usa la interfaz de voz. Mediante esta actualización constante del ML se permite el conocimiento previo del vocabulario para reducir la perplejidad. Por lo tanto, el **vocabulario** y el **ML** se consideraron como la **segunda variable** de control a ser manipulada por el usuario.

Adicionalmente, una **tercer variable** fue considerada, el **factor de escala gramática** del ML [61]. Este factor regula la presencia que tiene el ML sobre la señal acústica al momento del reconocimiento de la misma. Cuando este factor aumenta, el SRAH le da más importancia al ML sobre la señal de voz para predecir lo que dijo el usuario (p.e., las restricciones del ML tienen más importancia). Por lo tanto, el factor de gramática se puede usar también para reducir la perplejidad del ML durante el reconocimiento de voz.

4.2 Corpus de Entrenamiento

La construcción del núcleo del reconocedor (HMM's) con pocos recursos acústicos, y que sea robusto, se puede lograr si se tienen suficientes muestras de voz de fonemas del lenguaje. Esto incluso si sólo un usuario se considera como fuente del corpus oral [4]. Como se presentó en la Figura 2.2, la selección de un texto representativo es el paso inicial para obtener las muestras de voz del corpus, y por lo tanto, el más importante para obtener la diversificación de fonemas [4].

Para esta interfaz, el texto representativo del corpus de entrenamiento se obtuvo de las siguientes fuentes:

- 49 palabras diferentes usadas para evaluar el nivel de disartria de un paciente mexicano (ver Tabla A.1, Anexo A). Estas palabras contienen los fonemas del español mexicano mostrados en la Tabla 2.1. Esta selección de palabras fue proporcionada por los terapeutas de lenguaje del Sistema Nacional para el Desarrollo Integral de la Familia (SNDIF) de la ciudad de Huajuapan de León, Oaxaca.

- Un fragmento del relato "Fiesta en la Montaña" (ver Tabla A.2, Anexo A) [38] que se encuentra fonéticamente balanceada y que consistió de 102 palabras diferentes.

- 16 frases fonéticamente balanceadas (ver Tabla A.3, Anexo A). Estas frases fueron diseñadas para que pudieran usarse como estímulo para obtener muestras de voz para adaptación de nuevo usuario.

En total, el texto representativo para el corpus constó de 205 palabras únicas. Como se presentó en la Sección 2.2.1, las secuencias de fonemas que definen cada palabra se obtuvieron con TranscribEMex [11, 47]. Basados en los resultados obtenidos en [23], en donde un mínimo de 6 muestras de voz fueron necesarias para obtener precisiones cercanas al 100% para el reconocimiento de comandos, se consideró que este corpus textual estaba balanceado para proveer suficientes muestras de fonemas. En la Figura 4.1 se muestran las ocurrencias de cada fonema del español mexicano en el texto representativo. Note que el fonema con menor número de ocurrencias es $/\_G/$ con 6. En la Figura 4.2 se muestran las ocurrencias de fonemas en la sección del corpus textual correspondiente a las 16 frases de adaptación. Ambas distribuciones se correlacionan con un coeficiente de 0.62.

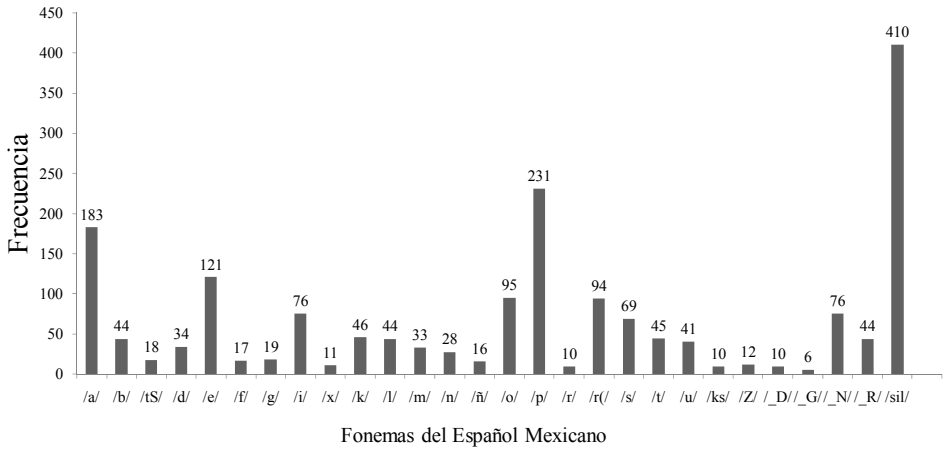


Figura 4.1: Frecuencia de fonemas en el Texto Representativo

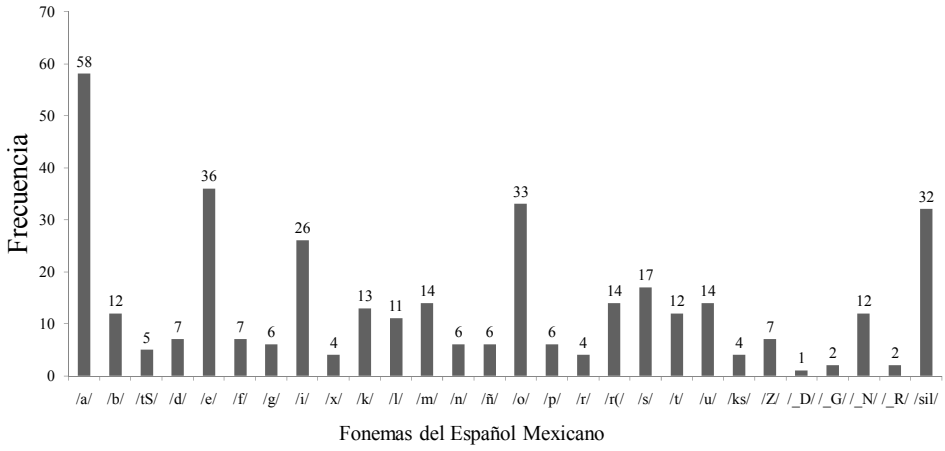


Figura 4.2: Frecuencia de fonemas en el estímulo para adaptación

Para obtener el corpus oral el texto representativo fue leído 5 veces por dos usuarios: (1) un usuario de referencia con voz normal, (2) un usuario con nivel bajo-medio de disartria (ver Sección 5.1.1, Tabla 5.3, usuario GJ). Por lo tanto, dos corpus de entrenamiento fueron desarrollados, uno con voz normal y otro con voz disártrica. Esto se realizó para evaluar el desempeño de dos metodologías de desarrollo de SRAHs (ver Sección 2.1):

- uso de un SRAH dependiente de usuario (DU) (entrenado con la voz del usuario con disartria que usará el sistema) como en [19, 23, 24, 26, 36];
- uso de un SRAH independiente de usuario (IU) (entrenado con la voz de un

usuario con voz normal) adaptado a la voz del usuario con disartria que usará el sistema (como en el caso de sistemas comerciales que fueron aplicados en [15, 34, 51, 57]).

Las lecturas fueron grabadas con un equipo Sony lcd-bx800 con frecuencia de muestreo de 8 kHz monoaural en formato WAV. Posteriormente esta información se etiquetó manualmente al nivel ortográfico (palabra) y fonético usando WaveSurfer y TranscribEMex como se describió en la Sección 2.2.1. Finalmente los archivos de audio se codificaron en MFCCs (ver Sección 2.2.1). La interfaz por lo tanto se inicializa con los siguientes recursos:

- Muestras de voz codificadas en formato MFCC para entrenamiento supervisado de HMM's usando HTK.
- Etiquetas fonéticas y ortográficas correspondientes a las muestras de voz.
- Modelo de Lenguaje (ML) Inicial, estimado a partir de las etiquetas ortográficas del corpus. Una vez que inicie el uso del módulo de "Reconocimiento" y se lleve a cabo la adaptación dinámica del SRAH este componente cambiará.
- Diccionario fonético construído usando TranscribEMex para las palabras de este corpus. Una vez que se comience a usar el módulo de "Reconocimiento" y se lleve a cabo la adaptación dinámica del SRAH este componente cambiará.
- Lista de fonemas del corpus (esta nunca cambiará a lo largo del uso del SRAH).

A continuación se presentan los detalles de diseño de la interfaz de voz.

4.3 Módulos de la Interfaz

En la Figura 4.3 se presenta la pantalla inicial de la interfaz, la cual consiste de tres módulos principales:

- **Creación y Adaptación del Reconocedor de Voz**: crea el SRH IU habilitando al usuario a especificar el número de componentes gaussianos del mismo, lo cual repercute en el desempeño del mismo. También proporciona una rutina de grabación de frases para adaptación del mismo a otro usuario.

- **Uso del Reconocedor de Voz**: permite al usuario usar el reconocedor de voz especificando los parámetros del mismo que mejor se adapten a sus necesidades: vocabulario, factor de gramática. Incluye la función de adaptación continua.

- **Patrones de Confusión Fonética**: presenta una matriz de confusión fonética correspondiente al usuario para identificar confusiones significativas en el reconocimiento.

Figura 4.3: Pantalla Principal de la Interfaz de Voz

El lenguaje de programación fue Matlab 2008 con el toolkit *GUIDE*, y el diseño se basó en los Principios de Diseño de Interfaces de Ben Shneiderman [55]. En este proyecto se consideró a Usuarios Novatos dado el perfil tecnológico del usuario y de sus familiares, que comúnmente son de pocos recursos económicos y baja escolaridad. Se asume que estos usuarios conocen muy poco de las tareas o los conceptos de la interfaz. Las recomendaciones para este tipo de usuario fueron consideradas:

- no se implementaron sub-menús o sub-ventanas que haya que configurar cada vez que se accede a la interfaz. El texto que se utiliza se muestra en pantalla en todo momento, y el cambio que se realiza en los documentos de un usuario se actualizan para todos los módulos de la interfaz;

- se implementaron mensajes de retroalimentación cuando un dato o selección no es válida;
- el rango de tareas se muestra en pantalla y pueden llevarse a cabo con sólo un click, son sencillas y todo el proceso de adaptación o de reconocimiento se hace de manera automático. Se hizo amplio uso de botones;
- la interfaz usa de manera alternada la Manipulación Directa (uso de apuntadores y botones), Selección de Menú (lista desplegable para seleccionar usuario o ver frases), y Llenado de Formas (para añadir vocabulario).

4.3.1 Adaptación de Usuario

Los sistemas comerciales se entrenan con las muestras de cientos o miles de hablantes diferentes. Cuando un nuevo usuario quiere usar dicho sistema, es común el preguntar a este usuario el que lea algunas palabras o textos (estímulo) para proveer muestras de voz al sistema. Esto para adaptar sus modelos acústicos a los patrones de voz del nuevo usuario.

Para este proyecto, MLLR [31] es la técnica de adaptación para hacer a un SRAH adaptable y usable por otros usuarios con voz normal y/o disártrica. Como se presentó en la Sección 2.1.5, MLLR se basa en la creación de un conjunto de transformaciones lineales que, aplicadas sobre los parámetros de los componentes gaussianos de los HMM del SRAH (media y varianza), puede reducir la diferencia entre estos HMM y los datos de adaptación. Un árbol de regresión con 32 nodos terminales fue utilizado para la implementación dinámica de MLLR (ver Sección 2.1.5).

16 frases fonéticamente balanceadas (ver Tabla A.3, Figura 4.2) fueron diseñadas para la primera adaptación del SRAH, la cual se definió como **Estática**. Esto porque sólo se realiza una sola vez en este módulo previo a utilizar el SRAH por un nuevo usuario. En la Figura 4.4 se presenta la pantalla del módulo de **Creación y Adaptación del Reconocedor de Voz**. En tanto, en la Figura 4.5 se muestra el flujo de operaciones realizadas por el código de programación realizado para este módulo, incluyendo el uso y configuración de las bibliotecas asociadas de HTK (ver Sección 2.2.2).

El primer panel, **Reconocedor Base** construye los HMM's del SRAH con diferentes componentes gaussianos (manipulación de la primer variable de control). Esto permite

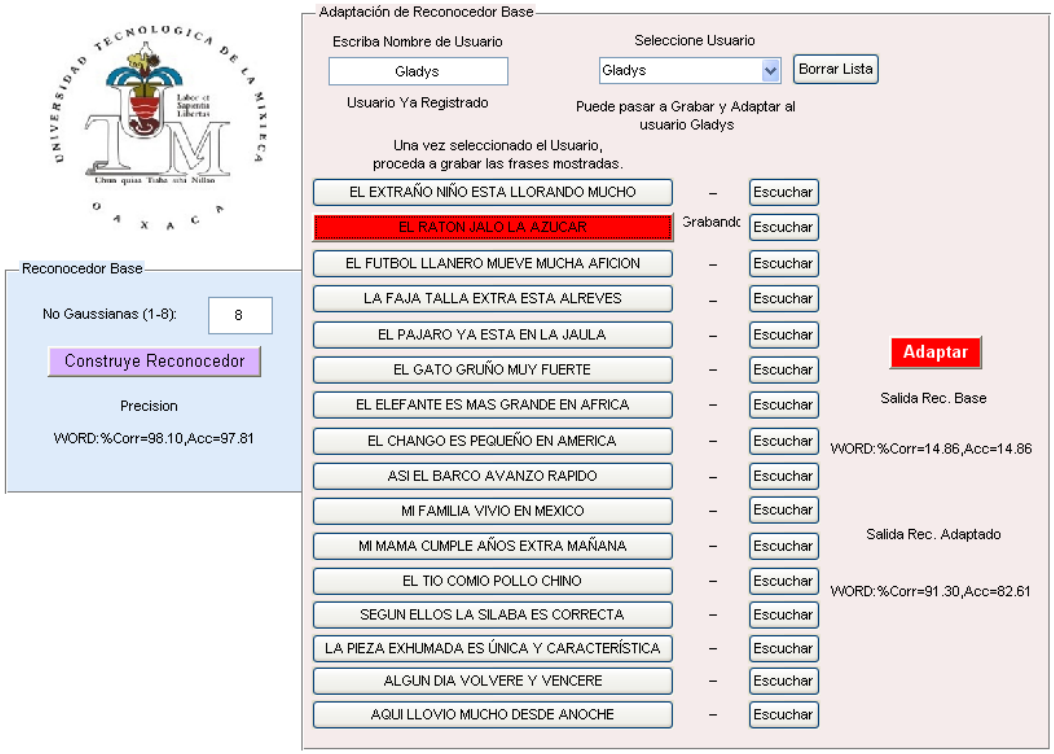


Figura 4.4: Interfaz del Módulo de Creación y Adaptación del Reconocedor de Voz.

al usuario crear el reconocedor en un solo paso y evaluar el desempeño del mismo conforme aumentan los componentes gaussianos de los HMM's. Los resultados de el reconocedor construído sobre el Corpus de Entrenamiento se muestran una vez que el proceso ha finalizado. Para generar estos resultados se utiliza el Modelo de Lenguaje estimado a partir del Corpus de Entrenamiento al igual que el diccionario. Como se muestra en la Tabla 4.1, el incremento en precisión (ver Ec. 2.6) es significativo conforme se aumentan los componentes.

Tabla 4.1: %WAcc del SRAH base entrenado con voz normal y con número variable de componentes gaussianos para el modelado acústico.

	No. de Componentes Gaussianos							
	1	2	3	4	5	6	7	8
% WAcc	93.02	94.92	97.39	97.25	97.81	98.45	97.88	97.81

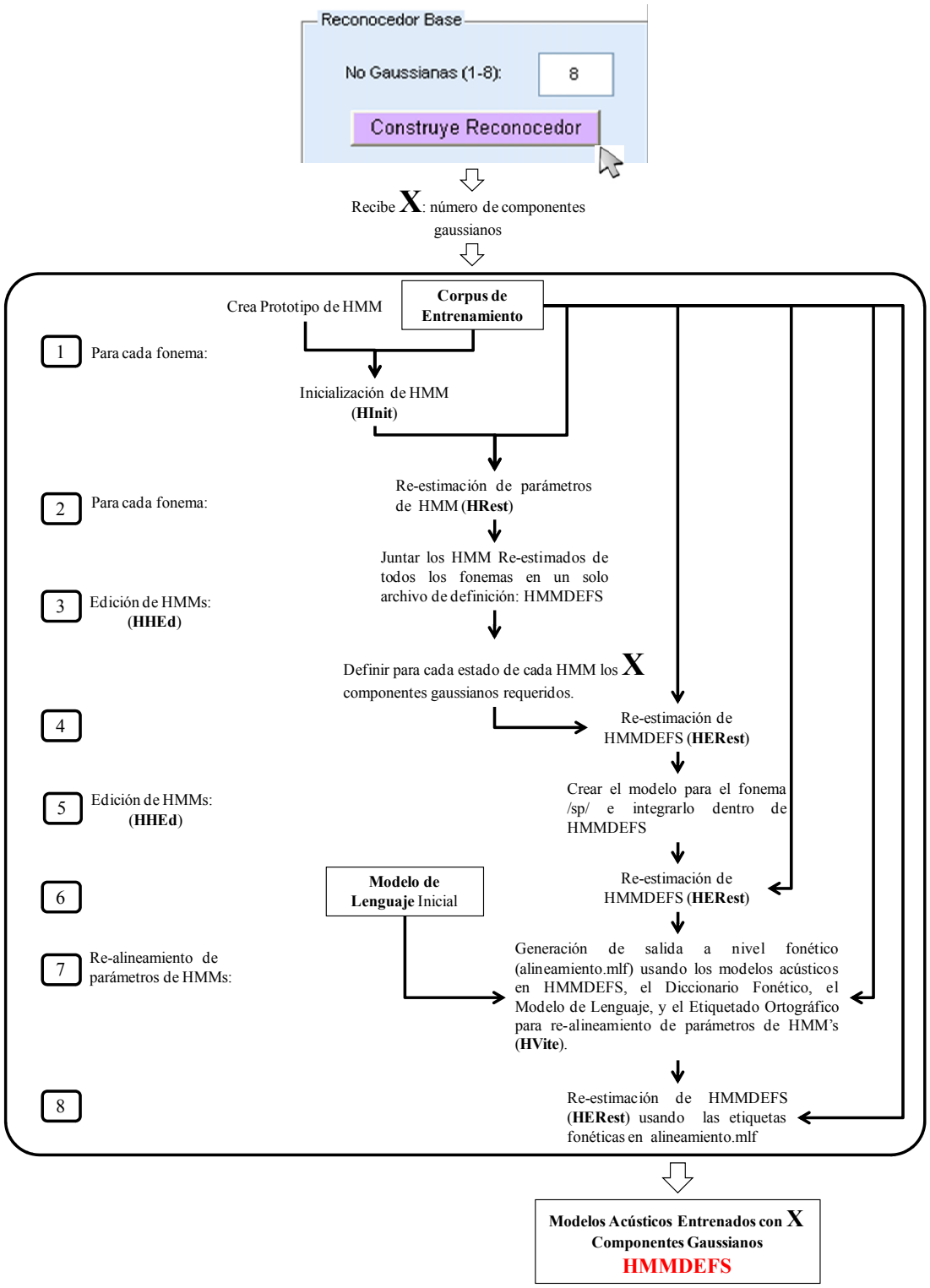


Figura 4.5: Flujo de operaciones internas del módulo de Creación del Reconocedor de Voz.

Una vez que los Modelos Acústicos (o HMM's) del SRAH se han construído, el usuario puede acceder al segundo panel, **Adaptación de Reconocedor Base**, para poder adaptar el sistema de manera **Estática** a su voz. Si se trata de un usuario nuevo se le pide que ingrese su nombre en la forma "Escriba Nombre de Usuario". Un ejemplo del nombre se muestra en el campo correspondiente. Una vez que el nombre se escribe y se presiona "Enter" éste se guarda automáticamente y se actualiza en la lista de usuarios del menú desplegable "Seleccione Usuario". Al seleccionar un usuario de esta lista se crean automáticamente los directorios y archivos correspondientes para inicializar la adaptación. Un mensaje explicativo de hacia dónde seguir se muestra abajo del menú de "Seleccione Usuario" una vez que se escogió al usuario. Si se trata de un usuario ya existente que quiere re-adaptar el SRAH no es necesario darse de alta (registro) en la primer forma, por lo tanto puede ir directamente a "Seleccione Usuario".

Para comenzar con la adaptación estática el usuario puede grabar en cualquier orden las frases del Corpus Textual de Adaptación (ver Tabla A.3) al dar click sobre el botón que representa la frase. Los principios de Diseño Centrado en el Usuario(UCD) [32] y Diseño de Interfaces se consideraron para estas operaciones de la siguiente manera:

- Inicialmente se tenía un botón para grabar, el cual tenía un tiempo pre-asignado de 5 segundos de duración. En la práctica esto fue muy problemático ya que un usuario con disartria toma mayor tiempo para articular una palabra. De igual manera no había un indicador del estado del proceso (por ejemplo, no se sabía cuándo iniciaba el programa a grabar o cuánto tiempo quedaba). Por lo tanto, en la interfaz actualizada cuando se presiona el botón de la frase a grabar, éste se ilumina de color rojo y junto al botón aparece la leyenda "Grabando". La grabación se detiene en el momento que el usuario vuelva a presionar el botón, regresando a su color original y desapareciendo la leyenda "Grabando". Para verificar que la muestra se ha grabado correctamente el usuario puede presionar el botón "Escuchar" para reproducir la grabación.

- El usuario puede re-grabar en cualquier momento cualquier frase y re-adaptar.

Al final de este proceso de grabación sólo es necesario que el usuario presione el botón de "Adaptar", el cual comienza la gestión de los módulos de HTK y de los componentes del SRH IU para crear los modelos adaptados y los folders correspondientes

para guardar la información del usuario (incluyendo las frases grabadas, que forman parte del Corpus de Adaptación personalizado del usuario). Los resultados del sistema original y del adaptado sobre el Corpus de Adaptación se muestra para propósitos de comparación de desempeño (como se observa en la Figura 4.4, el nuevo usuario obtiene un desempeño del SRH mucho menor cuando no está adaptado). En la Figura 4.6 se muestra el flujo de operaciones correspondientes a este módulo de la interfaz de voz.

Note que este tipo de adaptación es realizada una sola vez antes de que el nuevo hablante use el sistema (por eso se denomina estática). En los sistemas comerciales, si el usuario necesita mejorar la adaptación, es necesario que lea otros textos de estímulo. Para este sistema se incorporó esta tarea dentro del uso propio del SRAH. De tal manera que la adaptación puede llevarse a cabo mientras se realiza el reconocimiento de voz (definiéndola como adaptación dinámica). En este caso, se permite al usuario ingresar cualquier palabra o texto y realizar cualquiera de las siguientes acciones:

- añadir esta palabra o texto al Modelo de Lenguaje (ML) inicial del SRAH (y por lo tanto, reducir la perplejidad del ML);
- leer esta palabra o texto como estímulo para realizar adaptación. En este caso, además de que la palabra o texto es añadida al ML del SRAH, las nuevas muestras de voz se almacenan en los directorios personales del usuario dentro del SRAH. Entonces, la re-estimación de las transformaciones es llevada a cabo considerando todas las muestras grabadas del usuario (aquellas de la adaptación estática y aquellas grabadas mientras se usa el SRAH). Por lo tanto, la adaptación es dinámica y acumulativa.

4.3.2 Reconocimiento de Voz

Una vez que se ha construído los modelos acústicos del SRAH (HMMDEFS) y se ha adaptado para un nuevo usuario, éste ya puede comenzar a usarlo. Para ello accede al segundo módulo de la interfaz, **Reconocedor de Voz**, el cual se presenta en la Figura 4.7.

Esta interfaz comienza con el menú de "Seleccione Usuario" en donde al seleccionar su nombre la interfaz automáticamente cargará sus registros personales (HMMDEFS+Transformaciones Lineales MLLR). Adicionalmente se incluye el botón de

Figura 4.6: Flujo de operaciones internas del módulo de Adaptación Estática del Reconocedor de Voz.

Figura 4.7: Interfaz del Módulo de Reconocimiento de Voz para Comunicación.

"Crea Modelo de Lenguaje", el cual está estrechamente vinculado con el vocabulario pre-cargado que se muestra en el listado de "Frases de Vocabulario". Este listado tiene el objetivo de ser informativo acerca de las palabras que puede reconocer el sistema, las cuales corresponden al Corpus Textual de Entrenamiento y de Prueba. Esta interfaz permite al usuario añadir vocabulario mediante la forma de "Añadir Nuevas Frases o Palabras". Cualquier frase o palabra escrita en este campo se actualiza y guarda en la lista de "Frases de Vocabulario". Al presionar el botón de "Crea Modelo de Lenguaje" la interfaz automáticamente actualiza el diccionario fonético (mediante TranscribEMex) y el modelo de lenguaje del SRH. Mediante estas funciones se manipulan las segundas variables de control del SRAH, logrando el control de la perplejidad del ML.

La tercer variable de control, el factor de escala gramático, se integra para ajustar la respuesta SRAH a un nivel deseado. Es por esto que se asignó a la forma "Valor de Ajuste (1-30)" con un rango recomendado. Al ingresar el número (valor $-s$) se tiene el reconocedor configurado para su uso. Estos valores (modelo de lenguaje, factor de escala) se pueden cambiar en cualquier momento sin necesidad de re-iniciar la interfaz.

Para comenzar a reconocer la voz el usuario debe presionar el botón "Ejecuta Reconocedor" el cual cambiará a color rojo cuando esté listo para recibir voz. Cuando el usuario termine de hablar puede presionar de nuevo el botón el cual cambiará de nuevo a blanco. En ese momento se ejecutan las bibliotecas configuradas de HTK (HCopy, HVite) con los archivos generados y se proporcionan dos salidas: la original y la adaptada. La forma de onda de la voz se ilustra para fines de retroalimentación en el centro de la interfaz. El sintetizador de voz lee la frase obtenida con el SRAH adaptado de una manera más entendible. Para este propósito se usó el *Speech Application Programming*

Interface (SAPI) ver 5.0 de Windows XP como sistema de síntesis de voz, siendo *Isabel* de ScanSoft la voz para la articulación de español. El enlace de la interfaz de voz con el SAPI se implementó mediante la función **tts.m** de Siyi Deng (2007).

Otra función de esta interfaz es el permitir la adaptación dinámica del SRAH. Esto ha sido de beneficio en sistemas para usuarios con disartria avanzada [19] ya que permite el continuo modelado de las deficiencias de articulación del usuario. Adicionalmente se integra como medio de reducir la perplejidad del ML y reducir la tasa de error en el reconocimiento de voz.

Esto se realiza como opción adicional al ingreso de nuevo vocabulario en "Añadir Nuevas Frases o Palabras". Cualquier frase que se ingrese en este campo es candidata a ser grabada y ser añadida a las frases de adaptación guardadas del usuario. Si el usuario desea añadir frases para adaptación, debe presionar el botón que se encuentra abajo del campo de añadir texto, "Grabar para Adaptación" el cual graba la voz como en los casos anteriores. Internamente esta grabación queda asociada a la frase presente en el campo de "Añadir Nuevas Frases o Palabras". El límite en cuanto a las frases a ingresar o grabar, está en la capacidad de almacenaje del dispositivo computacional.

En este punto sólo estas frases de adaptación están grabadas. Para implementar la adaptación es necesario presionar el botón que está junto, "OK" el cual genera las transcripciones fonéticas correspondientes a dichas frases para realizar la re-adaptación de los modelos acústicos del usuario. Las frases grabadas en la sesión se añaden a las previamente guardadas por lo que el proceso de adaptación es acumulativo. Igualmente, al re-adaptar los modelos acústicos, actualiza el vocabulario, el diccionario, y el modelo de lenguaje del SRH. En la Figura 4.8 se muestra el esquema de operaciones correspondientes a este módulo de la interfaz de voz.

4.3.3 Patrones de Confusión Fonética

El tercer módulo de la interfaz provee de información de los patrones de confusión en la articulación de fonemas del usuario. Esta información se presenta visualmente en la forma de una matriz de confusión fonética como se presenta en la Figura 4.9. Esta puede ser usada por el terapeuta para detectar anormalidades significativas en la voz del usuario, y para definir actividades terapeuticas más específicas. También, puede ser

1 Carga archivos globales del SRAH:

- LISTA_PALABRAS.txt : lista con todo el vocabulario del SRAH.
- DICT.txt : diccionario fonético de todas las palabras del vocabulario.
- ML: Modelo de lenguaje estimado a partir del etiquetado ortográfico del corpus de entrenamiento
- PALABRAS.mlf : etiquetas ortográficas del corpus de entrenamiento.
- LISTA_FRASES.txt : lista con las frases de PALABRAS.mlf

Mostrar LISTA_FRASES.txt en "Frases Vocabulario"

2 Selecciona Usuario **Y** ⟶ Carga directorios y archivos personales:

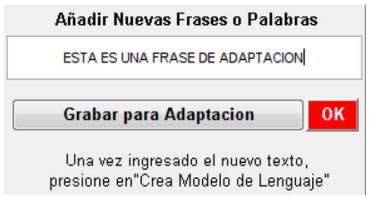
- adapt_Y (transformaciones lineales), adapt_mfc_Y (archivos de audio MFCC para adaptación), classes_Y (modelos acústicos del sistema HMMDEFS y árbol de regresión para adaptación)
- adapt_Y.scp (lista de archivos de audio MFCC para adaptación), adapt_fonemas_Y.mlf (etiquetas fonéticas de las frases usadas para adaptación).

3 Seleccionar Valor de Ajuste (Factor de Escala Gramática): **Z**

4 **Función de Reconocimiento:**

- Captura de Audio: [Ejecuta Reconocedor] **record, getaudiodata, wavwrite**
- Grafica forma de onda: **plot**
- Codifica archivo de audio WAV en MFCC (**HCopy**)
- Ejecuta **HVite** con ML, DICT.txt, **HMMDEFS** (para Salida Original), **HMMDEFS + Transformaciones Lineales MLLR** (para Salida Adaptada), y **Z**.
- Extrae frase reconocida como secuencia de caracteres y despliega en pantalla.
- Implementa **Síntesis de Voz** con la función **tts** usando la secuencia de caracteres correspondiente a la salida adaptada.

5 **Función de Adaptación Dinámica e Incremento de Vocabulario:**



- Recibe texto **W** en el campo "Añadir Nuevas Frases o Palabras": **get (hObject, 'string')**
- Escribe **W** en LISTA_FRASES.txt.
- Convierte **W** a formato de etiqueta ortográfica y añade a PALABRAS.mlf

Para Incremento de Vocabulario: [Crea Modelo de Lenguaje]

- Extrae y ordena alfabéticamente las palabras del archivo de etiquetas ortográficas PALABRAS.mlf.
- Actualiza LISTA_PALABRAS.txt
- Ejecutar **TranscribeMex** para extraer las secuencias fonéticas de las palabras en LISTA_PALABRAS.txt.
- Convertir la salida de TranscribeMex para integrarse con las palabras de LISTA_PALABRAS.txt para formar el diccionario fonético actualizado DICT.txt
- Ejecutar **HLStats** y **HBuild** con PALABRAS.mlf y LISTA_PALABRAS.txt para generar un ML actualizado con el nuevo texto **W**.

Para Adaptación Dinámica:

- Captura de Audio de la lectura del texto **W** [Grabar para Adaptacion] [OK] **record, getaudiodata, wavwrite**
- Codifica archivo de audio WAV en MFCC (**HCopy**) y guarda en adapt_mfc_Y
- Asignar un identificador al archivo MFCC (p.e., adapt_Y_17.mfc, en donde 17 es el número de frase de adaptación, el cual se incrementará conforme se añadan más frases.).
- Convierte **W** a formato de etiqueta fonética con **TranscribeMex** y añade a adapt_fonemas_Y.mlf . Esta etiqueta se nombra igual que el archivo MFCC que le corresponde (p.e., adapt_Y_17.phn).
- Actualiza adapt_Y.scp con la ruta y nombre del nuevo archivo de audio MFCC para adaptación.
- [OK] Ejecuta adaptación MLLR con **HERest** para el usuario Y con los siguientes registros actualizados: adapt_mfc_Y, adapt_fonemas_Y.mlf, adapt_Y.scp. Las nuevas transformaciones lineales MLLR se almacenarán en adapt_Y, reemplazando las anteriores.
- Se ejecuta actualización del vocabulario, ML, y diccionario: [Crea Modelo de Lenguaje]

Figura 4.8: Flujo de operaciones internas del módulo de Reconocimiento de Voz.

utilizada para medir el nivel de disartria de un usuario (ver Section 5.1.1).

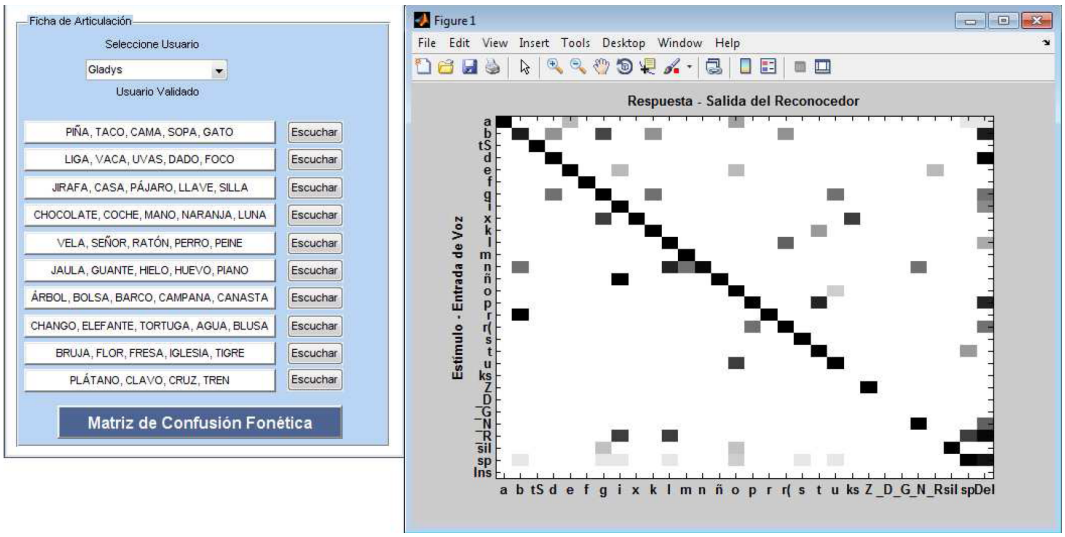


Figura 4.9: Interfaz del Módulo de Patrones de Confusión.

En el panel "Ficha de Articulación" el usuario selecciona su nombre del menú en "Seleccione Usuario". Al hacer esto la interfaz carga los modelos acústicos adaptados del usuario (HMMDEFS+Transformaciones Lineales MLLR). Una vez hecho esto, el usuario debe leer las secuencias de palabras mostradas en cada boton del panel (esta rutina es la misma que para la adaptación estática de la Sección 4.3.1).

Note que las palabras mostradas en la Figura 4.9 corresponden a la lista de 49 palabras usadas por los terapeutas para medir el nivel de disartria de un usuario cuyo idioma materno es el español mexicano (ver Tabla A.1). Después de que todas las muestras son grabadas el usuario solo necesita presionar el botón "Matriz de Confusión Fonética" para estimar los patrones de confusión sobre estos datos de diagnóstico.

La confiabilidad de esta herramienta para diagnóstico depende de la clasificación de fonemas entre el estímulo y la salida del reconocedor. Comunmente, la clasificación se realiza mediante análisis de percepción o mediante herramientas de alineamiento temporal. En la Tabla 4.2 se muestra un ejemplo de alineamiento temporal y clasificación de fonemas.

HTK puede estimar una matriz de confusión fonética del alineamiento de la transcripción fonética de la frase hablada (P), y de la salida fonética del SRAH ($\hat{P}$). El uso de herramientas de programación dinámica (PD) para alinear dos secuencias de car-

Tabla 4.2: Estimación de matriz fonética a partir de alineamiento de secuencias fonéticas.

									Eliminaciones	Inserciones	Sustituciones	Correctos
BRUJA												
P:	b	r(	u	x	a				r(→ ' '		u → o	b → b, x → x, a → a
$\widehat{P}$:	b		o	x	a							
IGLESIA												
P:	i	g	l	e	s	i	a		g → ' ', l → ' '			i → i, e → e, s → s
$\widehat{P}$:	i			e	s	i	a					i → i, a → a
CANASTA												
P:	k	a	n	a	s	t	a		k → ' ', s → ' ' ' ' → a			a → a, n → n, a → a
$\widehat{P}$:		a	n	a		t	a	a				t → t, a → a

Matriz de Confusión Fonética estimada del alineamiento de $P, \hat{P}$

	a	b	e	g	i	x	k	l	n	o	r(	s	t	u	Eliminación
a	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
e	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
i	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
x	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
n	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r(	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
s	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
t	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
u	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Inserción	1	0	0	0	0	0	0	0	0	0	0	0	0	0	

acteres (en este caso, secuencias de fonemas) puede dar resultados poco satisfactorios cuando un alineamiento en particular entre P y $\hat{P}$ es requerido. Esto es porque estas herramientas usualmente usan ponderaciones (medidas de distancias) que son "0" si un par de fonemas son iguales, y "1" en otro caso.

En el caso de HResults de HTK, tal alineamiento puede ser generado. Sin embargo, las ponderaciones son derivadas de manera empírica, asignando a un emparejamiento perfecto una ponderación de "0", a una inserción o eliminación una ponderación de "7", y a una sustitución una ponderación de "10" [61]. Aunque este alineamiento es más efectivo que los alineadores que usan ponderaciones de '1" and "0", éste puede ser mejorado como se presentó en [3], en donde las ponderaciones se basaron en la similitud acústica entre fonemas. En tanto que en [3] el idioma de uso fue el inglés británico, en este caso se hará para el español mexicano. Las ponderaciones correspondientes a este idioma para el alineador (clasificador de fonemas) se estimaron mediante:

$$Sim(p^j, \hat{p}^i) = 8Pr_{SI}(q^j, \hat{q}^i) - 3 \tag{4.1}$$

En la Ec. 4.1, $Sim(p^j, \hat{p}^i)$ es la matriz de ponderaciones de similitud entre pares de fonemas $\{p^j, \hat{p}^i\}$ del alineamiento de secuencias P y $\hat{P}$. En [3] estas ponderaciones se estimaron mediante la asignación de escalas a la matriz de confusión normalizada obtenida con un SRAH IU, $Pr_{SI}(q^j, \hat{q}^i)$, estimada de 92 hablantes del idioma inglés, en donde q^j y $\hat{q}^i$ son los elementos respectivos de la secuencia de fonemas $\{Q, \hat{Q}\}$ alineadas de los datos correspondientes. En este caso, $Pr_{SI}(q^j, \hat{q}^i))$ se obtuvo de la matriz de confusión fonética generada por HTK sobre el corpus de entrenamiento. Por lo tanto, un correcto alineamiento recibe la ponderación máxima, "+5" si la probabilidad de confusión en Pr_{SI} es alta (p.e., ≥ 0.95), con fonemas con muy poca confusión recibiendo la ponderación mínima, "-3" (p.e., < 0.05). La matriz de ponderaciones se presenta en la Tabla B.2.

Después de que las ponderaciones de similitud son calculadas, estas se integraron en un algoritmo de programación dinámica para realizar el alineamiento y clasificación de fonemas. El algoritmo PD usado en este módulo es una variante del algoritmo de Alineamiento Dinámico en el Tiempo (Dynamic Time Warping, DTW) simétrico presentado en [5]. Las secuencias de fonemas $\hat{P}$ fueron obtenidas de la ejecución del SRAH con un ML de bigramas a nivel fonema. Los resultados se discuten en la Sección 5.1.1. En la Figura 4.10 se muestra el esquema de operaciones correspondientes a este módulo de la interfaz de voz. El Pseudo-código del alineador fonético se presenta en el Anexo B.

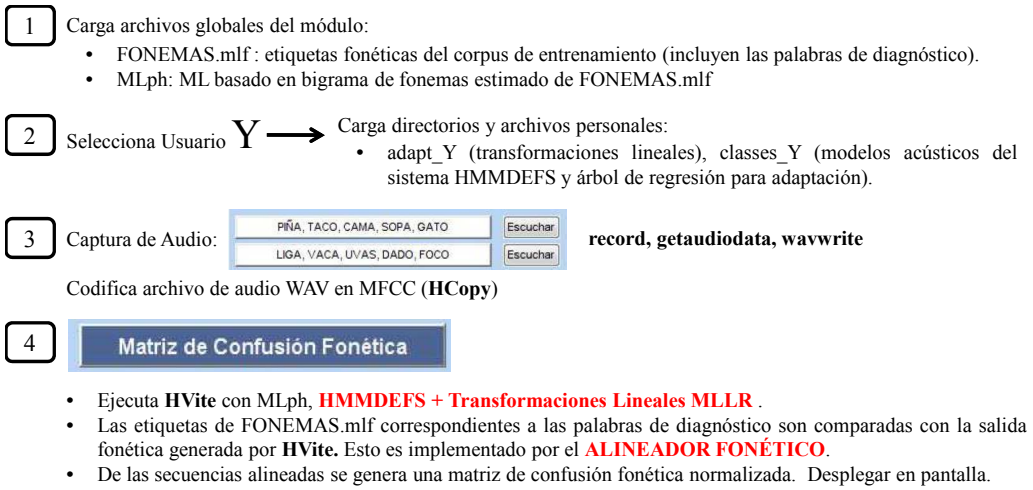


Figura 4.10: Flujo de operaciones internas del módulo de Patrones de Confusión Fonética.

Capítulo 5

Presentación de Resultados

En el capítulo 3 se presentó una reseña de proyectos de desarrollo de SRAHs para personas con discapacidad en el habla. Algunos de ellos, como en [15, 29, 34, 35], usaron software comercial para dictado con personas con diferentes niveles de disartria, logrando tasas de precisión de hasta 98%. Sin embargo, extensas sesiones de entrenamiento y continua asistencia de otra persona fue necesaria para mejorar y operar el sistema. Incluso después de esta preparación, para varios usuarios no fueron obtenidas tasas de reconocimiento similares [15, 51]. El factor de la variabilidad en la producción de voz fue evidente en estos estudios, incluso después de varias sesiones de entrenamiento.

Sistemas desarrollados especialmente para usuarios con disartria, como STARDUST [16, 23, 24], STRAPTK [19] y CanSpeak [21], han mostrado niveles de precisión cercanos al 85% sin sesiones de entrenamiento tan extensas. Sin embargo, en general, las evaluaciones de estos sistemas se hacen con vocabularios pequeños ($<$ 100 palabras), restringiendo sus aplicaciones al uso de palabras clave o comandos. Adicionalmente, la mayoría de estos proyectos se realizan para personas cuya lengua materna es el Inglés.

Es cuanto a las pocas aplicaciones encontradas en México para el objetivo de esta tesis, los SRAHs para el español mexicano fueron diseñados para transtornos del habla diferentes a la disartria como en [37] y [44]. De igual manera el vocabulario fue pequeño, no permitiendo agregar nuevas palabras, siendo su aplicación más para apoyo de terapias (a base de repeticiones) que de comunicación. Un proyecto similar al propuesto en esta tesis se presentó en [10] para terapias, sin embargo no se pudo constatar el

seguimiento del mismo.

Con excepción del sistema STRAPTK, estas aplicaciones (incluyendo aquellas para el español mexicano) no permiten modificar los parámetros de configuración de los componentes del reconocedor. Por lo que el ajuste de desempeño sólo depende de añadir más frases de entrenamiento o seleccionar un vocabulario que mejor sea reconocido por el sistema. STRAPTK habilita la creación de un sistema dependiente de usuario, pudiendo elegir el número de componentes gaussianos y las palabras a utilizar para su entrenamiento. Sin embargo esta lista de palabras se encuentra pre-definida y el sistema en general está restringido a reconocer pocas palabras. También es necesario ingresar a un módulo en específico para añadir nuevas muestras de voz. También, ninguna de estas aplicaciones está contemplada para comunicación, por lo que el reconocimiento de frases continuas no se encuentra implementada.

Finalmente, de los estudios realizados para otros idiomas, se obtuvo que si el SRAH se puede entrenar continuamente, su desempeño puede mejorar incluso para usuarios con niveles severos de disartria [9, 19, 51]. Por lo tanto, el que el SRAH reciba información acústica continuamente y que la pueda administrar para re-entrenamiento de sus componentes es una de las funciones del sistema propuesto. Con esto se considera que se puede reducir las tasas de confusión de palabras, eliminando la necesidad de re-seleccionar vocabulario, y aumentar el tamaño del vocabulario para reconocer frases de uso cotidiano. A continuación se presentan los resultados de las pruebas realizadas con la interfaz propuesta, la cual aborda los puntos mencionados.

5.1 Pruebas con Voz Normal

La interfaz de voz fue instalada en una PC portátil del tipo Netbook con el siguiente hardware: 1GB de memoria RAM y procesador Intel Atom N570 a 1.66 GHz. El micrófono fue integrado en una diadema la cual fue conectada a la PC.

Inicialmente los módulos de Adaptación y Reconocimiento se probaron con 10 usuarios con voz normal (5 hombres, 5 mujeres). El vocabulario consistió de 12 frases (ver Tabla 5.1) usadas para el control de un robot. Cada usuario leyó 10 veces cada frase, por lo tanto, 120 frases fueron grabadas por cada usuario. Solo frases reconocidas completamente fueron consideradas, mostrando el desempeño mostrado en la Tabla 5.1.

Tabla 5.1: Porcentajes de frases reconocidas correctamente por el SRAH con usuarios con voz normal.

No.	Frase de Control
1	BOT AVANZA RAPIDO DOS METROS
2	BOT RETROCEDE LENTO
3	BOT GIRA CUARENTA Y CINCO GRADOS A LA IZQUIERDA
4	BOT GIRA CUARENTA Y CINCO GRADOS A LA DERECHA
5	CUBE SIRVE BOTELLA
6	CUBE TOMA EL VASO
7	BOT SAL POR PUERTA UNO
8	BOT ENTRA POR PUERTA DOS
9	BOT SIRVE LA COPA
10	BOT AVANZA LENTO DOS METROS
11	CUBE INICIO
12	BOT DETENTE

Usuario (H=Hombre, M= Mujer)	Fallas/Total	% Éxito
Usuario H1	4/120	96.67
Usuario H2	3/120	97.50
Usuario H3	5/120	95.83
Usuario H4	5/120	95.83
Usuario H5	2/120	98.33
Usuario M1	7/120	94.17
Usuario M2	5/120	95.83
Usuario M3	8/120	93.33
Usuario M4	6/120	95.00
Usuario M5	3/120	97.50
	% Total	96.00

Estos resultados dieron confianza acerca del desempeño del sistema para usarse con usuarios diferentes al que se uso para el entrenamiento de los modelos acústicos (ver Sección 4.2). Como el porcentaje de reconocimiento correcto de frases completas fue mayor a 96%, se puede asumir que el porcentaje de precisión en el reconocimiento de palabras es significativamente mayor. Para los experimentos con voz disártrica la métrica de desempeño fue la precisión de reconocimiento de palabras (% WAcc, Ec. 2.6) y la tasa de error (%WER, Ec. 2.7).

5.1.1 Pruebas con Voz Disártrica

El centro del Sistema Nacional para el Desarrollo Integral de la Familia (SNDIF), ubicado en Blv. Tierra del Sol Esq. Calle Pedro Sepulveda, en la Ciudad de Huajuapan de León (Oaxaca) proporcionó el apoyo para buscar y reclutar voluntarios para participar en este trabajo. En la Tabla 5.2 se muestra una semblanza del personal del centro que colaboró con este proyecto.

Durante el proceso de búsqueda algunos requerimientos fueron establecidos de acuerdo a las recomendaciones de los terapeutas del centro SNDIF. De esta manera, se definieron los siguientes requerimientos básicos para los posibles candidatos para este trabajo:

- Diagnóstico de disartria no causado por enfermedad neurodegenerativa (por ejem-

plo, disartria no causada por Alzheimer).

- Preservación de facultades cognitivas.
- Sin diagnóstico de problemas de entendimiento de lenguaje.
- Que tengan más de 15 años de edad (participantes más jóvenes requieren de atención especial).

Tabla 5.2: Personal del centro SNDIF que colaboró en la realización del proyecto.

Nombre	Actividad
Dra. María Luisa Gutiérrez	Coordinadora del centro SNDIF de la ciudad de Huajuapan de León, Oaxaca. México.
Rocío Bazán Pacheco	Terapia del Lenguaje
Diana Pérez Hernández	Terapia Ocupacional

En un período de tres meses se trabajó en la búsqueda de candidatos, y aunque hubieron acercamientos, en la mayoría de los casos los candidatos no cubrían los requerimientos básicos. En el Anexo C se presentan algunos de estos casos.

Finalmente se pudo concretar la colaboración con dos participantes que cubrían los requerimientos básicos. Por confidencialidad estos participantes se identifican como GJ y MM. En la Tabla 5.3 se muestra el cuadro clínico general de GJ y MM. Note que para MM falta alguna información. Esto es porque MM fue contactado por referencias personales fuera del centro SNDIF, por lo tanto no había registros formales de su condición. Sin embargo, su nivel de disartria fue evaluado mediante grabaciones de voz por los terapeutas.

Para este proyecto se consideró significativo el trabajar con dos usuarios. Esto dado que en la mayoría de los trabajos de investigación en este campo se trabaja con un sólo usuario como en [29, 34, 35]. Esto es por la dificultad de los pacientes en dar seguimiento a las pruebas de evaluación por el esfuerzo y movilidad requerida, al igual que la continua asistencia del técnico y terapéuta para llevar a acabo las mismas.

Dos SRAHs fueron probados con estos usuarios:

- SRAH base (HMMDEFS) entrenado con muestras de voz de un usuario con voz normal (Independiente de Usuario, IU). Con este SRAH se obtuvieron los resultados reportados en la Tabla 5.1.

Tabla 5.3: Perfil de los usuarios con disartria GJ y MM.

Nombre	GJ		
Edad	64	Género	Masculino
Patologías	Disartria leve-moderada causada por un accidente vascular cerebral. Hemiplegia del lado izquierdo (parálisis del brazo, pierna, y torso izquierdo). 90% de pérdida de la visión. Escoliosis (columna vertebral curva de lado a lado)		

Nombre	MM		
Edad	37	Género	Masculino
Patologías	Disartria moderada causada por una traqueotomía.		

- SRAH base (HHMDEFS) entrenado con muestras de voz del usuario que utilizará el sistema (Dependiente de Usuario, DU).

Como se comentó en la Sección 4.2, esto se realizó para evaluar dos enfoques que se han empleado en otros proyectos de RAH para voz disártrica. Un **SRAH DU** fue construído y evaluado por GJ sin el módulo de Adaptación de Usuario ya que este sistema fue entrenado con su voz. Sin embargo, no fue posible hacer lo mismo para el usuario MM por restricciones de tiempo y disponibilidad para proveer la cantidad de muestras de voz necesarias. Porque GJ tiene una visibilidad casi nula, la interfaz fue operada por un familiar para las actividades de adaptación y evaluación.

Para evaluar el **SRAH IU**, inicialmente GJ y MM pasaron por el módulo de Adaptación de Usuario antes de usar el módulo de Reconocimiento de Voz. El SRAH IU fue evaluado con diferentes cantidades de frases de adaptación para estudiar el efecto de la adaptación estática y dinámica sobre el desempeño del sistema. Tres condiciones de adaptación fueron consideradas, y para esto se establecieron las siguientes configuraciones:

- **SRAH IU I:** SRAH base adaptado con sólo las 16 frases del módulo de Adaptación de Usuario (adaptación estática).

- **SRAH IU II:** SRAH IU I adaptado con 11 frases adicionales (distintas a las

de la adaptación estática) mientras se usa el módulo de Reconocimiento de Voz (adaptación dinámica I).

- **SRAH IU III:** SRAH IU II adaptado con 11 frases adicionales (distintas a las 11 usadas para la adaptación dinámica I), mientras se usa el módulo de Reconocimiento de Voz (adaptación dinámica II).

Las 22 frases para la adaptación dinámica fueron frases espontáneas relacionadas con las actividades cotidianas de los usuarios GJ y MM (estas se muestran en el Anexo D, Tabla D.1). Estas fueron añadidas al diccionario del SRAH y el ML antes de las sesiones de evaluación. Finalmente, cada configuración del SRAH de la interfaz fue evaluado con 50 frases espontáneas con un total de 275 palabras diferentes (ver Anexo D, Tabla D.2). Estas frases fueron diferentes de aquellas usadas para la adaptación estática y dinámica.

La versión latinoamericana de Dragon NaturallySpeaking (**LTN Dragon**) [42] fue usada para propósitos de comparación de desempeños. Para este sistema IU, el usuario llevó a cabo la sesión de adaptación correspondiente, la cual consistió en leer uno de diez relatos. En este caso, se escogió el relato "Las Aventuras de Pinocho" (Anexo D, Tabla D.3), el cual consistió de 310 palabras diferentes. Una vez que esta sesión fue completada, el LTN Dragon fue evaluado en modo de dictado con las mismas 50 frases espontáneas.

Dada las posibles combinaciones de valores para el factor de escala gramática y el número de componentes gaussianos, se optó por utilizar ocho (8) componentes gaussianos basados en los resultados presentados en la Tabla 4.1. El factor de gramática se escogió de acuerdo al desempeño del SRAH DU, siendo el mismo para el SRAH IU para una comparación más fiel.

En la Figura 5.1 se muestra el desempeño del SRAH DU con diferentes valores de factor de escala gramática con el usuario GJ. Como se puede observar, inicialmente con un valor s=5, una precisión de aproximadamente 50% es obtenida con las frases de evaluación. Conforme este valor aumentó se obtuvo un máximo de 75% cuando s=20, lo cual muestra el efecto de reducir la perplejidad del ML mediante el aumento en la restricción del mismo. Después de este punto el desempeño tiende a disminuir. Esto es por un sobreuso de la información del ML, lo cual tiende a descartar la información de

la señal de voz. Sin embargo, cabe señalar que para voz normal usualmente un valor s=5 es suficiente [61], pero para voz disártrica este valor es mayor [3]. La información presentada en la Figura 5.1 corrobora esta información para el caso de voz disártrica en español mexicano.

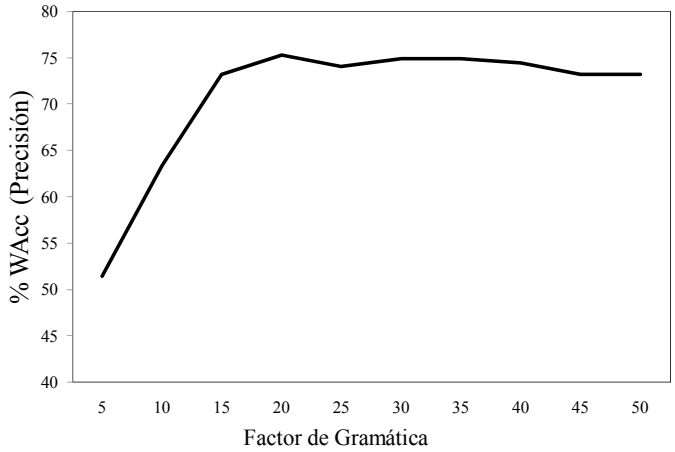


Figura 5.1: Desempeño del SRAH DU con 50 frases de evaluación y diferentes valores de factor de gramática.

El desempeño del SRAH IU, con un factor s=20, es comparado con el de otros sistemas, comerciales y de investigación, incluyendo la transcripción humana de voz normal. Los resultados de las sesiones de evaluación se presentan en la Tabla 5.4.

En la Figura 5.2 se muestra gráficamente el análisis comparativo de resultados mostrado en la Tabla 5.4. Como se presenta, los sistemas IU tuvieron un desempeño de 93.67%-94.94% para GJ, y de 90.04%-94.70% para MM. Este desempeño es comparable a la percepción (transcripción) humana (96%-98%) y al de SRAHs comerciales para voz normal bajo condiciones similares de tamaño de vocabulario [41]. Cuando se compara con SRAHs adaptados (o desarrollados) para usuarios con disartria, la interfaz de voz tiene un desempeño comparable al de SRAHs con muy pequeño vocabulario ($<$ 100 palabras) [15, 24] y niveles similares de disartria [15]. Al considerar un SRAH con tamaño similar de vocabulario de evaluación (300 palabras, 77.28%-82.20%) [34] y un usuario con nivel similar de disartria, la interfaz propuesta tiene mejor desempeño.

Considerando ambas metodologías o enfoques, IU (Independiente de Usuario) y DU (Dependiente de Usuario), el SRAH DU tuvo un desempeño bajo con WAcc = 75% para el usuario GJ. En cambio, la metodología de usar un SRAH IU mostró mejoras cuando se evaluó el SRAH comercial Dragon NaturallySpeaking: 83.50% para GJ, y 82.40%

para MM. Mejoras adicionales fueron obtenidas sobre este SRAH con la interfaz de voz propuesta usando la metodología IU.

Tabla 5.4: Precisión (WAcc) y tasa de error (WER) de la interfaz de voz y su comparación con otros sistemas: percepción humana y SRAHs comerciales: *[41]; SRAHs comerciales y de investigación usados con voz disártrica con diferentes niveles de inteligibilidad: ** alta [15], *** moderada [34], y **** baja [24].

Voz	Sistema	WAcc	WER
Normal	* Reconocimiento Humano	96% - 98%	2% - 4%
	* SRAH Comerciales (≤ 1000 palabras)	80% - 96%	4% - 20%
Disártrica (idioma extranjero)	** Dragon Dictate (Version 1.01A) ("Pledge of Allegiance", 24 palabras).	80% - 98%	2% - 20%
	*** Dragon Naturally Speaking (≈ 300 palabras).	77.28% - 82.20%	17.80% - 22.72%
	**** STARDUST (Environmental Control System, 10 palabras).	88.5% - 95.4%	4.6% - 11.5%
		GJ	
Disártrica (idioma español mexicano) 275 palabras	SRAH DU	75.00%	25.00%
	SRAH IU I	93.67%	6.33%
	SRAH IU II	94.51%	5.49%
	SRAH IU III	94.94%	5.06%
	LTN Dragon	83.50%	16.50%
		MM	
	SRAH IU I	90.04%	9.96%
	SRAH IU II	93.54%	6.46%
	SRAH IU III	94.70%	5.30%
	LTN Dragon	82.40%	17.60%

En la Figura 5.3 se muestran las matrices de confusión fonética para GJ y MM. Esta información fue comparada con las pruebas de percepción realizadas por los terapeutas usando el grupo de 49 palabras usadas en el módulo correspondiente. Para GJ se corroboraron las confusiones significativas observadas para los fonemas $/b/$, $/r(/$, $/u/$, $/f/$, $/l/$, $/e/$, $/Z/$ y $/g/$. Para los demás fonemas como */p/*, */ñ/*, y */t/*, */a/* e */i/*, muy pocas deficiencias fueron percibidas. Por lo tanto, hubo un acuerdo con los patrones mostrados en la matriz de confusión. Para el usuario MM más confusiones fueron observadas, aunque con menor número de inserciones o eliminaciones. Estas confusiones fueron corroboradas por los terapeutas, notando sólo un desacuerdo con los patrones de los fonemas */b/* y */ñ/*. Note que aunque hay confusiones, eliminaciones e inserciones significativas, el patrón observado no es muy diferente al de un usuario con voz nor-

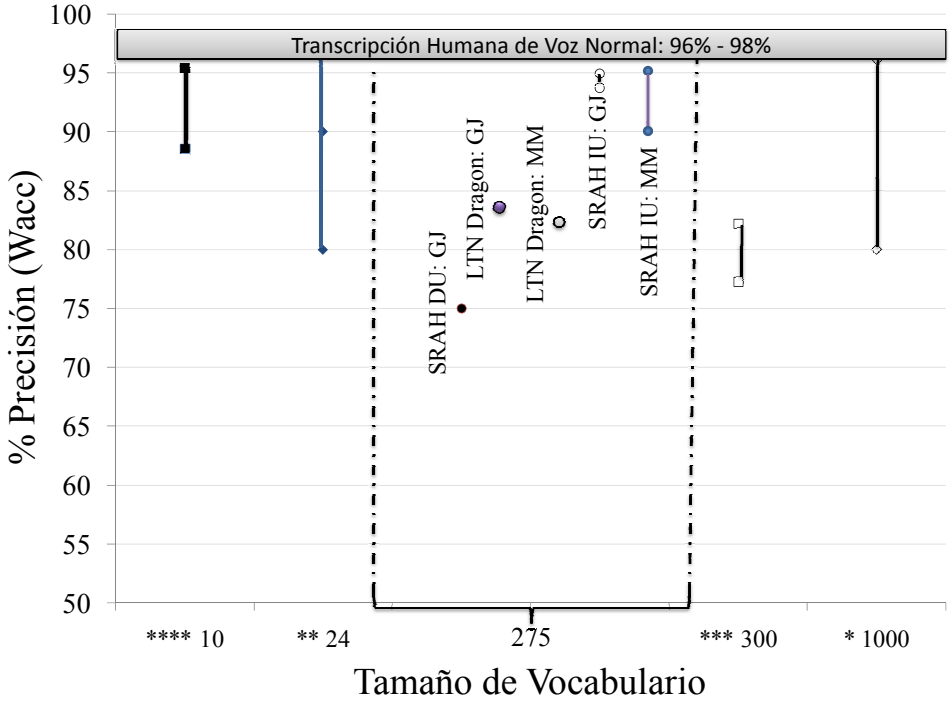


Figura 5.2: Análisis visual de los resultados presentados en la Tabla 5.4.

mal (ver Figura 4.9). Esto puede ser causado por el nivel bajo-moderado de la disartria de los usuarios, en donde severas anormalidades en la articulación de fonemas no son evidentes.

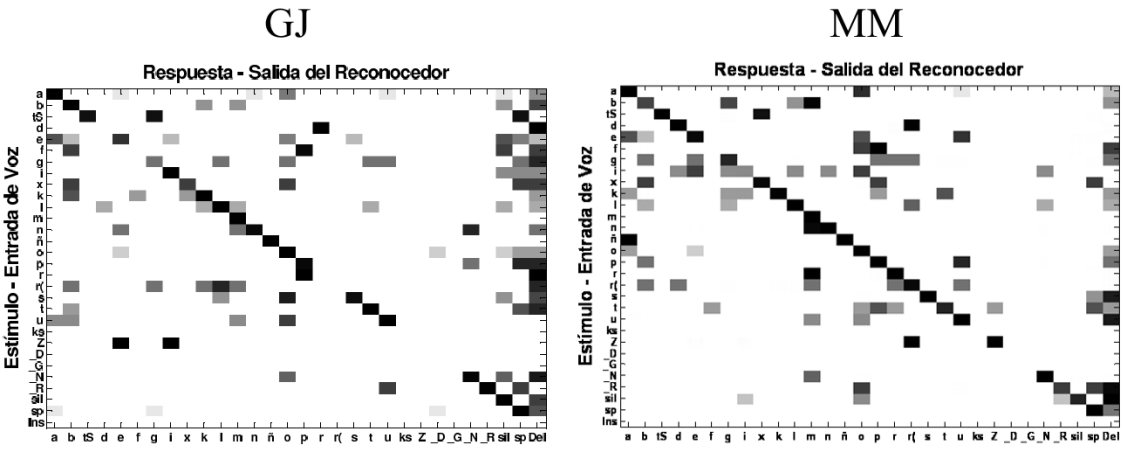


Figura 5.3: Matrices de confusión fonética para los usuarios GJ y MM.

Capítulo 6

Conclusiones

En los capítulos anteriores se explicó el desarrollo de esta tesis y se mostró cómo se fueron logrando uno a uno los objetivos. El desarrollo de la interfaz de voz para personas con disartria cuya lengua es el español mexicano fue el objetivo general que engloba este proyecto.

El elemento más importante de la interfaz es el SRAH, que consta de los modelos acústicos, modelo de lenguaje (ML), corpus textual/oral, diccionario fonético, y algoritmo de búsqueda (decodificación). La interfaz por lo tanto involucra el medio para la administración y configuración de todos estos elementos (en tiempo de ejecución y de manera automática) para cubrir las funciones de comunicación y apoyo a diagnóstico de usuarios con disartria.

Para el desarrollo de la interfaz se abordaron dos enfoques principales para la creación del SRAH: Dependiente de Usuario (DU) e Independiente de Usuario (IU). Aunque se ha argumentado en los trabajos citados que el desarrollo de SRAH's DU son mejores para usuarios con disartria, éstos requieren más tiempo y trabajo de cerca con el paciente para ser desarrollados. Esto porque este enfoque implica el desarrollar un corpus de entrenamiento con la voz del usuario que va a usar el sistema, el cual debe estar etiquetado correctamente.

Los SRAH's IU son aquellos que necesitan una técnica de adaptación para que puedan ser usados por un usuario diferente. Con este enfoque, un usuario nuevo puede comenzar a usar el SRAH mucho más rápido que si se construye uno especial para él. En este trabajo, se obtuvo un alto rendimiento en la tasa de precisión con el enfoque IU,

debido la técnica de adaptación dinámica que se implementó en el sistema sin dejar de considerar el aporte de las variables que se manejan en la misma. Como se resume en la Figura 6.1, la tasa de precisión fue del 93.67/90.04 - 94.94/94.70% para el SRAH IU, y 75% para el SRAH DU.

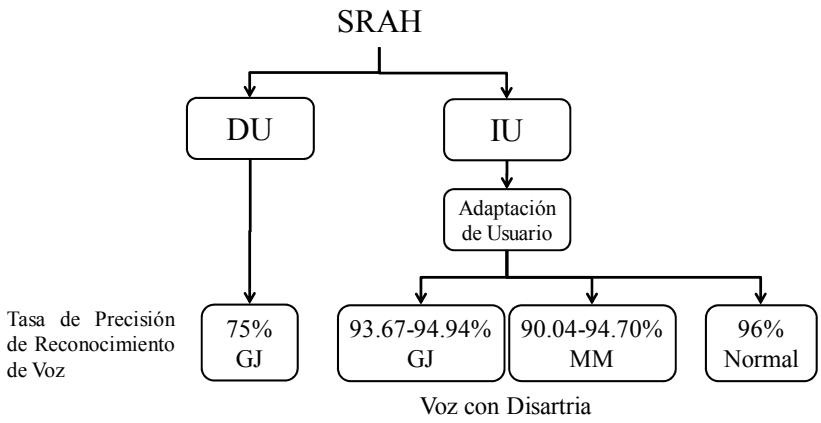


Figura 6.1: Comparación de SRAH: DU y IU.

Mientras que para el desarrollo del SRAH DU para el usuario con disartria GJ tomó cerca de tres semanas, la adaptación para que pudiera usar el SRAH IU tomó sólo algunos minutos (al igual que para el usuario MM). Por lo tanto, la interfaz (y el SRAH) puede ser probado con mayor facilidad por otros usuarios sin restricción de un tiempo largo de preparación. Por ejemplo, las pruebas con frases de control y usuarios con voz normal sólo tomaron cerca de 5 minutos de preparación (proveer muestras de voz para adaptación) para utilizar la interfaz de voz, obteniendo una tasa de precisión del 96%.

En cuanto a los factores (o variables) a considerar para el desarrollo de la interfaz para comunicación de usuarios con voz disártrica, se identificaron los siguientes: los componentes gaussianos de los modelos acústicos (HMM's), el vocabulario para el control de la perplejidad del ML, y el factor de escala gramatical. Los componentes gaussianos repercuten en un mejor modelado de la señal de voz, por lo que el primer módulo de la interfaz habilita la construcción del SRAH IU con el número que componentes que el usuario requiera.

El segundo módulo de la interfaz habilita la adición de nuevo vocabulario en tiempo real, y el control del factor de gramática. Esto para reducir la perplejidad del ML, lo cual se ha reportado como importante para el reconocimiento de voz disártrica (la perplejidad

es inversamente proporcional a la tasa de precisión del SRAH, ver Figura 5.1). De igual manera se implementaron funciones para la adaptación dinámica y acumulativa de los HMM's del SRAH, al igual que de construcción del diccionario fonético. Con esto, se tiene un proceso de refinamiento acumulativo del SRAH para la voz del usuario disártrico. También, el vocabulario y ML de la interfaz se vuelve escalable para otras aplicaciones o contextos.

Con estos módulos, al asignar valores específicos a cada una de las variables identificadas, se obtuvo de manera consistente para dos usuarios con disartria tasas máximas cercanas al 95% (ver Tabla 5.4). Sin embargo, en cuanto a la comparación de este desempeño con otros SRAH's, ésta sólo se pudo realizar con sistemas desarrollados principalmente para el idioma inglés. Esto dado que en México no se ha desarrollado una aplicación igual o similar a la propuesta.

Por lo tanto, la comparación se hizo considerando trabajos en otro idioma pero con usuarios de nivel similar de disartria (en este caso, bajo-moderado) y tamaño de vocabulario. Con el control de las tres variables definidas, las tasas de reconocimiento obtenidas con la interfaz son equiparables, o mejores, a las obtenidas por dichos sistemas con vocabularios más pequeños. Incluso, el desempeño es muy cercano al del reconocimiento de voz humano y el de SRAH's comerciales para voz normal.

El tercer módulo de la interfaz, propuesto para asistir el diagnóstico de deficiencias de pronunciación, se desarrolló mediante un alineador de secuencias fonéticas. Este alineador se basó en la similitud acústica de fonemas del español mexicano. Esta similitud se estimó a partir de la matriz de confusión fonética obtenida con el corpus de entrenamiento y la respuesta (salida) del SRAH. Con este módulo se proporcionaron visualmente patrones representativos de las deficiencias fonéticas reales del usuario, las cuales fueron corroboradas por los terapeutas.

De esta manera se considera que esta interfaz, con las funciones implementadas, contribuye al campo de desarrollo de tecnologías para personas con discapacidad en el habla. Siendo que este campo no se ha explorado de manera significativa en México (o para el español mexicano).

6.1 Contribuciones

El desarrollo del proyecto de tesis tiene las siguientes contribuciones:

- Creación de un corpus transcrito de manera formal (usando la fonética descrita en [11]) a nivel ortográfico y fonético de un usuario con disartria del español mexicano.

- Interfaz de voz para comunicación y diagnóstico de personas con disartria con las siguientes características:

  - permite la adaptación dinámica e ingreso de vocabulario en tiempo real para el español mexicano;
  - realiza la construcción automática de los componentes del SRAH con parámetros especificados por el usuario;
  - genera una matriz de confusión fonética como apoyo a terapeutas para la observación de patrones de deficiencias en la articulación de fonemas.

- Se hace énfasis en que no existe un sistema similar al contexto de uso abordado por esta tesis para el español mexicano, y con el enfoque hacia la comunicación y apoyo al diagnóstico de la disartria. Comparado con otros sistemas [19, 21] con vocabularios muy pequeños ($<$70 palabras) y en los que sólo se reconocen comandos (palabras discretas), la interfaz propuesta reconoce frases de varias palabras.

6.2 Trabajo a Futuro

Se consideran los siguientes puntos como temas para trabajo a futuro:

- Extender el trabajo multidisciplinario para mejorar el módulo para el diagnóstico de la disartria: refinar la clasificación de fonemas para el módulo de confusión fonética y proveer de una métrica para evaluar los diferentes niveles de disartria.

- Extender el diseño gráfico y usabilidad del módulo para comunicación de voz disártrica.

- Evaluar el rendimiento de la interfaz con un vocabulario más extenso ($>$ 1000 palabras) y la adaptación dinamica de ésta.
- Analizar los efectos del control de la perplejidad y la adaptación dinámica para usuarios con disartria más severa.
- Adaptación de la interfaz para su implementación en dispositivos móviles.
- Incrementar el tamaño del corpus de entrenamiento con usuarios disártricos de diferentes edades y de ambos géneros.
- Adaptar la interfaz para usuarios con otras discapacidades (como en el caso de GJ).
- Desarrollo de un sub-módulo para la creación en línea de listas de palabras (con fonemas en específico) para el diagnóstico de la disartria.
- Explorar el enfoque DU e IU para usuarios con disartria más severa.
- Desarrollo e integración de un sintetizador de voz para el español mexicano (femenino y masculino).

Bibliografía

[1] Aguilar, E. and Serra, M. *A-RE-HA. Análisis del Retraso del Habla (2a Edición)*. Universitat de Barcelona (UBe), España, ISBN: 978-84-475-3161-5, 2007.

[2] Beskow, J. and Sjolander, K. *Wavesurfer v.1.8.8.3p3*. http://www.speech.kth.se/wavesurfer/, Consultado el 12/03/2012.

[3] Caballero, S.O. and Cox, S.J. Modelling Errors in Automatic Speech Recognition for Dysarthric Speakers. *EURASIP J. Adv. Signal Processing*, 2009:1–14, 2009.

[4] Cal, M., Núñez, P., and Palacios, I.M. *Nuevas Tecnologías en Lingüística, Traducción y Enseñanza de Lenguas*. Universidade de Santiago de Compostela, ISBN: 84-9750-518-2, 2005.

[5] Cannarozzi, G.M. String alignment using dynamic programming. In *Institute of Computational Science, ETH Zürich*, http://www.biorecipes.com/DynProgBasic/code.html, Consultado el 12/03/2012.

[6] Carrillo, R. Diseño y Manipulación de Modelos Ocultos de Markov Utilizando Herramientas HTK. *Ingeniare: Revista chilena de ingeniería*, 15(1):18–26, 2007.

[7] Chen, S., Beeferman, D., and Rosenfeld, R. Evaluation metrics for language models. *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[8] Cohen, S.M., Elackattu, A., Noordzij, J.P., Walsh, M.J., and Langmore, S.E. Palliative Treatment of Dysphonia and Dysarthria. *Otolaryngologic Clinics Of North America*, 42:107–121, 2009.

[9] Coleman, C. L. and Meyers, L.S. Computer recognition of the speech of adults with cerebral palsy and dysarthria. *Augmentative and Alternative Communication*, 7:34–42, 1991.

[10] Copalcua-Pérez, Ma. de la Paz. *Sistema de procesamiento de fonemas para la rehabilitación de personas con problemas de habla mediante técnicas de aprendizaje automático*. Tesis de Maestría en Sistemas Computacionales del Instituto Tecnológico de Apizaco, México., 2009.

[11] Cuétara, J. *Fonética de la Ciudad de México: Aportaciones de las Tecnologías del Habla*. Tesis de Maestría, Universidad Nacional Autónoma de México, 2004.

[12] Darley, F., Aronson, A., and Brown, J. Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, 12:462–496, 1969.

[13] Davis, S.B. and Mermelstein, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28(4):357–366, 1980.

[14] Fabián-Aguilar, Aldo-Ernesto. *Desarrollo e Implementación de un Párser Semántico para el Módulo Golem-Universum*. Universidad Tecnológica de la Mixteca, Tesis de Ingeniería en Computación, 2011.

[15] Ferrier, L. J., Shane, H. C., Ballard, H. F., Carpenter, T., and Benoit, A. Dysarthric speaker's intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication*, 11:165–175, 1995.

[16] Foundation for Assistive Technology (FAST). Register Charity Number 1061636. In *http://www.fastuk.org/research/projview.php? trm=STARDUST& id=216*, Consultado el 12/03/2012.

[17] Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., and Zue, V. *TIMIT Acoustic Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia, 1993.

[18] Green, P., Carmichael, J., Hatzis, A., Enderby, P., Hawley, M.S, and Parker M. Automatic speech recognition with sparse training data for dysarthric speakers. In

Proc. European Conference on Speech Communication Technology, pages 1189–1192, 2003.

[19] Green, P, Hatzis, A, Parker, M, Carmichael, J, Cunningham,S, O'Neill, P, and Palmer, R. An Integrated Toolkit Deploying Speech Technology for Computer Based Speech Training with Application to Dysarthric Speakers. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 03), Geneva, Switzerland*, pages 2213–2216, 2003.

[20] Livingston, N. Hamidi, F., Baljko, M. and Spaltecholz, L. KeySurf a character controlled browser for people with physical disabilities. In *In: Proc. of WWW, ACM Press, New York*, 2008.

[21] Livingston, N. Hamidi, F., Baljko, M. and Spaltecholz, L. CanSpeak: A custimizable speech interface for people with dysatric speech. In *K. Miesenberger et al. (Eds.): ICCHP 2010, Part I, LNCS 6179,Springer -Verlag Berlin Heidelberg*, 2010.

[22] Hawley, M., Cunningham, S., Cardinaux, F., Coy, A., O'Neill, P., Seghal, S., and Enderby, P. Challenges in developing a voice input voice output communication aid for people with severe dysarthria. In *Proc. European Conference for the Advancement of Assistive Technology in Europe*, 2007.

[23] Hawley, M, Enderby, P, Green, P, Brownsell, S, Hatzis, A, Parker, M, Carmichael, J, Cunningham,S, O'Neill, P, and Palmer, R. STARDUST Speech Training And Recognition for Dysarthric Users of assistive Technology. In *Proceedings of the 7th European Conference for the Advancement of Assistive Technology in Europe, Dublin, Ireland*, 2003.

[24] Hawley, M. S., Enderby, P., Green, P., Cunningham, S., Brownsell, S., Carmichael, J., Parker, M., Hatzis, A., O'Neill, P., and Palmer, R. A speech-controlled environmental control system for people with severe dysarthria. *Medical Engineering & Physics*, 29:586–593, 2007.

[25] Hawley, M.S., Green, P, Enderby, P., Cunningham, S., and Moore R.K. Speech Technology for e-Inclusion of People with Physical Disabilities and Disordered

Speech. In *Proc. of the 9th European Conference on Speech Communication and Technology (Interspeech 05, Lisbon, Portugal)*, pages 445–448, 2005.

[26] Jayaram, G. and Abdelhamied, K. Experiments in dysarthric speech recognition using artificial neural networks. *Journal of Rehabilitation Research and Development*, 42:162–169, 1995.

[27] Jurafsky, D. and Martin, J.H. *Speech and Language Processing*. Pearson: Prentice Hall, 2009.

[28] Karnjanadecha, M. and Zahorian, S. Signal Modeling for High-Performance Robust Isolated Word Recognition. *IEEE Transactions On speech and Audio Processing*, 9(6), 2001.

[29] Kotler, A. and Thomas-Stonell, N. Effects of speech training on the accuracy of speech recognition for an individual with a speech impairment. *Augmentative and Alternative Communication*, 13:71–80, 1997.

[30] Kotler, A., Thomas-Stonell, N., Doyle, P., Leeper, H. A., Dylke, M., O'Neill, C., and Rolls, K. Comparative perceptual and computerized speech recognition functions for dysarthric speakers. *American Speech Language and Hearing Association*, 1993.

[31] Leggetter, C.J. and Woodland, P.C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9(2):171–185, 1995.

[32] Lewis. C.H. and Rieman, J. *Task-Centered User Interface Design: A Practical Introduction*. Shareware, http://oldwww.acm.org/perlman/uidesign.html, Consultado el 12/02/2011.

[33] Lizandra Laplaza, Rafael. Dificultades en el Desarrollo del Lenguaje Oral e Intervención. In *http://eoepzar1.educa.aragon.es/PROFESORADO/Dificultades_lenguaje_oral.pdf*, Consultado el 12/03/2012.

[34] Manasse, N. J., Hux, K., and Rankin-Erickson, J. Speech recognition training for enhancing written language generation by a traumatic brain injury survivor. *Brain Injury*, 14:1015–1034, 2000.

[35] Manasse, N. J., Hux, K., Rankin-Erickson, J., and Lauritzen, E. Accuracy of three speech recognition systems: Case study of dysarthric speech. *Augmentative and Alternative Communication*, 16:186–196, 2000.

[36] Matsumasa, H., Takiguchi, T., Ariki, Y., LI, I., and Nakabayashi, T. Integration of metamodel and acoustic model for speech recognition. In *Proc. of Interspeech 2008*, pages 2234–2237, 2008.

[37] Miranda, P.C, Camal, U.R, Cen, M.J, González, S.C, González, S.S, García, M, and Narvaez, D.L. Un Juego de Gravedad con Reconocimiento de Voz para Niños con Problemas de Lenguaje. In *Workshop on Perspectives, Challenges and Opportunities for Human-Computer Interation in Latin America*, 2007.

[38] Montalto, L. *Fiesta en la Montaña*. http://home.cc.umanitoba.ca/ fernand4//fiesta.html, Consultado el 12/03/2012.

[39] Moriana, M.J. La disartria. In *DEP. LEGAL 2922-2007. ISSN 1988-6047*, 2009.

[40] Moya-García, Edith. *Entrenamiento Dinámico de Modelos Acústicos de Reconocedores de Voz para los Corpora en Español de México: DIMEx100 niños y adultos*. Universidad Tecnológica de la Mixteca, Tesis de Ingeniería en Computación, 2011.

[41] National Institute of Standards and Technology (NIST). *The History of Automatic Speech Recognition Evaluations at NIST*. http://www.itl.nist.gov/iad/mig/publications/ASRhistory/ index.html, Consultado el 12/03/2012.

[42] Nuance Communications, Inc. *Dragon NaturallySpeaking, Español*. Version 10.00.200.161, 2008.

[43] Peña-Casanova, J. *Introducción a la Patología y Terapéutica del Lenguaje*. Elsevier, España, 2002.

[44] Perea, G.G and Miranda, P.C. Diseño de un corpus de voz en español para niños en edad escolar con problemas de lenguaje. *Faz, Revista Diseño de Interacción. ISSN 0718-526X*, pages 26–37, 2009.

[45] Pineda, L. A. *El proyecto DIME y el robot conversacional Golem: Una experiencia multidiciplinaria entre la computación y la lingüística*. Universidad Autónoma de México, UNAM, 2008.

[46] Pineda, L.A., Villaseñor, L., Cuétara, J., Castellanos, H., Galescu, L., Juárez, J., Llisterri, J., and Pérez, P. The corpus dimex100: Transcription and evaluation. *Language Resources and Evaluation*, 44:347–370, 2010.

[47] Pineda, L.A., Villaseñor, L., Cuétara, J., Castellanos, H., and López, I. DIMEx100: A new phonetic and speech corpus for Mexican Spanish. In *Advances in Artificial Intelligence, Iberamia-2004*, 2004.

[48] Prater, J. and Swiff, R. *Manual de Terapia de la Voz*. Salvat, 1989.

[49] Rabiner, L. A tutorial on Hidden Markov Models and selected applications in speech recognition. In *Proc. IEEE*, volume 37, pages 257–286, 1989.

[50] Rabiner, L. and Juang, B.H. *Fundamentals of Speech Recognition*. Prentice Hall, NY, USA, 1993.

[51] Raghavendra, P., Rosengren, E., and Hunnicutt, S. An investigation of different degrees of dysarthric speech as input to speaker adaptive and speaker dependent recognition systems. *Augmentative and Alternative Communication*, 17:265–275, 2001.

[52] Resch, B. *Automatic Speech Recognition with HTK*. Signal Processing and Speech Communication Laboratory. Inffeldgase. Australia, 2003.

[53] Saltillo Corporation. In *http://www.saltillo.com/products*, Consultado el 12/03/2012.

[54] Sánchez, María Gabriela. Desórdenes Motores del Habla y PROMPT (Parte II). In *http://www.espaciologopedico.com/articulos2.php?Id_articulo=1692*, Consultado el 12/03/2012.

[55] Shneiderman, B. *Design the User Interface: Strategies for Effective Human Computer Interaction*. Addison-Wesley Longman, 2004.

[56] Speech Enhancer. Voicewave Technology Inc. In *http://www.speechenhancer.com/equipment.htm*, Consultado el 12/03/2012.

[57] Strik, H., Sanders, E., Ruiter, M., and Beijer, L. Automatic recognition of dutch dysarthric speech: a pilot study. *ICSLP*, pages 661–664, 2002.

[58] Swanberg, M.M., Nasreddine, Z.S., Mendez, M.F., and Cummings, J.L. *Speech and Language*. Goetz CG, Ed. Textbook of clinical Neurology: 3rd. Ed. Philadelphia, Pa: Elsevier, 2007.

[59] Villamil-Espinosa, I.H. *Aplicaciones en Reconocimiento de Voz utilizando HTK*. Tesis de Maestría en Electrónica. Pontificia Universidad Javeriana. Santa Fe de Bogota, DC., 2005.

[60] Voice Input Voice Output Communication Aid (VIVOCA). *Clinical Applications of Speech Technology, Speech and Hearing Group, Department of Computer Science, University of Sheffield*. http://www.shef.ac.uk/cast/projects/vivoca, 2008.

[61] Young, S. and Woodland, P. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2006.

Anexo A

Texto Representativo para Corpus de Entrenamiento

Tabla A.1: Ficha de Articulación: Selección de palabras para diagnóstico de disartria.

Palabras	Fonemas a Observar	Palabras	Diptongos a Observar	Palabras	Heterosílabas a Observar
Piña	(p)(ñ)	Peine	(ei)	Blusa	(blu)
Taco	(t)(k)	Jaula	(au)	Bruja	(bru)
Cama	(k)(m)	Guante	(ua)	Flor	(flo)
Sopa	(s)(p)	Hielo	(ie)	Fresa	(fre)
Gato	(g)(t)	Huevo	(ue)	Iglesia	(gle)
Liga	(l)(g)	Piano	(ia)	Tigre	(gre)
Vaca	(b)(k)	Árbol	(rb)	Plátano	(pla)
Uvas	(b)(s)	Bolsa	(ls)	Clavo	(cla)
Dado	(d)	Barco	(rk)	Cruz	(cru)
Foco	(f)(k)	Campana	(mp)	Tren	(tre)
Jirafa	(x)(r)(f)	Canasta	(st)		
Casa	(k)(s)	Chango	(ng)		
Pájaro	(p)(x)(r)	Elefante	(nt)		
Llave	(Z)(b)	Tortuga	(rt)		
Silla	(s)(Z)	Agua	(ua)		
Chocolate	(tS)(k)(l)(t)				
Coche	(k)(tS)				
Mano	(m)(n)				
Naranja	(n)(r)(x)				
Luna	(l)(n)				
Vela	(b)(l)				
Señor	(s)(ñ)(r)				
Ratón	(r)(t)(n)				
Perro	(p)(r)				

Tabla A.2: Fragmento del relato “Fiesta en la Montaña”.

"El día amaneció gris y a través de la ventana se podía ver caer la nieve que difundía en la pieza del hotel un color mágico, en el que se desvanecían las paredes manchadas, adquiriendo una suerte de dignidad de obra de arte la horrible marina de mueblería colgada frente a la cama, a la que admiré como si se tratara de un Gauguin.
De pronto, descubrí que no estaba solo y con fastidio me preparé para el inevitable diálogo con la ocasional pasajera de mediana edad que encontré no tantas horas antes en el bar del residencial y que luego de tres o cuatro whiskies y casi por inercia llevé a mi habitación para cumplir en forma cuasi chacarera con el deber del momento, mediocre proeza que inexplicablemente fue festejada con entusiasmo por esa mujer rubia, pálida y de lacia cabellera que dormía a mi lado y cuyo nombre me resultaba imposible recordar."

Tabla A.3: Frases diseñadas para adaptación.

No	Frase
1	EL EXTRAÑO NIÑO ESTÁ LLORANDO MUCHO
2	EL RATÓN JALÓ LA AZÚCAR
3	EL FÚTBOL LLANERO MUEVE MUCHA AFICIÓN
4	LA FAJA TALLA EXTRA ESTÁ ALREVÉS
5	EL PAJARO YA ESTA EN LA JAULA
6	EL GATO GRUÑÓ MUY FUERTE
7	EL ELEFANTE ES MAS GRANDE EN ÁFRICA
8	EL CHANGO ES PEQUEÑO EN AMÉRICA
9	ASÍ EL BARCO AVANZÓ RÁPIDO
10	MI FAMILIA VIVIÓ EN MÉXICO
11	MI MAMÁ CUMPLE AÑOS EXTRA MAÑANA
12	EL TÍO COMIÓ POLLO CHINO
13	SEGÚN ELLOS LA SÍLABA ES CORRECTA
14	LA PIEZA EXHUMADA ES ÚNICA Y CARACTERÍSTICA
15	ALGÚN DÍA VOLVERÉ Y VENCERÉ
16	AQUÍ LLOVIO MUCHO DESDE ANOCHE

Anexo B

Alineador Fonético

El Alineador Fonético usado para este proyecto es una variación del algoritmo de Alineamiento Dinámico en el Tiempo (DTW, Dynamic Time Warping) presentado en [5]. Los cambios implementados en este algoritmo se presentan en la Sección B.1.

El problema abordado consiste en encontrar el alineamiento óptimo de dos cadenas de fonemas:

$$P = \{p^j, \cdots, p^n\},\ j = 1, \cdots, n \tag{B.1}$$

$$\tilde{P} = \{\tilde{p}^i, \cdots, \tilde{p}^m\},\ i = 1, \cdots, m \tag{B.2}$$

El proceso de alineamiento consta de dos fases:

1. *Fase Hacia Adelante (Forward Phase).* La matriz $D(i, j)$ es creada para guardar las puntuaciones asignadas al alineamiento a cada par de sub-cadenas de fonemas $\{p^j, \tilde{p}^i\}$. Esta matriz tiene dimensiones $(m + 1)$ por $(n + 1)$.

2. *Fase de Rastreo (Trace-back Phase).* El alineamiento óptimo es reconstruído al rastrear en D cualquier camino desde $D(m, n)$ hasta $D(1, 1)$ que genere la puntuación máxima.

Detalles de cada fase se muestran en las siguientes secciones.

B.1 Fase Hacia Adelante

La matriz D contiene las diferentes puntuaciones asignadas al alineamiento de dos cadenas de fonemas P=$\{p^{j=1,\ldots,n}\}$ y $\tilde{P}$=$\{\tilde{p}^{i=1,\ldots,m}\}$. En donde $D(i, j)$ contiene la mejor

puntuación (distancia) de alinear un par de fonemas $\{p^j, \tilde{p}^i\}$ de las cadenas $P, \tilde{P}$. Si ambos fonemas son iguales, se tendría un emparejamiento óptimo, lo cual daría una puntuación alta. Si ambos son diferentes, la puntuación se espera que sea baja. En la Tabla B.1 se muestra el pseudo-código para el cálculo de los elementos de la matriz de distancias D. Durante el cálculo de D, hay algunas variables: *valor_espacio*, que es un valor constante de -2 que equivale a asignar un vacío ("-") a algún caractér de P o $\tilde{P}$. Emparejar un fonema de P a "-" equivale a definir una eliminación, en tanto que emparejar "-" a un fonema en $\tilde{P}$ equivale a definir una inserción.

Tabla B.1: Pseudo-código de la Fase Hacia Adelante

(a) Inicialización de la primer fila y columna de *D*
D=ceros (*m*+1, *n*+1) % Inicializa en ceros la matriz *D*
for *j*=1 hasta *n*
D(1, *j*+1) = *valor_espacio***j*
end
for *i*=1 hasta *m*
D(*i*+1, 1) = *valor_espacio***i*
end
(b) Calcular todos los valores para *D*(*i,j*)
for *i*=2 hasta *m*+1
for *j*= 2 hasta *n*+1
$Emparejamiento$ = *D*(*i*-1, *j*-1) + $Sim\left(p^{j-1}, \tilde{p}^{i-1}\right)$
$Espacio\_en\_P$= *D*(*i*, *j*-1) + *valor_espacio*
$Espacio\_en\_\tilde{P}$= *D*(*i*-1, *j*) + *valor_espacio*
D(*i,j*) = *max* ($Emparejamiento$, $Espacio\_en\_P$, $Espacio\_en\_\tilde{P}$)
end
end
Puntuacion = *D*(*m*+1, *n*+1) % Mejor puntuación del alineamiento local

$D(i, j)$ considera la puntuación en base a una similitud entre P y $\tilde{P}$. Si son muy diferentes entonces $D(i, j)$ tendrá un valor muy bajo, lo cual conllevará a clasificar un fonema como eliminación o inserción. Sin embargo es necesario considerar la similitud acústica entre fonemas para evitar la discriminación de fonemas, e identificar sustituciones de manera más adecuada (emparejamiento de fonemas diferentes). Esto es, considerar qué tan lejos o cerca, acústicamente hablando, se encuentran ciertos fonemas. Para esto, el cálculo de $D(i, j)$ considera una puntuación adicional, que es proporcionada por una matriz $Sim(p^j, \tilde{p}^i)$. Esta matriz considera la similitud acústica entre

fonemas y la pondera como una probabilidad de confusión. Fonemas muy parecidos, por ejemplo, $/a/$ y $/e/$, tendrán una probabilidad de confusión significativa como 0.50. En tanto, fonemas idénticos tendrán la máxima probabilidad (> 0.90). Fonemas muy diferentes tendrán la mínima probabilidad ($/a/$ y $/k/$, < 0.10). En [3] se mostró que $Sim(p^j, \tilde{p}^i)$ se podía estimar de manera eficiente a partir de la salida de un SRAH independiente de usuario. En este trabajo, estas probabilidades se normalizaron en base a la siguiente expresión empírica:

$$Sim(p^j, \tilde{p}^i) = 8Pr_{SI}(q^j, \tilde{q}^i) - 3 \tag{B.3}$$

En la Tabla B.2 se muestra la matriz de puntuaciones (o ponderaciones) de similitud para los fonemas del español mexicano. Una vez que se calcula D, se procede a rastrear el camino a través de esta matriz que contenga el máximo de puntuaciones acumuladas (cuyo valor se almacena en $D(m+1, n+1)$). Esto se hace en la siguiente fase.

Tabla B.2: Matriz de ponderaciones para el alineador fonético.

	a	e	i	o	u	b	d	_D	f	g	_G	k	ks	l	m	n	_N	ñ	p	r	r(	_R	s	t	tS	x	Z	sil	sp
a	5	1.25	0.75	0.5	0.25	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
e	1.25	5	1.25	0.75	0.5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
i	0.75	1.25	5	1.25	0.75	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
o	0.5	0.75	1.25	5	1.25	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
u	0.25	0.5	0.75	1.25	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
b	-3	-3	-3	-3	-3	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
d	-3	-3	-3	-3	-3	-3	5	1.75	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
_D	-3	-3	-3	-3	-3	-3	1.75	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
f	-3	-3	-3	-3	-3	-3	-3	-3	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
g	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	1.75	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
_G	-3	-3	-3	-3	-3	-3	-3	-3	-3	1.75	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
k	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	0.25	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
ks	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	0.25	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
l	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
m	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	0.25	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
n	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	0.25	5	1.75	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
_N	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	1.75	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
ñ	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
p	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
r	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	1.75	1.75	-3	-3	-3	-3	-3	-3	-3
r(	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	1.75	5	1.75	-3	-3	-3	-3	-3	-3	-3
_R	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	1.75	1.75	5	-3	-3	-3	-3	-3	-3	-3
s	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	-3	-3	-3	-3	-3	-3
t	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	0.25	-3	-3	-3	-3
tS	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	0.25	5	-3	-3	-3	-3
x	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	-3	-3	-3
Z	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	-3	-3
sil	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	1.75
sp	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	1.75	5

B.2 Fase de Rastreo

En la Tabla B.3 se muestra el pseudo-código para esta fase. Note que dos arreglos son generados para almacenar el alineamiento final, $Alineamiento\_en\_P$ y $Alineamiento\_en\_\tilde{P}$, en donde se realizan las siguientes asignaciones:

- Si $D(i-1, j-1)$ pertenece al camino óptimo, el siguiente elemento de P y $\tilde{P}$ se añade (o guarda) en $Alineamiento\_en\_P$, y $Alineamiento\_en\_\tilde{P}$.
- Si $D(i, j-1)$ pertenece al camino óptimo, el siguiente elemento de P se añade a $Alineamiento\_en\_P$, y un vacío ("-") se añade a $Alineamiento\_en\_\tilde{P}$.
- If $D(i-1, j)$, pertenece al camino óptimo, el siguiente elemento de $\tilde{P}$ se añade a $Alineamiento\_en\_\tilde{P}$, y un vacío ("-") se añade a $Alineamiento\_en\_P$.

Tabla B.3: Pseudo-código de la Fase de Rastreo

```
w = 1 % índice del arreglo alineado
while i>1 y j>1 do
    if D(i,j) - Sim(p^{j-1}, p̃^{i-1}) = D(i-1, j-1) then
        Alineamiento_en_P(w) = p^{j-1}
        Alineamiento_en_P̃(w) = p̃^{j-1}
        i = i-1
        j = j-1
        w = w+1
    else
        if D(i,j) - valor_espacio= D(i, j-1) then
            Alineamiento_en_P(w) = p^{j-1}
            Alineamiento_en_P̃(w) = " − "
            j = j-1
            w = w+1
        else
            if D(i,j) - valor_espacio= D(i-1, j) then
                Alineamiento_en_P(w) = " − "
                Alineamiento_en_P̃(w) = p̃^{j-1}
                i = i-1
                w = w+1
            end
        end
    end
end
```

Anexo C

Perfiles de Candidatos

- Señora **JC**, 82 años (SNDIF). JC había sufrido una Embolia, presentando un nivel diagnosticado de disartria alto. Recibía terapia una vez a la semana. Para iniciar el proceso de aceptación se platicó con ella y con los familiares para estar presente en algunas de sus terapias como observador y explicarles las razones de la visita.

  Al estar como observador en algunas terapias, se pudo percatar del vocabulario y/o la forma de interactuar con el terapeuta, el paciente, y los familiares, al igual que la comodidad del paciente con mi presencia. Posterior a la observación se inicio la interacción con JC, se platicó, se hizó contacto físico a parte del saludo habitual. También se conversó con el familiar al respecto de cómo se sentía anímicamente JC, el tiempo que llevaba en esas condiciones, la actitud de JC en sus terapias y en su vida cotidiana, cómo era JC antes de su embolia. Esto es muy importante ya que todo trabajo con usuarios, especialmente si tienen alguna discapacidad, debe basarse en la confianza y comprensión de la persona.

  Aunque JC y sus familiares aceptaron colaborar en la realización del proyecto, JC comenzó a perder la paciencia en la realización de sus actividades cotidianas por sus propias limitantes físicas, mostrando molestia e irritación incluso mientras tomaba su terapia ocupacional. En ese punto, al observar la situación de JC y no haber comenzado ningún trabajo práctico con ella, se optó por descartarla como candidato.

- Niño **UO**, 9 años (SNDIF). UO fue considerado a pesar de ser menor de edad dado

que daba la impresión de que padecía disartria ya que emitía gemidos para señalar cosas o personas, aunque llegaba a articular algunas palabras. Sin embargo llevaba sólo algunas sesiones en el SNDIF y no tenía un diagnóstico concreto, teniendo como posible discapacidad en el habla la dislalia sin identificar la causa de la misma. También se observó que aunque tenía conocimiento de ciertas palabras las asociaba con diferentes significados. Por lo tanto no se sabía si su discapacidad consistía en un problema de lenguaje neurológico, del habla, o si era causado porque no había recibido atención por parte de su familia para el desarrollo de sus habilidades de lenguaje. Por estas razones se descartó como candidato.

- Señor **MM**, 37 años (Personal). Por medio de referencias personales se contactó con MM, quien presentaba disartria leve debido a un accidente automovilístico hace 20 años en el cual sufrió daños en sus cuerdas vocales por una inadecuada traqueotomía. MM decidió colaborar con el proyecto y proporcionó muestras de voz para su análisis. Al ir avanzado en el proyecto MM dejó de colaborar dado que tuvo una oferta de trabajo lo cual le demandó más tiempo. Sin embargo dejó abierta la posibilidad de cooperación para el futuro, la cual fue concretada posteriormente y cuyos resultados se reportan en este proyecto.

- Señora **MC**, 32 años (SNDIF). MC asistía como familiar de apoyo a un paciente del SNDIF, y de acuerdo a la observación del terapeuta presentaba un problema del habla. Al conversar con ella, comentó que creció con el problema en la forma de hablar (voz cortada) y era la única persona en su familia que presentaba el problema, habiendo recibido ayuda de terapias de lenguaje en su niñéz sin resultados. Comentó que de recién nacida tuvo una caída de una hamaca y que pudo haber sido la razón de su problema. También debido a los escasos recursos económicos de la familia no se trató o indagó más en un diagnóstico adecuado. En esta charla se concretó otra cita, pero posterior a ello dejo de asistir al SNDIF y no se tuvo más contacto con MC.

- Señor **GJ**, 64 años. Oriundo y radicado en el Municipio de Santa Cruz Tacache de Mina, quien presentaba disartria causada por un accidente vascular cerebral. Se platicó con él y su familia para poder acceder a alguna de sus terapias que recibe (una vez a la semana) en el SNDIF, explicándole el propósito del proyecto. Al

igual que en el caso del candidato JC se observó la forma de interacción con el terapeuta y sus familiares. Desde el inicio GJ mostró interés en el proyecto y su colaboración fue concretada. Sus resultados se reportan en este proyecto.

Anexo D

Frases de Adaptación y Evaluación

Tabla D.1: Grupos de Frases para Adaptación Dinámica I y II de la Interfaz de Voz.

	Frases para Adaptación Dinámica I
1	UNA COCA FRIA
2	UN VASO DE REGRESCO DE MANZANA
3	SALIR MAÑANA AL PARQUE
4	ENCIENDE LA LUZ
5	APAGA LA LUZ
6	DAME LA SILLA
7	DAME EL PEINE O CEPILLO
8	SUBE EL VOLUMEN AL RADIO
9	BAJA EL VOLUMEN A LA TELEVISION
10	QUIERO BAÑARME
11	LAVARME LAS MANOS

	Frases para Adaptación Dinámica II
1	VER LA TELEVISION
2	VER LAS NOTICIAS
3	ESCUCHAR EL RADIO
4	HACER EJERCICIOS
5	UN PEDAZO DE SANDIA
6	UNA TAZA DE CHOCOLATE TIBIO
7	UNA TORTILLA CON QUESO
8	IR A DORMIR TEMPRANO
9	TOMAR UN POCO DE AGUA DE NARANJA
10	TOMAR UNA TAZA DE ATOLE CALIENTE
11	IR A TERAPIA

Tabla D.2: Grupo de Frases para Evaluación de la Interfaz de Voz y de LTN Dragon

1	COMER FRIJOLES CON HUEVO	26	COMER HUEVOS ESTRELLADOS CON SALSA
2	COMER GELATINA DE PIÑA	27	UN VASO DE AGUA FRIA
3	COMER SOPA DE POLLO	28	UNA TAZA DE LECHE CALIENTE
4	COMER UNA TAJADA DE PAPAYA	29	UNA GELATINA DE UVA
5	TOMAR UN VASO DE AGUA DE HORCHATA	30	COMER CALDO DE RES
6	TOMAR UN VASO DE JUGO DE MANZANA	31	COMER TORTILLAS
7	TOMAR REFRESO DE NARANJA	32	COMER ARROZ CON POLLO
8	TOMAR UNA TASA DE CHOCOLATE CALIENTE	33	TOMAR UNA TAZA DE CAFE CALIENTE
9	TOMAR UN VASO DE LECHE FRIA	34	QUIERO COMER UN PAN DE DULCE
10	UNA TAJADA DE MELON	35	PASAME UNA COBIJA
11	UNA REBANADA DE PAPAYA	36	QUIERO ESCUCHAR MUSICA
12	UN TACO DE HUEVO	37	COMPRAME UN DISCO DE MUSICA RANCHERA
13	UNA REBANADA DE SANDIA	38	DAME LA ANDADERA
14	UNA TAZA DE TE CALIENTE	39	ME SUBO SOLO AL CARRO
15	UN VASO DE JUGO DE UVA	40	ME SIENTO SOLO EN LA SILLA
16	UN VASO DE REFRESCO DE MANZANA	41	PONGAME EL PAÑAL
17	UN TACO DE HUEVO	42	DAME LA PLAYERA, EL SHORT O PANS
18	UN PAN CON HUEVO Y FRIJOLES	43	PASAME LOS LENTES
19	UN TACO DE POLLO Y CAFE CON PAN	44	LA PASTILLA DE LA PRESION
20	SALIR A TOMAR AIRE	45	COMPRAME UN BIMBO
21	COMER GELATINA DE LIMON	46	VENGAN PARA JUGAR UN RATO
22	COMER SOPA DE RES	47	COMER UN PEDAZO DE JICAMA
23	COMER UNA REBANADA DE PIÑA	48	COMER UN POCO DE ARROZ
24	COMER UN PEDAZO DE PAPAYA	49	UNA TAZA DE LECHE CALIENTE CON PAN
25	TOMAR UN VASO DE JUGO DE NARANJA	50	UN TACO DE ARROZ CON POLLO

Tabla D.3: Texto de adaptación para LTN Dragon

AVENTURAS DE PINOCHO

CAPITULO UNO | DE COMO EL CARPINTERO MAESTRO CEREZA ENCONTRO UN TROZO DE MADERA QUE LLORABA Y REIA COMO UN NIÑO | PUES SEÑOR HABIA UNA VEZ UN REY | DIRAN ENSEGUIDA MIS PEQUEÑOS LECTORES | PUES NO MUCHACHOS NADA DE ESO | HABIA UNA VEZ UN PEDAZO DE MADERA | PERO NO UN PEDAZO DE MADERA DE LUJO | SINO SENCILLAMENTE UN LEÑO DE ESOS CON QUE EN EL INVIERNO SE ENCIENDEN LAS ESTUFAS Y CHIMENEAS PARA CALENTAR LAS HABITACIONES
PUES SEÑOR ES EL CASO QUE DIOS SABE COMO EL LEÑO DE MI CUENTO FUE A PARAR CIERTO DIA AL TALLER DE UN VIEJO CARPINTERO | CUYO NOMBRE ERA MAESTRO ANTONIO PERO AL CUAL LLAMABA TODO EL MUNDO MAESTRO CEREZA | PORQUE LA PUNTA DE SU NARIZ SIEMPRE COLORADA Y RELUCIENTE PARECIA UNA CEREZA MADURA
CUANDO MAESTRO CEREZA VIO AQUEL LEÑO SE PUSO MAS CONTENTO QUE UNAS PASCUAS | TANTO QUE COMENZO A FROTARSE LAS MANOS MIENTRAS DECIA PARA SU CAPOTE | HOMBRE LLEGAS A TIEMPO | VOY A HACER DE TI LA PATA DE UNA MESA | DICHO Y HECHO COGIO EL HACHA PARA COMENZAR A QUITARLE LA CORTEZA Y DESBASTARLO | PERO CUANDO IBA A DAR EL PRIMER HACHAZO SE QUEDO CON EL BRAZO LEVANTADO EN EL AIRE PORQUE OYO UNA VOCECITA MUY FINA | MUY FINA QUE DECIA CON ACENTO SUPLICANTE NO | NO ME DES TAN FUERTE | FIGURENSE COMO SE QUEDARIA EL BUENO DE MAESTRO CEREZA | SUS OJOS ASUSTADOS RECORRIERON LA ESTANCIA PARA VER DE DONDE PODIA SALIR AQUELLA VOCECITA Y NO VIO A NADIE
MIRO DEBAJO DEL BANCO Y NADIE | MIRO DENTRO DE UN ARMARIO QUE SIEMPRE ESTABA CERRADO Y NADIE EN EL CESTO DE LAS ASTILLAS Y DE LAS VIRUTAS Y NADIE | ABRIO LA PUERTA DEL TALLER SALIO A LA CALLE Y NADIE TAMPOCO
QUE ERA AQUELLO | YA COMPRENDO DIJO ENTONCES SONRIENDO Y RASCANDOSE LA PELUCA | ESTA VISTO QUE ESA VOCECITA HA SIDO UNA ILUSION MIA | REANUDEMOS LA TAREA Y TOMANDO DE NUEVO EL HACHA PEGO UN FORMIDABLE HACHAZO EN EL LEÑO | AY ME HAS HECHO DAÑO DIJO QUEJANDOSE LA MISMA VOCECITA | ESTA VEZ SE QUEDO MAESTRO CEREZA COMO SI FUERA DE PIEDRA CON LOS OJOS ESPANTADOS LA BOCA ABIERTA Y LA LENGUA DE FUERA | COLGANDO HASTA LA BARBA COMO UNO DE ESOS MASCARONES TAN FEOS Y TAN GRACIOSOS POR CUYA BOCA SALE EL CAÑO DE UNA FUENTE. SE QUEDO HASTA SIN VOZ | CUANDO PUDO HABLAR COMENZO A DECIR TEMBLANDO DE MIEDO Y BALBUCEANDO | PERO DE DONDE SALE ESA VOCECITA QUE HA DICHO AY | SI AQUI NO HAY UN ALMA | SERA QUE ESTE LEÑO HABRA APRENDIDO A LLORAR Y A QUEJARSE COMO UN NIÑO | YO NO PUEDO CREERLO ESTE LEÑO | AQUI ESTA ES UN LEÑO DE CHIMENEA COMO TODOS LOS LEÑOS DE CHIMENEA | BUENO PARA ECHARLO AL FUEGO Y GUISAR UN PUCHERO DE HABICHUELAS | ZAMBOMBA SE HABRA ESCONDIDO ALGUIEN DENTRO DE EL | AH PUES SI ALGUNO SE HA ESCONDIDO DENTRO PEOR PARA EL AHORA LE VOY A ARREGLAR YO
Y DICIENDO ESTO AGARRO EL POBRE LEÑO CON LAS DOS MANOS Y EMPEZO A GOLPEARLO SIN PIEDAD CONTRA LAS PAREDES DEL TALLER | DESPUES SE PUSO A ESCUCHAR SI SE QUEJABA ALGUNA VOCECITA | ESPERO DOS MINUTOS Y NADA CINCO MINUTOS Y NADA | DIEZ MINUTOS Y NADA
YA COMPRENDO DIJO ENTONCES TRATANDO DE SONREIR Y ARREGLANDOSE LA PELUCA | ESTA VISTO QUE ESA VOCECITA QUE HA DICHO AY HA SIDO UNA ILUSION MIA | REANUDEMOS LA TAREA | Y COMO TENIA TANTO MIEDO SE PUSO A CANTURREAR PARA COBRAR ANIMOS | ENTRE TANTO DEJO EL HACHA Y TOMO EL CEPILLO PARA CEPILLAR Y PULIR EL LEÑO | PERO CUANDO LO ESTABA CEPILLANDO POR UN LADO Y POR OTRO OYO LA MISMA VOCECITA QUE LE DECIA RIENDO | PERO HOMBRE QUE ME ESTAS HACIENDO UNAS COSQUILLAS TERRIBLES | ESTA VEZ MAESTRO CEREZA SE DESMAYO DEL SUSTO
CUANDO VOLVIO A ABRIR LOS OJOS SE ENCONTRO SENTADO EN EL SUELO QUE CARA DE BOBO SE LE HABIA PUESTO | LA PUNTA DE LA NARIZ YA NO ESTABA COLORADA DEL SUSTO SE LE HABIA PUESTO AZUL